Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13246
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Determining The Optimal Number of K-Means Clusters Using The Calinski Harabasz Index and Krzanowski and Lai Index Methods for Groupsing Flood Prone Areas In North Sumatra

Ziana Syahputri[1)*], Sutarman[2)], Machrani Adi Putri Siregar[3)]
[1]State Islamic University of North Sumatra [2]University of North Sumatra, [3]State Islamic University of North Sumatra, Indonesia
[1)] ziana0703182096@uinsu.ac.id, [2)] sutarman@usu.ac.id, [3)] machraniadiputri@uinsu.ac.id

Submitted :Dec 16, 2023 | **Accepted** : Jan 12, 2024 | **Published** : Jan 14, 2024

Abstract: The k-means algorithm is a partitional clustering method. K-Means has several advantages, including being easy to implement, having a high level of convergence and producing denser clusters. Meanwhile, the drawback is that it is difficult to determine the optimal number of clusters. The K-Means method will be used to solve problems in areas prone to flood disasters in North Sumatra. This research aims to find the optimal number of clusters with the Calinski Harabasz Index and Krzanowski And Lai Index based on the Cluster Tightness Measure (CTM) value. There are eleven variables used in this research. Based on the research results, it was concluded that the CTM CH result of 0.376 was smaller than the CTM KL of 0.7843. So it can be said that determining the optimal number of clusters using CH with k = 6 is better than KL with k = 2.

Keywords: Cluster, K-Means, CH Index, KL Index, Cluster Tightness Measure (CTM), Flood.

INTRODUCTION

Finding new patterns in enormous data sets is a technique known as data mining. Data mining combines database systems, statistics, machine learning, and artificial intelligence techniques. Since data quantities are multiplying through conventional means and are becoming more complicated and varied, data mining has become increasingly necessary in recent years. Big Data, as it is now defined, refers to the massive amounts of organized or unstructured data that are currently flooding the business and other industries. Naturally, if you try to interpret and understand the patterns and relationships in Big the manually, you will find it challenging. (Suyanto, 2019).

One of the methods in the data mining process is cluster analysis, which looks for homogeneous items and divides them into groupings known as clusters (Ni'matuzzahroh et al., 2022). The process of grouping a data set into many clusters so that the characteristics of one cluster are the same while the characteristics of other groupings are different is called clustering. Partition clustering, a technique that divides a data set into several mutually exclusive groups with an initial number of randomly generated clusters, includes the K-Means algorithm. K-means has many advantages over hierarchical approaches, such as being easier to use, showing a high degree of convergence, and producing denser clusters. The K-Means approach is prone to outliers because it uses average values to identify cluster centers, which is one of its weaknesses. Another reason is the difficulty of determining the ideal number of clusters. By combining the Krzanowski and Lai (KL) Index and Calinski Harabasz (CH) Index techniques, this deficiency can be corrected.

In reality, a lot of K-Means users find it difficult to decide on the ideal number of clusters as a result of this flaw. As a result, numerous researchers have developed a Validity Index, such as the Silhouette Index, KL Index, Davies and Bouldin (DB) Index, Calinski and Harabasz (CH) Index, Dunn Index, Gap Index, and others, to ascertain the number of clusters in a data set.

Research results (Saitta et al., 2008) show that the Dunn Index approach is sensitive to noise. Random errors or data changes in the measured variable are referred to as noise Three indices—SF, MB, and CH—perform well

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13246
e-ISSN : 2541-2019
p-ISSN : 2541-044X

when dealing with overlapping clusters. The GE and SF approaches effectively manage noise; however, in most datasets, it is often seen that GE overestimates the true number of clusters. DB, GE, SF, and CH can all estimate five clusters in the case of sub-cluster hierarchy; however, CH and SF usually provide the best results. According to research (Sikana & Wijayanto, 2021) the best technique for finding out the ideal number of clusters in K-Means is the SI, CH and DB Index approaches.

Therefore, by incorporating the Calinski Harabasz (CH) Index and Krzanowski and Lai (KL) Index approaches, the weaknesses of the k-means method can be overcome. Both approaches are included in the Validity Index. When comparing the two indices, choose Cluster Tightness Measure (CTM) as the optimal measurement approach if you know the ideal cluster results. In North Sumatra, groups of areas prone to flood disasters will be used to determine the ideal number of clusters using CH and KL. In this case, flood tragedies occur quite frequently so that it is necessary to group flood-prone areas for mitigation measures.

LITERATURE REVIEW

The results of previous research explained that the Dunn Index method is sensitive to noise. CH, MB and SF are one of the three indexes that are successful when dealing with overlapping clusters. The GE and SF methods successfully handle noise but GE is often found to overestimate the actual number of clusters in most data sets. In the case of a sub-cluster hierarchy, CH, DB, GE, and SF can estimate 5 clusters, and in general, CH and SF provide the best results (Saitta et al., 2008).

Research (Sikana & Wijayanto, 2021) also explains that the SI, CH and DB Index methods are the best methods for determining the optimum number of clusters in K-Means.

In research (Charrad et al., 2014) provides a solution that in determining the optimal number of clusters there are 30 indices given in the research, one of which is the Calinski Harabasz (CH) Index and the Krzanowski Lai (KL) Index. According to his research, both methods have an optimal number of clusters if the index value is maximum. This index can be used to determine the number of clusters using any method, including K-Means.

Research that discusses flood disasters includes research conducted (Azizah et al., 2021). This research uses 9 variables, namely the number of RTs and RWs affected by flooding, the minimum water level during a flood, the maximum water level during a flood, the duration of inundation during a flood, the number of refugee camps for flood victims, the number of evacuees during a flood, the number of victims who died during a flood, and the number refugee camp. This research produced 3 clusters, namely sub-districts with high levels of vulnerability, sub-districts with medium vulnerability and sub-districts with low vulnerability.

METHOD

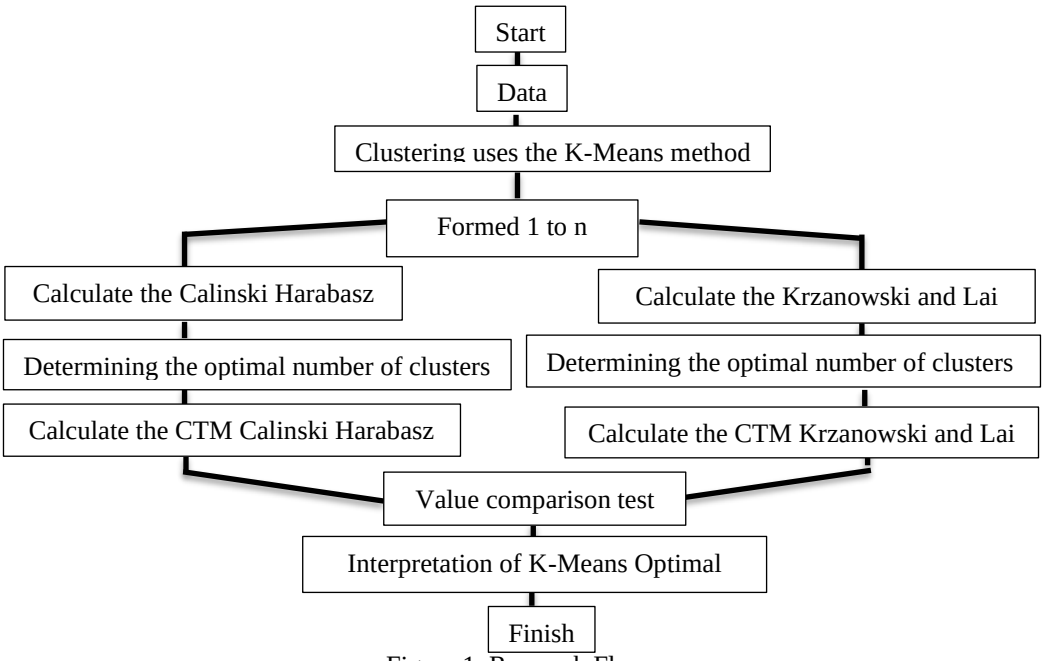

Figure 1. Research Flow

This research was carried out at the North Sumatra Regional Disaster Management Agency (BPBD), precisely on Jalan Medan-Binjai, KM 10, 3 No. 8 Medan Krio, Sunggal District, Deli Serdang Regency, North Sumatra. Time This research was conducted from January – May 2023.

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13246
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Based on the level, this type of research is applied research because this research is the development of knowledge (theory) used to solve problems. Based on the specific objectives of this research, descriptive research is used because after the case of this research is solved, the results of the research will be presented. And this research uses a quantitative approach because this research uses systematic calculations. The data used in this research is secondary data. The data used is data on the number of flood events, victim data and damage data from January 2020 to December 2020 obtained from the North Sumatra Regional Disaster Management Agency.

The variables that will be used in this research are indicators related to the impact of flood disasters, namely the number of flood events (X_1), the number of victims who died (X_2), the number of missing victims (X_3), the number of injured victims (X_4), the number of victims suffering (X_5), the number of victims displaced (X_6), the amount of damage to the house (X_7), the amount of damage to educational facilities (X_8), the amount of damage to health facilities (X_9), the amount of damage to worship (X_{10}), the amount of damage to other buildings (X_{11}).

This research started by searching for several literatures that discussed natural flood disasters as well as the methods used in analyzing flood disasters which were then used as references. After the literature has been collected, the next stage is to collect data that will be used in research. After the data is obtained and collected, the next step is data analysis and processing.

The K-means method used in this research combines the Calinski-Harabasz Index and Krzanowski and Lai Index methods to determine the optimal number of clusters. The results of this calculation method are in the form of clustering which determines the groups based on the results of the closest distance to each district/city in North Sumatra. The research design is as follows:

1. Data Exploration

   The first stage in the clustering process is data exploration. Data exploration is the process of understanding the data to be analyzed. By exploring the data first, it can be determined which technique to use. At this stage, the data obtained is data whose class label has not been determined, so the data is appropriate to use in the K-Means method calculations.

2. K-Means Clustering

   Because K-Means relies on determining the first centroid value to determine the initial number of groups, it is a partitioning algorithm. This approach attempts to organize data in such a way that information that has similar qualities is collected into one group and information that has different characteristics is collected into different groups (Ulinnuha & Sholihah, 2021). In other words, the K-Means technique attempts to minimize the objective function set during the clustering process by maximizing the variation in the number of groups represented by the k variables and reducing the variance between data in a cluster. The number of clusters required is represented by the variable k. Data without class labels is accepted as input for this approach (Marisa et al., 2021).

   Before calculating the optimal number of clusters, the clustering is first calculated using the K-Means method, the calculation is as follows.

   1. Enter the data that will be clustered
   2. Determine the number of *k* or the number of cluster. The *k* value that will be tried use $k = 2, 3, 4, 5, 6$. For this research, the *k* value starts from *k* = 3 after completing the calculation *k* = 3.
   3. Determine the initial centroid randomly
   4. Calculate the Euclidean distance using the following equation:

$$d_{(x,y)} = \sqrt{\sum_{i=1}^{n}(x_{ij} - y_{ij})^2}\,; i = 1,2,3,\cdots,n \quad (1)$$

Information:
$d_{(x,y)}$ = Distance of data x to the center of cluster y
x_{ij} = Data x in the observation i
y_{ij} = Center Point of y observation i
n = Number of observation

   5. Recalculate the centroid with cluster membership formed by calculating the average value of all data in a cluster using the equation (2).

$$C_{ij} = \frac{\sum_{i=1}^{p} x_{ij}}{P} \quad (2)$$

Information:
C_{ij} = newest center point in the next iteration

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13246
e-ISSN : 2541-2019
p-ISSN : 2541-044X

x_{ij} = cluster member
p = the number of cluster

6. Recalculate an object using the new centroid. If the cluster members do not experience any further changes, then the clustering process is declared complete.

(Ulinnuha & Sholihah, 2021)

2. K-Means Clustering using Calinski Harabasz Index method

If the cluster results have been formed then proceed with determining the optimal number of clusters using the Calinski Harabasz Index method with equation (3).

$$CH = \frac{trace\ (SSB)}{trace\ (SSB)} \times \frac{n-k}{k-1} \quad (3)$$

Index the centroid result from the last iteration are needed to calculate the *Sum of Square Within Cluster* (SSW) value in equation (4) and the last centroid to calculate the sum if square between *Sum of Square Between* (SSB value in equation (5). In CH Index the centroid result from the last iteration is needed to calculate the *Sum of Square Within Cluster* (SSW) value in equation (4) and the last centroid to calculate the sum if square between *Sum of Square Between* (SSB) value in equation (5).

$$SSW = \sum_{i=1}^{k} \sum_{\substack{j=1 \\ x_j \in w_i}}^{n_i} \|x_j - v_i\|^2 \quad (4)$$

$$SSB = \sum_{i=1}^{k} n_i \|v_i - \mu_{data}\|^2 \quad (5)$$

$$\mu_{data} = \frac{1}{N} \sum_{i=1}^{N} x_i \quad (6)$$

Information:
x_j = data x on the jth observation
v_i = vth center point jth observation
n_i = number of objects in cluster v
μ_{data} = average global centroid data

(Brito Da Silva et al., 2020)

3. K-Means Clustering using Krzanowski and Lai *Index* Method

Next, the calculation of the optimal number of clusters will be continued usingKrzanowski and Lai *Index* with equation (8). In equation (8) equation (7) will be calculate first. To find the value of equation (7) you will first look for the compactness value (W_k dan $W_{(k-1)}$) equation (5). Then divide by $DIFF_{(k+1)}$.

$$DIFF(k) = \left[(k-1)^{\frac{2}{p}} w(k-1) - k^{\frac{2}{p}} w(k)\right] \quad (7)$$

$$KL(k) = \frac{DIFF(k)}{DIFF(k+1)} \quad (8)$$

(Charrad et al., 2014).

4. Comparison CTM CH *Index* and KL *Index*

If the results of determining the optimal number of clusters have been formed, the calculation of the CTM CH Index and CTM Index will continue with equation 9.

$$CTM = \frac{1}{K} \sum_{t=1}^{k} \left(\frac{1}{P} \sum_{m=1}^{P} \frac{S_{stm}}{S_m} \right) \quad (9)$$

Information:
S_{tm} = t-th group standard deviation for the m-th variable
S_m = standard deviation of all data for the mth variable

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13246
e-ISSN : 2541-2019
p-ISSN : 2541-044X

K = many groups
P = many variables

(Madani, 2014)

5. After carrying out a comparison test, an interpretation of the optimal number of k-means clusters was carried out.

RESULT

The research results of the data that will be analyzed are sourced from BPBD North Sumatra, namely data on the impact of the 2020 flood disaster which will be used to find the optimal number of clusters with the help of Microsoft Excel and R. The data summary can be seen in the following image.

Table 1. Descriptive Statistics

Variable	N	Min	Max	Mean	Std. Deviation
Number of Flood events (X_1)	19	1	13	4,00	3,528
Number of Death Victims (X_2)	19	0	6	0,68	1,887
Number of Missing Victims (X_3)	19	0	3	0,21	0,713
Number of Injured Victims (X_4)	19	0	9	0,47	2,065
Number of Victims Suffering (X_5)	19	0	116939	17971,05	29898,781
Number of Victims Displaced (X_6)	19	0	10982	1194,32	3127,427
Number of House Damages (X_7)	19	0	218	20,00	52,819
Number of Damages to Educational Facilities (X_8)	19	0	8	0,95	2,094
Number of Damages to Health Facilities (X_9)	19	0	1	0,05	0,229
Amount of Worship Damage (X_{10})	19	0	7	0,58	1,710
Total Damage to Other Buildings (X_{11})	19	0	51	3,16	11,673
Valid N (*listwise*)	19				

source: Output Spss

Based on Table 1, it explains the descriptive statistics of the variables contained in the North Sumatra Province flood data used by researchers. In column N the eleventh variable has 19 data, meaning it can be seen that North Sumatra Province which was affected by the flood has 19 Regencies/Cities. In the variable number of flood events in 2020, the lowest (minimum) was 1 affected, while the highest was 13. Meanwhile in the variable of victims who died due to floods in 2020, the lowest was 0 and the highest was 6, namely in Medan City and Deli Serdang Regency. In Table 1 it can be seen that of the eleven variables, the fifth variable, namely the number of victims suffering, has the highest number of cases compared to the other variables. In the variable number of victims who suffered, there were 116,939 victims affected. The victim suffering variable has a greater average than the other variables, this shows that the victim suffered more than the other variables due to the flood disaster in North Sumatra. Then, looking at the standard deviation value, the number of victims suffering was higher, namely 29,898 victims, while the standard deviation was small in the health facility damage variable, namely 0 damage. This can be interpreted as meaning that the numbers for each variable are not evenly distributed in each Regency/City in North Sumatra.

K-Means

In the first experiment, k = 3 grouping will be carried out. The first step in grouping using the K-Means method requires finding the initial centroid. Determining the initial centroid value is done using data taken randomly from all data. After the results of the first experiment there are results then proceed to experiments k = 4 to k = 6.

Table 2. Initial Centroid

Centroid (C)	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
C1	13	0	0	0	17.630	65	0	0	0	0	0
C2	1	0	0	0	56.083	0	0	0	0	0	0
C3	3	0	0	0	303	0	9	0	0	0	0

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13246
e-ISSN : 2541-2019
p-ISSN : 2541-044X

The second step is to calculate the distance from the centroid using the Euclidean distance formula in equation 1.

Calculation of Euclidean distance from data 1 to data 19 with centroid 1 (Batu Bara Regency):

$$d_{(1,1)} = \sqrt{\sum_{i=0}^{n}(x_i - y_i)}$$

$$d_{(1,1)} = \sqrt{\begin{matrix}(4-13)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (490-17630)^2 + (0-65)^2 + (5-0)^2 + \\ (1-0)^2 + (0-0)^2 + (1-0)^2 + (6-0)^2\end{matrix}}$$

$$d_{(1,1)} = 17.140$$

$$d_{(2,1)} = \sqrt{\sum_{i=0}^{n}(x_i - y_i)}$$

$$d_{(2,1)} = \sqrt{\begin{matrix}(1-13)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (710-17630)^2 + (0-65)^2 + (0-0)^2 + \\ (0-0)^2 + (0-0)^2 + (0-0)^2 + (1-0)^2\end{matrix}}$$

$$d_{(2,1)} = 16.920$$

$$d_{(1,1)} = \sqrt{\sum_{i=0}^{n}(x_i - y_i)}$$

$$d_{(1,1)} = \sqrt{\begin{matrix}(3-13)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (303-17630)^2 + (0-65)^2 + (0-0)^2 + \\ (9-0)^2 + (0-0)^2 + (0-0)^2 +\end{matrix}}$$

$$d_{(1,1)} = 17.327$$

Then do the same thing to calculate the euclidean distance from data 1 to data 19 with centroid 2 (Tanjung Balai city) and centroid (Padang Sidimpuan city).

After calculating the Euclidean distance of each data to the three centroids. Next, the minimum distance between the data and one of the three centroids is selected. The next step is that the data is allocated to the nearest centroid as a cluster result. Table 3 is the result of calculating the Euclidean distance for iteration 1.

Table 3. Results of The 1st Iteration of Cluster Center Distance Calculations

Regency/City	C1	C2	C3	Shortest Distance	Group
Mandailing Natal	17140	55593	187	187	C3
Tapanuli Selatan	16920	55373	407	407	C3
Labuhan Batu	13130	51583	4197	4197	C 3
Sharpening	1471	36983	18797	1471	C 1
Karo	17345	55798	51	51	C 3
Deli Serdang	17257	55693	1036	1036	C 3
Step up	6571	31883	23897	6571	C 1
South Nias	16920	55373	407	407	C 3
Serdang Bedagai	30984	7566	48307	7566	C 2
Coal	0	38453	17327	0	C 1
North Labuhan Batu	11705	50158	5623	5623	C 3
North Nias	17630	56083	303	303	C 3
West Nias	17630	56083	306	306	C 3
Cape City Hall	38453	0	55780	0	C 2
Pematang Siantar City	17630	56083	303	303	C 3
Tebing Tinggi City	23718	14750	41044	14750	C1
Medan City	99705	61510	116979	61510	C2
Binjai City	17280	52998	11665	11665	C3
Padang Sidimpuan City	17327	55780	0	0	C3

Based on table 3, it can be seen that the 1st data (Mandailing Natal Regency) is included in the 3rd cluster, because the 1st data has the minimum or shortest distance to centroid 3, this is applied to all data.

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13246
e-ISSN : 2541-2019
p-ISSN : 2541-044X

The third step determines the newest centroid using equation 2. The newest centroid is determined by calculating the average of each cluster. then calculate the Euclidean distance using equation 1. It turns out that in the 2nd iteration there was 1 data that experienced a cluster shift, namely the city of Tinggi Cliff because in the 1st iteration, the city of Tinggi Cliff was included in the 2nd cluster (C2) and moved to cluster 1 (C1). then the next iteration must be carried out until the data does not move to another cluster, namely by averaging each cluster to determine the next centroid. here is the latest centroid to calculate the 3rd iteration.

Table 4. Latest Centroid For Iteration 3

Centroid	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
C1	8	0	0	0	25568	154	0	1	0	1	13
C2	4	2	0	0	73873	3348	0	0	0	0	1
C3	3	1	0	1	1463	1003	32	1	0	1	1

If you have recalculated the latest centroid for the 3rd iteration as in table 4 then the distance calculation will be carried out for the 3rd iteration. The results can be seen as in table 5 below.

Table 5. The 3rd Iteration.

Regency/City	C1	C2	C3	Iteration 3	Iteration 2
Mandailing Natal	25079	73459	1398	C3	C3 _
Tapanuli Selatan	24859	73240	1254	C3 _	C3 _
Labuhan Batu	21069	69454	3198	C3 _	C3 _
Asahan	6470	54875	17665	C1 _	C1 _
Karo	25284	73664	1547	C3 _	C3 _
Deli Serdang	25184	73511	1079	C 3	C 3
Step up	1377	49786	22759	C 1	C 1
South Nias	24859	73240	1254	C 3	C 3
Serdang Bedagai	23048	25376	47134	C 2	C 2
Coal	7939	56339	16194	C 1	C 1
North Labuhan Batu	19644	68031	4573	C 3	C 3
North Nias	25569	73949	1774	C 3	C 3
West Nias	25569	73947	1752	C 3	C 3
Cape City Hall	30515	18102	54629	C 2	C 2
Pematang Siantar City	25569	73949	1774	C 3	C 3
Tebing Tinggi City	15780	32650	39882	C 1	C 1
Medan City	91793	43428	115749	C 2	C 2
Binjai City	23924	70055	10357	C3 _	C3 _
Padang S idempuan City	25266	73646	1534	C3 _	C3 _

The final cluster results in Table 5 are cluster results that have not changed or are the same as the cluster results in the previous iteration. Because in the last iteration there were no clusters that moved to another cluster, it can be concluded that the members of Cluster 1 are Asahan, Langkat, Batu Bara, and Tebing Tinggi. *Cluster 2* is Serdang Bedagai, Tanjung Balai, Medan. *Cluster 3* is Mandailing Natal, Tapanuli Selatan, Labuhan Batu, Karo, Deli Serdang Nias Selatan, Labuhan Batu Utara, Nias Utara, Nias Barat, Pematang Siantar, Binjai, and Padang Sidimpuan.

K-Means Clutering Using Calinski Harabasz Index Method

The assessment of cluster results is given by theCalinski Harabasz (CH) Index based on a comparison of the values of sum of square between cluster (SSB) as separation and sum of square within cluster (SSW) as compactness which are multiplied by a normalization factor, namely the difference between the amount of data and the number of clusters and divided by the number of clusters minus one. The greater the Calinski-Harabasz (CH) Index, the better the cluster results (Saidah et al., 2022).

The Calinski-Harabasz (CH) Index equation can be seen in Equation 3. The optimization value with a larger CH value criterion indicates a better (maximum) solution grouping. The calculation results of the Calinski-Harabasz (CH) Index are as follows.

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13246
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 6. Calinski Harabasz Index Value

The Number of Cluster	Calinski Harabasz Index
2	41,90659197
3	30,59084208
4	66,7947015
5	248,2641389
6	309,21495

Table 6 is the result of calculating cluster values from the Calinski Harabasz Index. The calculation results for each cluster produce CH values that are not the same. In the CH Index the optimal number of clusters is indicated by the CH Index value which has the most optimum value. In Table 6 it is identified that the optimum number of clusters obtained is 6 because in the 6th cluster the CH value is 309.21495. So the optimum number of clusters obtained using the CH Index is 6 clusters.

K-Means Clustering Using Krzanowski and Lai Index Method

According to the Krzanowski-Lai (1988) Index KL is based on decreasing values of square numbers in a cluster (Achmad & Fernandes, 2021), which can be defined by equation 4. The p value shows the number of variables. The k value is optimal if it maximizes KL (k) (Charrad et al., 2014). The Krzanowski and Lai Index can be seen in table 7 below.

Table 7. Krzanowski And Lai Index Value

The Number of Cluster	**Krzanowski and Lai Index**
2	9,08517383
3	0,454148177
4	2,548998295
5	4,560819729
6	-51,08766963

Table 7 is the result of calculating cluster values from the Krzanowski and Lai Index. The calculation results for each cluster produce different KL values. In the CH Index the optimal number of clusters is shown by the KL Index value which has the most optimum value. In Table 7 it is identified that the optimum number of clusters obtained is 2 because in the 2nd cluster the KL value is obtained. So the optimum number of clusters obtained using the KL Index is 2 clusters.

Comparison of CTM Values

According to Epss and Ambikairajah (2004) in research (Madani, 2014) stated that to measure the level of goodness of optimal grouping an algorithm can use CTM benchmarking, because it is simpler and uses a measure based on the standard deviation of several groups with several variables. The equation can be seen in equation 5 above.

The results of comparing the optimal number of clusters with Calinski Harabasz and Krzanowski and Lai can be seen in Table 8

Table 8. Calinski Harabasz and Krzanowski and Lai CTM Values

Method	**CTM Value**
Calinski Harabasz	0,376
Krzanowski and Lai	0,784

Based on Table 8 of the CTM values, it can be concluded that the Calinski Harabasz Index is better in determining the optimal number of clusters because the CTM value of the Calinski Harabasz Index is smaller than the Krzanowski and Lai Index in the K-Means method cluster analysis. So from the CTM results it is found that the CH Index is the optimal one to use with the optimal number of clusters being 6 clusters, then these 6 clusters will be used for clustering results.

From the analysis results, the CH Index always produces the maximum number of clusters. This is in accordance with the CH Index formula which is the ratio between SSB and SSW, where the SSB value tends to

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

e-ISSN : 2541-2019
p-ISSN : 2541-044X

increase for data that is not clearly clustered and the SSW value always decreases. This causes the more clusters that are formed, the greater the CH Index value and the CH Index tends to always produce many clusters. According to (Milligan & Cooper, 1985), the CH Index can only be used for data that is clearly clustered. However, according to (Heer & Chi, 2002), the CH Index is unable to determine the appropriate number of clusters for data that is not clearly clustered. Because the data in this study clustered clearly, the CH Index produced very good clusters.

DISCUSSIONS

After determining the optimal number of clusters, namely k = 6, the cluster members can be determined as in table 9 below.

Table 9. Number of Cluster Members at k = 6

Group	Number of Cluster	Regency/City
1	1	Binjai City
2	2	Labuhan Batu Selatan, Labuhan Batu Utara
3	3	Sel Be ng Balai City , Tel Tinggi City
4	1	Mel and City
5	3	Asahan, Langkat, Batu Bara
6	9	Mandailing Natal, Tapanu li , Karo, Del Sel , Nias , Nias Ul , West Nias , Pe mature Siantar, Padang Sidimpu City .

To find out the characteristics of the formation of flood vulnerability clusters in North Sumatra, you can look at the average of each cluster. For cluster 1 the average is 1,384.0, cluster 2 the average is 478.9, cluster 3 the average is 4,475.4, cluster 4 the average is 11,445.1, cluster 5 the average is 1,850.9, cluster 6 the average the average is 43.4. In this case, it can be concluded that the city of Medan is in cluster 4, which is a very vulnerable zone, therefore the city area must be extra vigilant when the rainy season has entered. The Serdang Bedagai area, Tanjung Balai City, Tebing Tinggi City in cluster 3 are vulnerable zones. The Asahan, Langkat, Batu Bara areas are quite vulnerable zones. The Binjai City area is a less vulnerable zone. The non-vulnerable areas are Labuhan Batu, North Labuhan Batu, while the areas of Mandailing Natal, South Tapanuli, Karo, Deli Serdang, South Nias, North Nias, West Nias Pematang Siantar City, Padang Sidimpuan City are very non-vulnerable zones.

CONCLUSION

The test results for determining the optimal number of K-Means clusters with the Calinski Harabasz Index showed that the optimal number of clusters was 6 clusters with a Calinski Harabasz (CH) value of 309.21495, while the Krzanowski And Lai Index found that the optimal number of clusters was 2 clusters with a Krzanowski And Lai (KL) Index is 9.0851.

Based on the results of the Cluster Tightness Measure (CTM) comparison. The CTM CH result of 0.376 is smaller than the CTM KL of 0.7843. So it can be said that determining the optimal number of clusters using CH is better than KL.

REFERENCES

Azizah, Oscarini, D. R., Saputra, F. M., & Multazam, H. (2021). *Pengelompokan Kecamatan di Jakarta Berdasarkan Tingkat Kerawanannya Terhadap Banjir Menggnakan K-Means Clustering*. *3*, 150–159.

Brito Da Silva, L. E., Melton, N. M., & Wunsch, D. C. (2020). Incremental Cluster Validity Indices for Online Learning of Hard Partitions: Extensions and Comparative Study. *IEEE Access*, *8*, 22025–22047.

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). Nbclust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, *61*(6), 1–36.

Fernandes, A. A. R. (2021). Comparison of Cluster and Linkage Validity Indices in Integrated Cluster Analysis with Structural Equation Modeling War-PLS Approach. *Journal of Hunan University (Natural Sciences)*, *48*(4).

Heer, J., & Chi, E. H. (2002). Mining the Structure of User Activity using Cluster Stability. *Proceedings of the Workshop on Web Analytics SIAM Conference on Data Mining*, *February*.

Madani, B. J. (2014). *Analisis Hybrid Hirerchical Clustering Melalui Mutual Cluster, Bottom-Up dan Top Down Menggunakan Jarak Euclidean dan Mahalanobis*.

Marisa, F., Maukar, A. L., & Akhriza, T. M. (2021). *Data Mining Konsep dan Penerapannya*. Budi Utama.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, *50*(2).

Ni'matuzzahroh, L., Andrea Tri Rian, D., & Adrianingsih, N. Y. (2022). Clustering Regencies / Cities in

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Kalimantan Island Based on Poverty Indicators using Agglomerative Hierarchical Clustering (AHC). *Jurnal Matematika, Statistika, Dan Komputasi*, *19*(1), 79–89.

Saidah, D. A., Santoso, R., & Widiharih, T. (2022). Pengelompokan Provinsi Di Indonesia Berdasarkan Indikator Kesehatan Lingkungan Menggunakan Metode Partitioning Around Medoids Dengan Validasi Indeks Internal. *Jurnal Gaussian*, *11*(2), 302–312.

Saitta, S., Raphael, B., & Smith, I. F. C. (2008). A comprehensive validity index for clustering. *Intelligent Data Analysis*, *12*(6), 529–548.

Sikana, A. M., & Wijayanto, A. W. (2021). Analisis Perbandingan Pengelompokan Indeks Pembangunan Manusia Indonesia Tahun 2019 dengan Metode Partitioning dan Hierarchical Clustering. *Jurnal Ilmu Komputer*, *14*(2), 66.

Suyanto. (2019). *DATA MINING; Untuk KLasifikasi dan Klasterisasi Data*. Informatika Bandung.

Ulinnuha, N., & Sholihah, S. A. (2021). Analisis Cluster Untuk Pemetaan Data Kasus Covid - 19 Di Indonesia Menggunakan K - Means. *Jurnal MSA (Matematika Dan Statistika Serta Aplikasinya)*, *9*(2).

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.