# Davies Bouldin Index Algorithm for Optimizing Clustering Case Studies Mapping School Facilities

Yudhistira Arie Wijaya [1], Dedy Achmad Kurniady [2], Eddy Setyanto [3], Wahdan Sanur Tarihoran [4], Dadan Rusmana [5], Robbi Rahim [6]

*[1]STMIK IKMI Cirebon, Cirebon, Indonesia*
*[2]Universitas Pendidikan Indonesia, Bandung, Indonesia*
*[3]Universitas Indraprasta PGRI, Jakarta, Indonesia*
*[4]Universitas Islam Negeri Sultan Maulana Hasanuddin Banten, Indonesia*
*[5]UIN Sunan Gunung Djati Bandung, Bandung, Indonesia*
*[6]Sekolah Tinggi Ilmu Manajemen Sukma, Medan, Indonesia*

***Abstract*** **– The lower Davies Bouldin Index (DBI) is considered the best clustering algorithm based on the criteria that yields a cluster set. The purpose of this research is to optimize the clustering results using DBI. The data sources used are the number of villages that have school facilities and the level of education is obtained from the government website (https://www.bps.go.id). The level of education in question is senior high school and vocational high school. The method used is k-means. The results show that from the number of clusters (k = 2, 3, 4, 5, 6) the optimal DBI for (k = 2) is obtained with a value of 0.168 for Measure Type = Mixed Measures. For the value of k = 2, a mapping of areas with L0 (low) = 31 province and L1 (high) = 3 provinces is obtained. The final centroids obtained for each cluster are L0 (315 and 155) and L1 (1710 and 1259). Based on the results of mapping by optimizing k-means and DBI, more than 90% of the villages still have school facilities, especially at the high school and vocational high school levels.**

DOI: 10.18421/TEM103-13
https://doi.org/10.18421/TEM103-13

**Corresponding author:** Robbi Rahim,
*Sekolah Tinggi Ilmu Manajemen Sukma, Medan, Indonesia.*
**Email:** usurobbi85@zoho.com

*Received: 24 April 2021.*
*Revised: 25 June 2021.*
*Accepted: 01 July 2021.*
*Published: 27 August 2021.*

***Keywords*** **– clustering, k-means, Davies Bouldin Index, School Facilities, Mapping.**

## 1. Introduction

The appropriate evaluation metrics have to be found to measure the clustering performance. Most of the clustering metrics are designed to maximize class similarity and minimize class similarity [1]. The Davies Bouldin index (DBI) is a popular measure to assess clustering performance by dividing clusters [2] [3], [4]. DBI can show clustering quality with intracluster similarity and cluster-like similarity [5]. Clustering is the grouping process of an object group based on a number of similarities. Clustering is a data mining component [2], [6], [7], [8]. There are several familiar clustering methods [9], [10]. These include k-means [11], [12], [13]. In this research, we do not solve the problem of clustering explicitly, but rather, how the DBI is used to evaluate how well the clustering is formed [5]. DBI works by calculating the average value of each item in the data set. The value of each point is calculated as the sum of the compactness values divided by the distance between the two center points of the group as separation. The smaller DBI value indicates the best number of clusters [14]. The case raised is the number of villages with educational facilities in Indonesia. Schools are formal institutions which are expected to improve people. Schools are closely linked to schools as an educational institution [15]. School facilities are equipment or facilities used directly to support the education process, in particular education and learning processes such as buildings, rooms, tables, chairs and media tools. This facility is a means and infrastructure for facilitating educational activities in schools [16]. The research aims to optimize the mapping in the form of clusters for the number of school settlements in Indonesia using the DBI and k-means.

Several related studies, such as [17] on the centeroid cluster analysis x-means algorithm with the DBI assessment. This paper proposes to modify the x-means algorithm by evaluating the DBI to determine the number of centroid clusters. The results of the research indicate that the DBI value is close to 0. Next, [2] research on the DBI clustering assessment of cereal data using K-means. This paper proposes the k-means method and the DBI using the R-language. The results indicate that DBI can be used with results of k = 5 being better than k = 3, i.e. 1,498871 and 1,667952. In addition, research has been conducted [1] on the hierarchical initialization of the k-means of the DBI. This paper proposes a hierarchical initialization approach to determine the center of the initial cluster using the DBI hierarchical k-means (DHIKM) algorithm. The results suggest that the proposed algorithm can integrate DBI metrics into a hierarchical k-means algorithm and can determine the number of clusters at a low cost.

## 2. Methodology

### 2.1. *Data Set*

The data used for this analysis are data on the number of villages that have educational facilities in Indonesia. The data shall be from 2011, 2014 and 2018. These data are obtained from the official website of the *Badan Pusat Statistik* (BPS). The variables used are two variables, namely High School and High School. The data are processed using the Rapid Miner software. The data processed is the average number of villages with school-level facilities at senior high schools and vocational colleges. One can access https://osf.io/zmdjh pre-processing data.

### 2.2. *Davies Bouldin Index*

Davies Bouldin Index (DBI) [7], [18], [19] is a measure to evaluate the clustering performance. DBI has the positive correlation for the "within-class" case and negative correlation for the "between-class" case. Use DBI as a clustering metric because of the general way it is clustering Validation contains two main categories - external validation and internal validation which is used to assess the performance of clustering results [1].

### 2.3. *Design of the System*

The stage of the process in this study is as follows:

a) Prepare a data set for the number of villages that have educational facilities in Indonesia.
b) Indicator data will be processed for input using the RapidMiner software.
c) The school facility data process shall use the k-means clustering method and shall be stored under the weight name, including the calculation of the distance using several Measure Types between the weights.
d) Weights that have been calculated using the Measure Type will be shown (k = n). From the clustering process, an assessment was carried out using the DBI to determine the optimum number of clusters in the clustering process.

## 3. Results and Discussion

### 3.1. *Cluster Evaluation Results (Measure Type = Mixed Measures)*

The clustering results were obtained from five experiments that were performed (k= 2, 3, 4, 5, 6) for Measure Type = Mixed Measures, then the next step is to calculate the DBI value for each clustering experiment. The results of the DBI value can be found in Table 1 below.

*Table 1. The resulting cluster (Measure Type = Mixed Measures)*

| Measure Type | Cluster Set | DBI | Number of cluster members |
|---|---|---|---|
| Mixed Measures | 2 | 0.168 | L0: 31 items<br>L1: 3 items<br>Total number of items: 34 |
| | 3 | 0.248 | L0: 22 items<br>L1: 9 items<br>L2: 3 items<br>Total number of items: 34 |
| | 4 | 0.232 | L0: 22 items<br>L1: 2 items<br>L2: 2 items<br>L3: 8 items<br>Total number of items: 34 |
| | 5 | 0.26 | L0: 12 items<br>L1: 2 items<br>L2: 2 items<br>L3: 8 items<br>L4: 10 items<br>Total number of items: 34 |
| | 6 | 0.197 | L0: 12 items<br>L1: 2 items<br>L2: 10 items<br>L3: 1 items<br>L4: 8 items<br>L5: 1 items<br>Total number of items: 3 |

In Table 1, the DBI evaluation results obtained from k-means are 0.168 with the number of k = 2. Then the number of clusters k = 3 has a DBI value of 0.248. For the number of clusters k = 4, it has a DBI value of 0.232. Meanwhile, the number of clusters k = 5 and k = 6, has a DBI value of 0.26 and 0.197.

### *3.2. Cluster Evaluation Results (Measure Type = Bregmann Divergences - Mahalanobis Distance)*

The clustering results were obtained from five experiments that were performed (k= 2, 3, 4, 5, 6) for Measure Type = Bregmann Divergences-Mahalanobis Distance, followed by the calculation of the DBI value for each clustering experiment. The results for the DBI value can be found in Table 2 below.

*Table 2. The resulting cluster (Measurement Type = Bregmann Divergences-Mahalanobis Distance)*

| Measure Type | Cluster Set | DBI | Number of cluster members |
|---|---|---|---|
| Bregmann Divergences - Mahalanobis Distance | 2 | 0.208 | L0: 32 items<br>L1: 2 items<br>Total number of items: 34 |
| | 3 | 0.484 | L0: 5 items<br>L1: 2 items<br>L2: 27 items<br>Total number of items: 34 |
| | 4 | 0.333 | L0: 23 items<br>L1: 1 items<br>L2: 8 items<br>L3: 2 items<br>Total number of items: 34 |
| | 5 | 0.468 | L0: 14 items<br>L1: 2 items<br>L2: 3 items<br>L3: 14 items<br>L4: 1 items<br>Total number of items: 34 |
| | 6 | 0.582 | L0: 3 items<br>L1: 5 items<br>L2: 2 items<br>L3: 13 items<br>L4: 10 items<br>L5: 1 items<br>Total number of items: 34 |

In Table 2, the DBI evaluation results obtained from k-means are 0.208 with the number of k = 2. Then the number of clusters k = 3 has a DBI value of 0.484. For the number of clusters k = 4, it has a DBI value of 0.333. Meanwhile, the number of clusters k = 5 and k = 6, has a DBI value of 0.468 and 0.582.

### *3.3. Cluster Evaluation Results (Measure Type = Bregmann Divergences - Squared Euclidean Distance)*

The clustering results were obtained from five experiments that were performed (k= 2, 3, 4, 5, 6) for Measure Type = Bregmann Divergences – Squared Euclidean Distance, followed by the calculation of the DBI value for each clustering experiment. The results for the DBI value can be found in Table 3 below.

*Table 3. The resulting cluster (Measurement Type = Bregmann Divergences- Bregmann Divergences - Squared Euclidean Distance)*

| Measure Type | Cluster Set | DBI | Number of cluster members |
|---|---|---|---|
| Bregmann Divergences - Squared Euclidean Distance | 2 | 0.226 | L 0: 30 items<br>L 1: 4 items<br>Total number of items: 34 |
| | 3 | 0.211 | L 0: 30 items<br>L 1: 2 items<br>L 2: 2 items<br>Total number of items: 34 |
| | 4 | 0.232 | L 0: 22 items<br>L 1: 2 items<br>L 2: 2 items<br>L 3: 8 items<br>Total number of items: 34 |
| | 5 | 0.162 | L0: 8 items<br>L1: 2 items<br>L2: 1 items<br>L3: 22 items<br>L4: 1 items<br>Total number of items: 34 |
| | 6 | 0.197 | L0: 12 items<br>L1: 2 items<br>L2: 8 items<br>L3: 1 items<br>L4: 10 items<br>L5: 1 items<br>Total number of items: 34 |

In Table 3, the DBI evaluation results obtained from k-means are 0.226 with the number of k = 2. Then the number of clusters k = 3 has a DBI value of 0.211. For the number of clusters k = 4, it has a DBI value of 0.232. Meanwhile, the number of clusters k = 5 and k = 6, has a DBI value of 0.162 and 0.197.

### *3.4. Implementation of the Davies- Bouldin Index (DBI)*

In research using experiments (k = n), different mappings are produced according to the type of measurement used. Of the three measurement types used, the lowest DBI value is 0.168 with cluster 2 resulting in 2 clusters. The following is a recapitulation of the tables and graphs of each DBI based on the measurement types shown in Table 4 and Figure 1 below.

*Table 4. Results of the Recapitulation of Each DBI*

| Measure Type | C=2 | C=3 | C=4 | C=5 | C=6 |
|---|---|---|---|---|---|
| Mixed Measures | 0.168 | 0.248 | 0.232 | 0.26 | 0.197 |
| Bregmann Divergences - Mahalanobis Distance | 0.208 | 0.484 | 0.333 | 0.468 | 0.582 |
| Bregmann Divergences - Squared Euclidean Distance: | 0.226 | 0.211 | 0.232 | 0.162 | 0.197 |

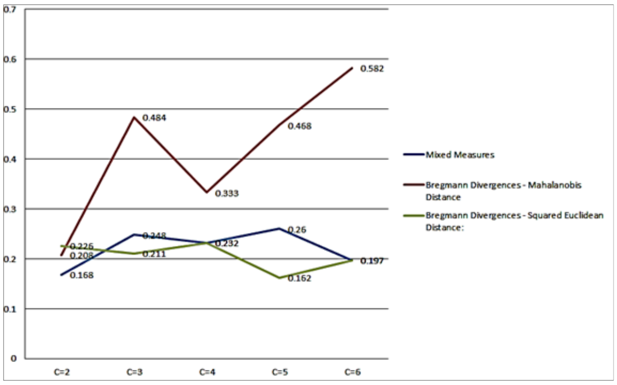
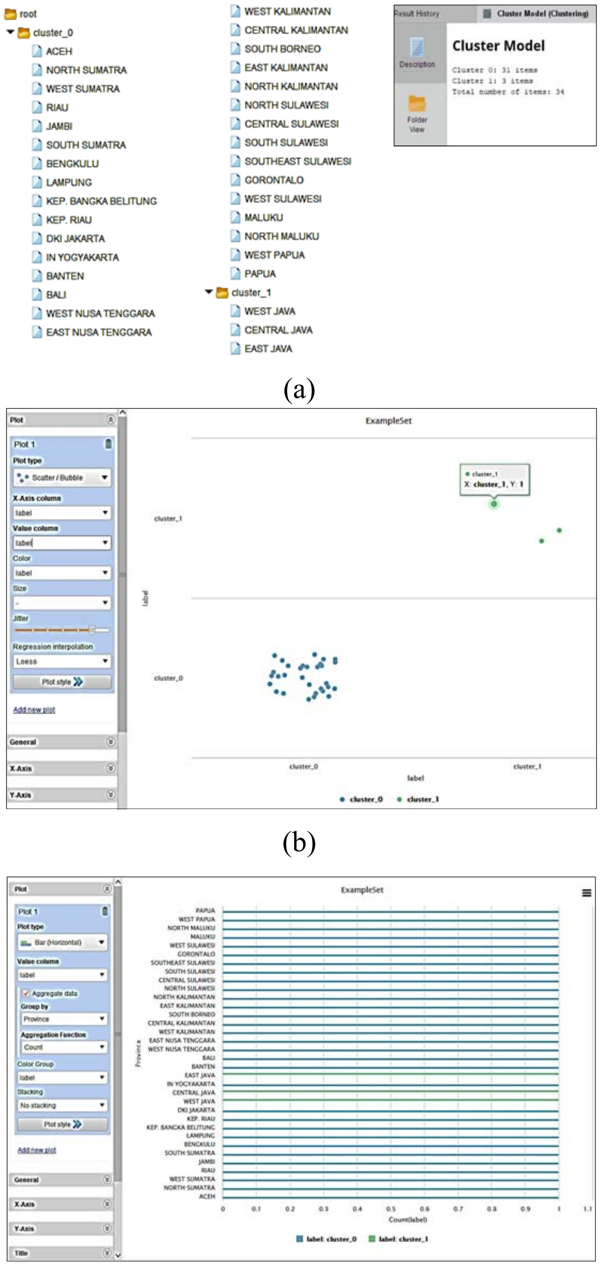


*Figure 1. Graph of DBI recapitulation results*

Figure 1 shows that the DBI value of each cluster is different. Starting from the sum of k = 2 to k = 6 based on the type of measurement. DBI is a tool for measuring the validity of clusters in a clustering method. The smaller the DBI value (DBI approaches 0), the more optimal the results of the cluster will be. For k = 2, the measure type = mixed measure has the closest DBI value to 0, i.e. 0.168. So that the results of mapping in the form of clusters against the number of villages with education-based school facilities are 31 provinces in cluster 0 (L0= low) and 3 provinces in cluster 1 (L1= high). Below are the cluster results using the RapidMiner software as shown below.



root
- cluster_0
  - ACEH
  - NORTH SUMATRA
  - WEST SUMATRA
  - RIAU
  - JAMBI
  - SOUTH SUMATRA
  - BENGKULU
  - LAMPUNG
  - KEP. BANGKA BELITUNG
  - KEP. RIAU
  - DKI JAKARTA
  - IN YOGYAKARTA
  - BANTEN
  - BALI
  - WEST NUSA TENGGARA
  - EAST NUSA TENGGARA
  - WEST KALIMANTAN
  - CENTRAL KALIMANTAN
  - SOUTH BORNEO
  - EAST KALIMANTAN
  - NORTH KALIMANTAN
  - NORTH SULAWESI
  - CENTRAL SULAWESI
  - SOUTH SULAWESI
  - SOUTHEAST SULAWESI
  - GORONTALO
  - WEST SULAWESI
  - MALUKU
  - NORTH MALUKU
  - WEST PAPUA
  - PAPUA
- cluster_1
  - WEST JAVA
  - CENTRAL JAVA
  - EAST JAVA

Result History | Cluster Model (Clustering)

Description | Folder View

Cluster Model

```
Cluster 0: 31 items
Cluster 1: 3 items
Total number of items: 34
```

(a)

(b)

(c)

*Figure 2. Cluster data set (a), Cluster graph by label (b), Cluster graph by region (c)*

Figure 2 shows the mapping results in the form of clusters at the minimum DBI value (k = 2; 0.168) for the number of villages with education-based school facilities, 3 provinces have high cluster results (L1), namely West Java, Central Java, and East Java. The rest of it goes to the low cluster (L0).

## 4. Conclusion

In this paper, we propose a k-means method optimized by the DBI to determine the optimum number of clusters. Our experiment is based on a dataset of the number of villages that have school facilities by level of education (high school and vocational high school). By performing different tests with the number of clusters (k = 2, 3, 4, 5, 6), the results of the k-means method with the type of measure are Mixed Measures with the most optimal DBI of 0.168 with cluster set 2 resulting in 2 clusters.

## References

[1]. Xiao, J., Lu, J., & Li, X. (2017). Davies Bouldin Index based hierarchical initialization K-means. *Intelligent Data Analysis*, *21*(6), 1327-1338. https://doi.org/10.3233/IDA-163129.

[2]. Singh, A. K., Mittal, S., Malhotra, P., & Srivastava, Y. V. (2020, March). Clustering Evaluation by Davies-Bouldin Index (DBI) in Cereal data using K-Means. In *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 306-310). IEEE.

[3]. Kärkkäinen, I., & Fränti, P. (2000, September). Minimization of the value of Davies-Bouldin index. In *Proceedings of the IASTED International Conference on Signal Processing and Communications (SPC'2000). IASTED/ACTA Press* (pp. 426-432).

[4]. Thomas, J. C. R., Peñas, M. S., & Mora, M. (2013, November). New version of Davies-Bouldin index for clustering validation based on cylindrical distance. In *2013 32nd International Conference of the Chilean Computer Science Society (SCCC)* (pp. 49-53). IEEE. https://doi.org/10.1109/SCCC.2013.29

[5]. Coelho, G. P., Barbante, C. C., Boccato, L., Attux, R. R., Oliveira, J. R., & Von Zuben, F. J. (2012, June). Automatic feature selection for BCI: an analysis using the davies-bouldin index and extreme learning machines. In *The 2012 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE. https://doi.org/10.1109/IJCNN.2012.6252500.

[6]. Wu, M., Li, X., Liu, C., Liu, M., Zhao, N., Wang, J., ... & Zhu, L. (2019). Robust global motion estimation for video security based on improved k-means clustering. *Journal of Ambient Intelligence and Humanized Computing*, *10*(2), 439-448. https://doi.org/10.1007/s12652-017-0660-8.

[7]. Feng, Z. K., Niu, W. J., Zhang, R., Wang, S., & Cheng, C. T. (2019). Operation rule derivation of hydropower reservoir by k-means clustering method and extreme learning machine based on particle swarm optimization. *Journal of Hydrology*, *576*, 229-238. https://doi.org/10.1016/j.jhydrol.2019.06.045

[8]. Sitompul, B. J. D., Sitompul, O. S., & Sihombing, P. (2019, June). Enhancement clustering evaluation result of davies-bouldin index with determining initial centroid of k-means algorithm. In *Journal of Physics: Conference Series* (Vol. 1235, No. 1, p. 012015). IOP Publishing.

[9]. Javadi, S., Hashemy, S. M., Mohammadi, K., Howard, K. W. F., & Neshat, A. (2017). Classification of aquifer vulnerability using K-means cluster analysis. *Journal of hydrology*, *549*, 27-37. https://doi.org/10.1016/j.jhydrol.2017.03.060

[10]. Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, *31*(3), 264-323. https://doi.org/10.1145/331499.331504

[11]. Nguyen, H., Bui, X. N., Tran, Q. H., & Mai, N. L. (2019). A new soft computing model for estimating and controlling blast-produced ground vibration based on hierarchical K-means clustering and cubist algorithms. *Applied Soft Computing*, *77*, 376-386. https://doi.org/10.1016/j.asoc.2019.01.042

[12]. Hossain, M. Z., Akhtar, M. N., Ahmad, R. B., & Rahman, M. (2019). A dynamic K-means clustering for data mining. *Indonesian Journal of Electrical engineering and computer science*, *13*(2), 521-526. https://doi.org/10.11591/ijeecs.v13.i2.pp521-526

[13]. Majhi, S. K., & Biswal, S. (2018). Optimal cluster analysis using hybrid K-Means and Ant Lion Optimizer. *Karbala International Journal of Modern Science*, *4*(4), 347-360. https://doi.org/10.1016/j.kijoms.2018.09.001

[14]. Vergani, A. A., & Binaghi, E. (2018, July). A soft davies-bouldin separation measure. In *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (pp. 1-8). IEEE. https://doi.org/10.1109/FUZZ-IEEE.2018.8491581

[15]. Lina, E. O. (2019). Pengaruh Jumlah Desa Yang Memiliki Fasilitas Sekolah Terhadap Penduduk Buta Huruf Di Provinsi Kepulauan Bangka Belitung. *AL-ISHLAH: Jurnal Pendidikan*, *11*(1), 71-81.

[16]. Subagyo, A., Anggrairi, E. S., & Radianto, D. O. (2019). Analisis Pengaruh Fasilitas Pendidikan Terhadap Tingkat Pengangguran dan Kemiskinan di Wilayah Indonesia Tahun 2018. Jurnal Pendidikan: Riset dan Konseptual, 3(2), 109-115.

[17]. Mughnyanti, M., Efendi, S., & Zarlis, M. (2020). Analysis of determining centroid clustering x-means algorithm with davies-bouldin index evaluation. In *IOP Conference Series: Materials Science and Engineering* (Vol. 725, No. 1, p. 012128). IOP Publishing. https://doi.org/10.1088/1757-899X/725/1/012128.

[18]. Tol, W. A., Komproe, I. H., Jordans, M. J., Susanty, D., & De Jong, J. T. (2011). Developing a function impairment measure for children affected by political violence: a mixed methods approach in Indonesia. *International journal for quality in health care*, *23*(4), 375-383. https://doi.org/10.1093/intqhc/mzr032.

[19]. Li, M., Xu, D., Zhang, D., & Zou, J. (2020). The seeding algorithms for spherical k-means clustering. *Journal of Global Optimization*, *76*(4), 695-708. https://doi.org/10.1007/s10898-019-00779-w.