Internal Cluster Validation on Earthquake Data in the Province of Bengkulu

D S Rini*, P Novianti, H Fransiska
Statistics Department, Faculty of Mathematics and Science, University of Bengkulu, Indonesia
*dyah.setyorini@unib.ac.id

Abstract. K-means method is an algorithm for cluster n object based on attribute to k partition, where k < n. There is a deficiency of algorithms that is before the algorithm is executed, k points are initialized randomly so that the resulting data clustering can be different. If the random value for initialization is not good, the clustering becomes less optimum. Cluster validation is a technique to determine the optimum cluster without knowing prior information from data. There are two types of cluster validation, which are internal cluster validation and external cluster validation. This study aims to examine and apply some internal cluster validation, including the Calinski-Harabasz (CH) Index, Sillhouette (S) Index, Davies-Bouldin (DB) Index, Dunn Index (D), and S-Dbw Index on earthquake data in the Bengkulu Province.

The calculation result of optimum cluster based on internal cluster validation is CH index, S index, and S-Dbw index yield k = 2, DB Index with k = 6 and Index D with k = 15. Optimum cluster (k = 6) based on DB Index gives good results for clustering earthquake in the Bengkulu Province.

1. Introduction
Cluster analysis is a multivariate technique that has the main purpose to classify objects based on similarity characteristics. The similarity between objects is measured using the distance. Characteristics of objects in a cluster have a high level of similarity, while the characteristics of objects in a cluster with other clusters have a low level of similarity. In other words, the variance in a cluster is minimum while the variance between clusters is maximum [1].

K-means method is an algorithm for cluster n object based on attribute to k partition, where k < n. The clustering of objects is seen from the distance of the object to the closest centroid. After knowing the nearest centroid, the object will be classified as a member of that category. The K-Means algorithm is as follows: a. Select k centroid point randomly, b. Group the data so that k clusters with the point of the centroid of each cluster is the centroid point that has been selected previously, c. Update the centroid point value, 4. Repeat steps 2 and 3 until the value from the centroid point no longer changes. This shows that there is a shortage of algorithms before the algorithm is executed, the points are initialized randomly so that the grouping of the resulting data can vary, so never know the actual cluster by using the same data. If the random value for initialization is not good, then the resulting grouping becomes less optimal.

Cluster validation is a technique for determining optimal cluster number without knowing prior information from data. In general, there are two types of cluster validation, which are external cluster
validation and internal cluster validation. In situations where external information is not available, the internal validation index is the only way to validate the cluster. Internal cluster validation consists of several types including the Calinski-Harabasz (CH) Index, Sillhouette (S) Index, S-Dbw, Davies-Bouldin (DB) Index, Krzanowski-Lai (KL) Index, Dunn Index (D), and etc.

Liu, et al (2010) performed a comparison of 11 internal cluster validations on crisp clustering. The result of the research indicates that the S-Dbw index is a good internal cluster validation index in all aspects, while other index has certain limitations [2]. Rendon, et al (2011) discusses the comparison of internal versus external cluster validation on the K-means and Bisecting K-means algorithms. The study found that internal cluster validation is more accurate in determining the optimal number of clusters. The cluster validation index used in the study is BIC index, CH index, DB index, S Index, NIVA index, and Dunn index [3]. Arbelaitz, et al (2013) examined a comparison of 30 internal cluster validation indexes on various data characteristics [4]. Novianti, et al (2016) classify earthquake events and create zones based on earthquake classification characteristics using K-means cluster analysis. The cluster validation used in this research is Krzanowski-Lai (KL) index with optimal cluster number of 7 clusters [5]. Therefore, the researcher is interested to apply the index of other internal cluster validation on the earthquake data of Bengkulu province so that it is expected to obtain the best clustering for earthquake area in Bengkulu province.

2. Literature Review

2.1. K-Means cluster analysis
K-means cluster analysis seeks to partition the n individuals in a set of multivariate data into K clusters, where each individual in the dataset is allocated entirely to a particular cluster. As a hard partitioning algorithm, K-means cluster analysis is an iterative process. First, data are initially partitioned. Each group is calculated its mean and then the data partitioned again by allocating each datum to its nearest mean cluster position [6,7].

MacQueen suggests the use of K-Means to describe the algorithm in the determination of an object in a certain cluster by the size of the nearest center (mean). In its simplest form, this process consists of three stages [8]:
a. Partition objects into K initial cluster
b. Begin by noting the objects, determine an object into a cluster that which has the closest centroid (mean). (The distance is usually calculated using the Euclidean distance with either standardized or unstandardized observation). Recalculate the cluster centroid to get a new object and for the cluster that lost object. Centroid of the group is calculated by calculating the average value of the data is as follows:

\[ C_{kj} = \frac{x_{kj} + x_{kj} + \cdots + x_{kj}}{a}, \quad j = 1, 2, \cdots, p \]

\[ C_{kj} = \text{centroid of group-k, variable-j} \]
\[ a = \text{the number of members in the group k} \]
c. Step b is repeated until no more transfer of the object.
Euclidean distance is the distance the most common type selected. Its simplicity is the geometric distance in multiple dimension of space. Euclidean distance is usually calculated from the raw data, and not of the standard data. This method has several advantages, including the distance of any two objects are not affected by the addition of new objects to be analyzed, which may be an outlier.

2.2. Cluster validation
Cluster validation is a technique for determining optimal cluster number without knowing prior information from data. In general, there are two types of cluster validation, which are external cluster validation and internal cluster validation. External cluster validation is a cluster validation technique based on prior knowledge of data. While internal cluster validation is a cluster validation technique
based on intrinsic information from the data itself [3]. In situations where external information is not available, the internal validation index is the only way to validate cluster [2].

Internal cluster validation index consists of several methods, among which are as follows:

- **Calinski-Harabasz (CH) Index**
  The Calinski-Harabasz (CH) index evaluates the cluster validity based on the average between and within cluster sum of square [2]. This Index is computed by [3]:
  \[
  CH = \frac{\text{trace}(S_B)}{\text{trace}(S_W)} \frac{n_p - 1}{n_p - k}
  \]
  dengan
  \( S_B \): the between-cluster scatter matrix
  \( S_W \): the internal scatter matrix
  \( n_p \): the number of clustered samples
  \( k \): the number of clusters

  The optimal cluster number is determined by maximizing the value of this index [2].

- **Silhouette (S) Index**
  Silhouette (S) Index [2] validates the clustering performance based on the pairwise difference between and within cluster distance. In addition, The optimal cluster number is determined by maximizing the value of this index. It is defined as [3]:
  \[
  S = \frac{b(i) - a(i)}{\text{Max}\{a(i), b(i)\}}
  \]
  where
  \( a(i) \): the average distance between the i-th sample and all the samples included in \( X_j \)
  \( b(i) \): the minimum average distance between the i-th and all of the samples clustered in \( X_k, k = 1, ..., c, k \neq j \)

- **Davies-Bouldin (DB) Index**
  This index aim to identify sets of clusters that are compact and well separated. The Davies-Bouldin index is defined as:
  \[
  DB = \frac{1}{c} \sum_{i=1}^{c} \text{Max}_{i \neq j} \left( \frac{d(X_i) + d(X_j)}{d(c_i, c_j)} \right)
  \]
  where
  \( c \): the number of clusters
  \( i, j \): cluster labels
  \( d(X_i) \): all samples in clusters i and j to their respective cluster centroid
  \( d(c_i, c_j) \): the distance between these centroid

  Smaller value of DB indicates a better clustering solution [3].

- **Dunn (D) Index**
  Dunn (D) index is defined as:
Large values of $D$ correspond to good clustering solution [3].

- **S-Dbw Index**
  [4] This is a ratio-type index that has a more complex formulation based on the Euclidean norm $\|X\|=(X^TX)^{1/2}$, the standard deviation of a set of objects $\sigma(X) = 1/|X| \sum_{x_i \in X} (x_i - \bar{x})^2$, and the standard deviation of a partition $\text{stdev}(C) = 1/K \sqrt{\sum_{c_k \in C} \|\sigma(c_k)\|}$. Its definition is:

$$S-\text{Dbw}(C) = \frac{1}{K} \sum_{c_k \in C} \frac{\|\sigma(c_k)\|}{\|\sigma(X)\|} + \frac{1}{K(K-1)} \sum_{c_k \in C} \sum_{c_l \in C \setminus c_k} \frac{\text{den}(c_k, c_l)}{\max\{\text{den}(c_k), \text{den}(c_l)\}}$$

where

$$\text{den}(c_k) = \sum f(x_i, \bar{c}_k)$$

$$\text{den}(c_k, c_l) = \sum_{x_i, x_j \in c_k \cup c_l} f\left(x_i, \frac{\bar{c}_k + \bar{c}_l}{2}\right)$$

$$f(x_i, c_k) = \begin{cases} 0 & \text{if } d_x(x_i, \bar{c}_k) > \text{stdev}(C) \\ 1 & \text{otherwise} \end{cases}$$

[3] The minimum value of S-Dbw indicates the optimal cluster number.

### 2.3. Earthquake events in Bengkulu Province

Bengkulu Province is one of the provinces on the island of Sumatra which lies to the west of the Bukit Barisan mountains. The total area of Bengkulu Province reaches approximately 1,991,933 hectares or 19,919.33 square kilometers. Bengkulu Province extends from the border of West Sumatra Province to the border of Lampung Province and its distance is approximately 567 kilometers.

Astronomically, Bengkulu Province lies between 2 ° 16' to 3 ° 31' S and between 101 ° 01' to 103 ° 41'E. Meanwhile, in terms of geographical location, Bengkulu Province in the north is bordered by West Sumatera Province, in the south bordering with Indonesia Ocean and Lampung Province, in the west by the Indonesian Ocean and to the east and Jambi Province and South Sumatra Province.

Bengkulu has shaken two powerful tectonic earthquakes in a relatively short period of time, i.e. in 2000 and 2007. On June 4, 2000, Bengkulu was shaken by a tectonic earthquake with a strength of 7.3 Ms. Then a major earthquake occurred in Bengkulu on 12 September 2007 with the strength of 7.9 Ms. The main earthquake following the aftershocks has caused heavy casualties, property, and destruction of public facilities. According to the history of Bengkulu Province has been shaken several times big earthquake such as on November 24, 1833 (VIII-IX MMI), August 18, 1938 (VII MMI),
August 18, 1871 (VI-VII MMI), June 26, 1914 (VII-VIII MMI), November 24, 1933 (VIII-IX MMI) and 15 December 1979 (VIII MMI) [9].

3. Methods
Data used in this research is secondary data obtained from Badan Meteorologi Klimatologi dan Geofisika (BMKG) Kepahiang Regency. The object of this research is Bengkulu Province which is astronomically located at 5º40'-2º0' S and 100º40'-104º0' E. The data used is the catalog of earthquake occurrences sourced from single-station BMKG Kepahiang, Bengkulu from January 1970 to December 2015.

Firstly, descriptive statistics of earthquake data in Bengkulu province will be presented. The goal is to see the variance of data and other general features. The next step is to apply the internal cluster validation method to the data to obtain the optimum number of clusters. Furthermore, the optimum number of clusters will be used as the number of initial clusters in the K-means cluster analysis. Then, compare the results of K-means cluster analysis to find out the optimum cluster of earthquake area in Bengkulu province.

4. Result and discussions

4.1. Description of earthquake data in Bengkulu Province
Bengkulu Province is one of the provinces located at the meeting of tectonic plates of IndoAustralia and Eurasia which is the main generator of high earthquake activity. Data of Badan Penanggulangan Bencana Daerah (BPBD) of Bengkulu Province shows that the epicenter of the earthquake that occurred in Bengkulu in 1970-2015 vary. The difference of earthquake epicenter and earthquake occurrence is caused by geology and seismicity in Bengkulu Province. Therefore, every seismic event has interconnectedness in space and time.

Figure 1. Distribution of earthquake events in Bengkulu Province 1970-2015

From 1970 to 2015, the largest magnitude earthquake occurred in 2007, at 7.9 Ms. The position of the epicenter was at -4.67°S and 101.13°E and the depth of 10 KM under the sea. The occurrence of an earthquake with magnitude of 4 Ms is the occurrence of earthquakes with minimum magnitude occurring since 1970 to 2015. Figure 1 shows that the epicenter of the earthquake that occurred in the province of Bengkulu mostly located in the sea area.
4.2. Internal cluster validation

Cluster analysis is a multivariate technique that aims to classify objects based on the similarity characteristics. K-means method is an algorithm for cluster n object based on attribute to k partition, where k < n. The determination of k on this method is based on the will of the researcher, so that the number of clusters is not necessarily the optimal number of clusters.

Internal cluster validation is a technique for determining the optimal number of clusters based on the prior information. The optimal number of clusters is determined by a given index. In this study, the internal cluster validation index used is Calinski-Harabasz (CH) index, Davies-Bouldin (DB) index, Silhouette (S) index, Dunn index (D), and S-Dbw (S-Dbw) index. Each index has certain criteria in determining the optimal number of clusters. The CH, S, and D indices require that the optimal number of clusters be determined by the largest index value. While the DB index and S-Dbw index using the smallest index value to determine the optimal number of clusters. These indices are used in determining the optimal number of clusters of earthquake event data in Bengkulu province from 1970 to 2015. Furthermore, R Program will be used in calculating the value of the internal cluster validation index.

Table 1. Optimum cluster number based on CH, S, DB, S-Dbw, and D

| k  | CH    | S     | DB    | S-Dbw | D     |
|----|-------|-------|-------|-------|-------|
| 2  | 1706.643 | 0.40665 | 0.90756 | 1.11447 | 0.00806 |
| 3  | 1419.745 | 0.31115 | 1.12526 | 3.17425 | 0.00881 |
| 4  | 1263.765 | 0.31846 | 1.12314 | 1.90366 | 0.01242 |
| 5  | 1127.464 | 0.29621 | 1.20120 | 5.08853 | 0.00819 |
| 6  | 1080.015 | 0.28789 | 0.61909 | 1.12314 | 0.00819 |
| 7  | 1033.560 | 0.28348 | 1.12808 | 5.72832 | 0.00362 |
| 8  | 978.760  | 0.27560 | 1.12096 | 15.0530 | 0.00927 |
| 9  | 953.066  | 0.27487 | 1.12200 | 9.63439 | 0.01167 |
| 10 | 915.266  | 0.26860 | 0.98330 | 13.7527 | 0.00851 |
| 11 | 890.888  | 0.26726 | 1.06585 | 35.3362 | 0.01111 |
| 12 | 863.364  | 0.26456 | 0.92488 | 23.6950 | 0.01203 |
| 13 | 831.544  | 0.25727 | 1.07730 | 12.9756 | 0.01508 |
| 14 | 815.435  | 0.26920 | 0.99723 | 6.67181 | 0.00860 |
| 15 | 797.447  | 0.26563 | 0.97879 | 24.6156 | 0.00502 |

Table 1 gives information that the optimal number of clusters based on the CH, S, and S-Dbw indices are two clusters (k = 2) with the index values being 1706.643, 0.40665, and 1.11447, respectively. All three indices require that the largest index value represents the optimal number of clusters. The smallest DB index of 0.61909 refers to the optimal cluster number of six clusters (k = 6). And the optimal number of clusters obtained based on index D is 13, because the largest value of index D is 0.01508. Thus, the earthquake events data in Bengkulu province from 1970 to 2015 can be cloned in 2, 6, and 13 clusters (k = 2, 6, 13).

4.3. K-means method for earthquake data in the Bengkulu Province with k = 2, 6, 13

K-Means Cluster Analysis is a non-hierarchical cluster analysis method that seeks to partition existing objects into one or more clusters of objects based on their characteristics, so that objects that share the same characteristics are grouped in the same cluster and objects that have the different characteristics grouped into another cluster. The purpose of clustering is to minimize the objective function set in the clustering process, which essentially seeks to minimize variation in one cluster and maximize the variation between clusters [10]. Clustering of earthquake data in Bengkulu Province using K-means method. The attributes used in this clustering are the epicenter (longitude and latitude) and magnitude.

Based on the index CH, S, and S-Dbw, the optimal cluster number is 2 (k = 2). The K-means method for k = 2 yields:
Table 2. K-means clustering for earthquake data with k = 2

| k   | Means of Longitude | Means of Latitude | Means of Magnitude | Sizes | SSE     |
|-----|--------------------|-------------------|-------------------|-------|---------|
| 1   | 102.67             | -4.99             | 4.91              | 1035  | 1874.81 |
| 2   | 100.62             | -3.10             | 4.99              | 725   | 1532.56 |

Table 2 shows that cluster 1 is around 102.670BT and -4.990LS. The average magnitude of cluster 1 is 4.91 Ms. Cluster 1 consists of 1035 members and the sum of squares of error is 1874.81. The 725 member of cluster 2 has an average magnitude of 4.99 Ms with its centroid of 100.62°E and -3.10°S.

The results of clustering the earthquake data in Bengkulu Province with k = 2 visualized using R program are as follows:

![Cluster Map](image)

Figure 2. K-means method for earthquake data in Bengkulu Province with k = 2

The optimal cluster number for earthquake data in Bengkulu Province based on DB index is 6 (k = 6). Furthermore, the optimal number of clusters will be applied to the K-means method as the initial number of groups. The clustering results using the R program are described as follows:

Table 3. K-means clustering for earthquake data with k = 6

| k   | Means of Longitude | Means of Latitude | Means of Magnitude | Sizes | SSE     |
|-----|--------------------|-------------------|-------------------|-------|---------|
| 1   | 100.58             | -3.45             | 4.98              | 332   | 380.77  |
| 2   | 102.76             | -5.04             | 5.01              | 505   | 356.35  |
| 3   | 101.98             | -3.84             | 4.76              | 397   | 333.49  |
| 4   | 104.49             | -5.85             | 5.04              | 137   | 168.14  |
| 5   | 101.29             | -5.28             | 4.64              | 195   | 207.46  |
| 6   | 99.85              | -1.86             | 5.30              | 194   | 195.51  |

Table 3 shows that cluster 1 has 332 members, an average magnitude of 4.98 Ms, and is around 100.58°E and -3.45°S. Cluster 2 has 505 members, a magnitude of 5.01 Ms, and is around 102.76°E and -5.04°S. Cluster 3 has 397 members, an average magnitude of 4.76 Ms, and is around 101.98°E and -3.84°S. Cluster 4 has 137 members, an average magnitude of 5.04 Ms, and is around 104.49°E and -5.85°S. Cluster 5 has 195 members, an average magnitude of 4.64 Ms, and is around 101.29°E
and -5.28°S. Cluster 6 has 194 members, an average magnitude of 5.3 Ms, and is around 99.85°E and -1.86°S. The result of earthquake data clustering in Bengkulu Province for k = 6 can also be seen in the following figure:

![Figure 3. K-means method for earthquake data in Bengkulu Province with k = 6](image)

Index D determines k = 13 as the optimal number of clusters in earthquake data in Bengkulu Province. The thirteen clusters are described by the R program as follows:

| k   | Means of Longitude | Means of Latitude | Means of Magnitude | Size | SSE  |
|-----|--------------------|-------------------|-------------------|------|------|
| 1   | 100.53             | -4.64             | 4.54              | 87   | 48.52|
| 2   | 101.06             | -3.71             | 5.81              | 109  | 84.60|
| 3   | 99.70              | -1.78             | 5.38              | 155  | 109.65|
| 4   | 99.82              | -3.49             | 4.95              | 106  | 77.51|
| 5   | 104.62             | -5.86             | 5.13              | 117  | 110.38|
| 6   | 102.07             | -4.69             | 5.38              | 198  | 89.80|
| 7   | 102.95             | -4.89             | 5.24              | 221  | 78.58|
| 8   | 101.56             | -5.96             | 4.62              | 84   | 86.46|
| 9   | 102.75             | -3.18             | 4.84              | 64   | 60.65|
| 10  | 102.94             | -5.46             | 4.51              | 165  | 68.01|
| 11  | 101.02             | -2.53             | 4.73              | 121  | 74.24|
| 12  | 102.12             | -4.39             | 4.36              | 180  | 55.37|
| 13  | 101.40             | -3.62             | 4.49              | 153  | 57.31|

Table 4 shows that cluster 7 has the most members, i.e. 221 members with an average magnitude of 5.24 Ms, and centered at 102.95°E and -4.89°S. While the number of cluster members in cluster 9 is 64 members. Cluster 9 is a cluster whose members are located around the centroid of 102.75°E and -3.18°S, and the average magnitude of this cluster is 4.84 Ms. The clustering results can also be seen in the following figure:
Based on the three clustering results, it was found that the number of clusters $k = 2$ resulted in considerable variation. It can be seen from the SSE value of each cluster 1 and cluster 2 which is quite big. The number of clusters $k = 13$ has a fairly low SSE value for each cluster. However, when viewed in Figure 4, the clusters formed tend to overlap. In other words, there are members of a cluster who are members of other clusters. The number of clusters $k = 6$ is the ideal number of clusters for earthquake data in Bengkulu Province from 1970 to 2015. This is shown in Figure 3, where the clusters formed have members of each cluster that are not crossed. Although the SSE value generated by 6 clusters is greater than the value of SSE generated by 13 clusters.

5. Conclusion

The conclusions obtained in this study are as follows: The Calinski-Harabasz index, the Silhouette index, and the S-Dbw index determine the optimal number of clusters for earthquake data in Bengkulu Province 1970 - 2015 for 2 clusters. Based on the Davies-Bouldin index, the optimal number of clusters for earthquake data in Bengkulu Province in 1970 - 2015 was 6 clusters. Based on the Dunn index, the optimal number of clusters for earthquake data in the Bengkulu Province from 1970 to 2015 were 13 clusters. The optimal number of optimal clusters for earthquake data in Bengkulu Province from 1970 to 2015 is 6 clusters. If the number of clusters is 2, then there will be considerable variation. Whereas if the number of clusters is 13, then members of a cluster will become members of other clusters.

References

[1] Mattjik AA and IM Sumertajaya 2011 IPBPress Bogor
[2] Liu Y, Li Z, Xiong H, Gao X, Wu J 2010IEEE International Conference on Data Mining DOI 10.1109/ICDM.2010.35
[3] Rendon E, Abundez I, Arizmendi A, Quiroz EM 2011 International Journal of Computers and Communications527
[4] Arbelaitz O, Ibai G, Javier M, Jesus MP, Inigo P 2013Pattern Recognition 46243
[5] Novianti P, Ulfasari R, and Dyah S 2016Laporan Hasil Penelitian Hibah Penelitian Pembinaan Universitas Bengkulu
[6] Weatherill and Burton PW 2009Geophysical Journal International176565
[7] EverittB and HothornT 2011An Introduction to Applied Multivariate Analysis with RSpringer Science + Business MediaUSA
[8] JohnsonR A and Wichern D W 2002 Applied Multivariate Statistical AnalysisFifthedition Pearson Education Inc USA
[9] BMG 2007 Database Gempa BMG Kepahiang BMG Kepahiang Bengkulu
[10] Ediyanto, Muhasah NM, Neva S 2013 Bimaster2 133