BAREKENG: Journal of Mathematics and Its Applications
December 2024 Volume 18 Issue 4 Page 2283–2292
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol18iss4pp2283-2292

# APPLICATION OF K-MEANS++ WITH DUNN INDEX VALIDATION OF GROUPING WEST KALIMANTAN REGION BASED ON CRIME VULNERABILITY

**Rifkah Alfiyyah Sary[1], Neva Satyahadewi[2*], Wirda Andani[3]**

[1,2,3]*Statistics Study Program, Faculty of Mathematics and Natural Sciences, Universitas Tanjungpura*
*Jl. Prof. Dr. H. Hadari Nawawi, Pontianak, 78124, Indonesia*

*Corresponding author's e-mail: * **neva.satya@math.untan.ac.id***

***Article History:***

*Received: 21st February 2024*
*Revised: 19th May 2024*
*Accepted: 15th July 2024*
*Published: 14th October 2024*

***Keywords:***

*Euclidean;*
*Non-Hierarchical;*
*VIF*
.

*ABSTRACT*

*Crime is an unlawful behavior that will be given a punishment or sanctions based on Kitab Undang-Undang Hukum Pidana (KUHP) or other regulations in Indonesia. One of the provinces in Indonesia, namely West Kalimantan reported that criminal cases are increasing in 2021 and 2022. One of the solutions to minimize that case is grouping the district and city in West Kalimantan based on the level of vulnerability so the authority can be more responsive in solving these problems. The grouping can be done by cluster analysis. This analysis aims to group some objects based on the similarity of characteristics. K-Means++ is one of the methods of cluster analysis. K-Means++ is the development of K-Means, in which K-Means++ is smarter than K-Means in selecting the initial centroid because only one initial centroid is chosen randomly, and the initial centroids of the other clusters are done through calculations. This research uses secondary data from BPS of West Kalimantan, consisting of 10 variables. This research aims to form clusters to determine the level of vulnerability of each district and city in West Kalimantan. The selection of the optimal cluster is done by evaluating the cluster. One of these evaluations is the Dunn Index. Based on the analysis results, the optimum number of clusters is $k = 3$ with a Dunn Index value of 0.55. The first cluster is categorized as non-vulnerable with ten members, the second cluster as vulnerable with three members, and the third cluster as very vulnerable with one member.*



***How to cite this article:***

R. A. Sary, N. Satyahadewi and W. Andani., "APPLICATION OF K-MEANS++ WITH DUNN INDEX VALIDATION OF GROUPING WEST KALIMANTAN REGION BASED ON CRIME VULNERABILITY," *BAREKENG: J. Math. & App.*, vol. 18, iss. 4, pp. 2283-2292, December, 2024.


Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

**Research Article · Open Access**

## 1. INTRODUCTION

In general, criminality is an act that contradicts the values and norms that have been legalized by law and can harm certain parties. Criminality can include immoral acts, theft, pickpocketing, mugging, murder, drug abuse, fraud, and corruption. A person who commits a crime may be affected by several factors, including economic factors. These factors can be in the form of population density. The high population can cause many problems that arise in the area, including limited employment opportunities. This condition causes unemployment to be higher, which also has the potential to increase the number of poor people and economic inequality in the local community so that it is not impossible for people to commit criminal acts to meet their needs [1]. In addition to economic factors, educational factors also cause a person to commit a crime. A person with more excellent knowledge will consider the risks of his actions to avoid things that violate the law. Higher education that a person successfully pursues is also an opportunity to get a job to fulfill their life. Besides these factors, according to Badan Pusat Statistik (BPS), there are some indicators to measure the seriousness of the crime. Some indicators are crime total, crime rate, crime clock, and crime clearance.

Based on data published by the National Criminal Information Center, the number of crime cases for 2020 to 2022 in Indonesia has increased. In 2020, there were 188648 cases, and in 2021 there were 275258. Also 2022, the cases increased again to 322200 [2]. One of the provinces in Indonesia that has experienced an increase in cases is West Kalimantan. BPS of West Kalimantan stated that the number of reported cases in 2020 was 3470 cases. In 2021, there were 3622 cases, and in 2022, there were 4100 cases [3].

This increasingly widespread criminality, if not immediately resolved, will provide a frightening specter to residents in society. The government and other authorities need to find solutions by grouping districts/cities in West Kalimantan based on indicators and crime factors to determine each region's vulnerability level so that related parties can be more responsive in monitoring and overcoming these problems. Analysis to grouping the districts/cities can be done with multivariate analysis. One of the analysis techniques in multivariate analysis is cluster analysis, which aims to group objects into one group that tends to have similar characteristics. Methods in cluster analysis are divided into two, namely hierarchical and non-hierarchical methods. These two methods have differences in the initial stage of grouping objects. The hierarchical method begins by grouping objects with the closest similarity so that the final result will form a tree with precise levels between objects. Meanwhile, the non-hierarchical method starts with determining the number of clusters to be created and then placing objects into clusters [4].

One of the non-hierarchical methods that is often used is K-Means. However, this method has a weakness in the selection of the initial centroid, which is random for each cluster. Based on that reason, the K-means method was then developed and refined into another method. the method in question is K-Means++, which is developed in such a way that the selection of centroids is done more intelligently because the selection of the initial centroid is done through calculations to improve the quality of clustering. [5]. However, the drawback of this non-hierarchical method is that it is difficult to determine the optimal number of clusters, so cluster evaluation, known as index validation of a cluster, is needed. Dunn Index is one method of validation where this validation can calculate the best score in producing clusters that have high similarity within a cluster and low similarity between clusters [6].

There have been many previous studies related to the K-Means++ method. A study was done to group the regions in Java Island according to COVID-19 data by comparing the K-Means and K-Means++ methods [7]. This research aims to form a zone cluster based on positive patient data, the number of recipients of the first dose of vaccine, and the number of recipients of the second dose of vaccine in Java Island. The number of clusters analyzed was $k = 3$ to $k = 5$ and validated using the Silhouette Coefficient. The results of this study show that the K-Means++ method produces better clusters when the value of $k$ is larger, namely $k = 5$. Meanwhile, K-Means will be better if the value of $k$ used is smaller, namely when $k = 4$.

The other research was conducted in Bangkalan Regency in mapping crime using the K-Means and K-Means++ methods [5]. This research uses a dataset of types of crime in 2021 from 18 sub-districts in Bangkalan Regency. The method used in determining the optimum cluster is the Elbow method. Then, the clusters formed for each crime type are tested for quality using the Silhouette Coefficient to determine the best method. The result is that K-Means++ has better results in the crime types of assault and fraud. Meanwhile, K-Means has good results only in the kind of robbery crime, and the other seven types have the same results.

Also a research of grouping provinces in Indonesia based on poverty level in 2019 using partitioning and hierarchical methods with cluster validation Connectivity, Dunn, Silhouette, and Davies Bouldin [8]. The result showed that the hierarchical method with the number of clusters of $k = 2$ was the best because the Silhouette and Dunn Index values are close to 1. The higher the value of both methods, the more optimal the cluster is.

In previous studies, no one has discussed the clustering of districts/cities in West Kalimantan based on indicators and factors of crime, so this study will use 10 variables where four variables consist of crime indicators and six variables consist of crime factors in 2022. The number of clusters to be analyzed is $k = 2, 3,$ and 4.

## 2. RESEARCH METHODS

### 2.1 Materials and Data

The data used in this study are secondary data obtained from BPS West Kalimantan with 10 variables. The variables consist of crime indicators and factors in districts/cities in West Kalimantan in 2022. Table 1 below is a table of research variables.

**Table 1. Research Variables**

| Variable | Description | Unit |
|---|---|---|
| $X_1$ | Crime Total | Case |
| $X_2$ | Crime Clearance | % |
| $X_3$ | Crime Rate | Person |
| $X_4$ | Crime Clock | Second |
| $X_5$ | Population | Person |
| $X_6$ | The percentage of Poor People | % |
| $X_7$ | Open Unemployment Rate | % |
| $X_8$ | Gini Ratio | Ratio |
| $X_9$ | Average Years of Schooling | Year |
| $X_{10}$ | Human Development Index | Index |

***Data source*: *BPS West Kalimantan***

### 2.2 Cluster Analysis

Cluster analysis is a multivariate technique that aims to group some objects based on their similar characteristics. The clusters formed should have high similarity within clusters and high differences between clusters. In cluster analysis, some assumptions must be fulfilled, the first is that the sample must represent the population, and the second is that there is no correlation between variables [9]. If there is a correlation between variables, the clustering result is not good. One of the criteria that can be used to determine the correlation between variables is the Variance Inflation Factor (VIF). There is no correlation if the value of VIF is $\geq 10$ [10].

$$\text{VIF}_y = \frac{1}{1 - R_y^2} \tag{1}$$

Each variable has a VIF value and the value is found using Equation (1) above. $R_y^2$ is the coefficient of determination for the $y$-th. Variables with a VIF value $\geq 10$ indicate that the variable is highly correlated with other variables, so it must be overcome.

Besides these assumptions, standardization data is required in cluster analysis if data units have a difference. The difference can make the calculations invalid. Data can be standardized by transforming data to $Z$ Score [10].

where:

$$Z_{iy} = \frac{x_{iy} - \bar{x}_y}{s_y} \quad (2)$$

$Z_{iy}$ : transformation value on the $i$-th object of the $y$-th variable
$x_{iy}$ : data from the $i$-th object of the $y$-th variable
$\bar{x}_y$ : average of the $y$-th variable data
$s_y$ : standard deviation of the $y$-th variable data

The purpose of cluster analysis is to group the objects with similarity into one cluster. This similarity can be expressed in the distance between objects. A smaller distance means they are similar to each other. A commonly used distance measure is the Euclidean distance [11].

$$d_{i,j} = \sqrt{\sum_{y=1}^{n} (x_{iy} - x_{jy})^2} \quad (3)$$

where:
$d_{i,j}$ : distance between $i$-th object and $j$-th object
$x_{iy}$ : data from the $i$-th object on the $y$-th variable
$x_{jy}$ : data from the $j$-th object on the $y$-th variable
$n$ : number of variables

In cluster analysis, there are two methods, namely Hierarchical Method and Non-Hierarchical Method. The method for this research is Non-Hierarchical Method, this method starts by selecting the desired number of $k$ and then the objects will join into that cluster.

### 2.3 K-Means++ Clustering

K-Means++ is a cluster analysis method that overcomes the shortcomings of the K-Means method. K-means is one of the most commonly used cluster analysis methods, but in reality, this method still has weaknesses in analyzing the initialization of the centroid [12]. These two methods aim to divide data with a number of $N$ objects into $k$ clusters so the objects in the same cluster have a similarity [13]. K-Means aims to make the object have the largest similarity in one cluster and the smallest similarity between clusters [14]. The K-Means method, which is easy to apply in practice, needs to be improved in selecting a random initial center point (centroid) for each cluster. The cluster results are highly dependent on the initial centroid selection stage. Finding the ideal cluster result is quite difficult if the centroid selection is not done properly.

The development of the K-Means++ method overcomes this shortcoming. A more intelligent selection of initial centroids will have more consistent cluster results and minimize iteration [12]. The cluster center point selection in K-Means++ is done by selecting only one centroid randomly and then calculating the centroid distance with other objects to find the initial centroid for different clusters. The object with the largest distance value has the highest chance of becoming the new centroid [5]. The determination of the new centroid in K-Means++ can be calculated using the following formula [7]:

$$K = \max_{x_i \in X} K(x_i) \quad (4)$$

Where the formula $K(x_i)$ can be found with the following formula:

$$K(x_i) = \frac{D(x_i)^2}{\sum_{x_i \in X} D(x_i)^2} \quad (5)$$

where:
$K$ : the latest initial centroid value
$D(x_i)$ : minimum distance of the $i$-th object to the centroid
$\sum_{x_i \in X} D(x_i)^2$ : total minimum distances between objects and the centroid

After $k$ centroid has been selected, the next step is like the K-Means method, where a new centroid value will be obtained from each cluster's average value. The average value can be seen in the following equation [15].

$$C_{py} = \frac{\sum_{i=1}^{a_p} x_{iy}}{a_p} \quad (6)$$

where:
$C_{py}$ : average centroid of cluster $p$ for the $y$-th variable
$a_p$ : number of cluster members $p$
$x_{iy}$ : the value of the $i$-th object on the $y$-th variable for the cluster

### 2.4 Dunn Index Validation

One of the difficulties in analyzing clusters is determining the optimal number of clusters, so a validity index, which evaluates clustering results in choosing the best number of clusters is needed. Index validation is divided into external and internal, where external validation requires information outside the data. Meanwhile, internal validation focuses on information from the data itself [16].

Dunn Index will calculate the ratio of the minimum value of inter-cluster dissimilarity as separation and the maximum value of intra-cluster diameter as compactness. The optimal cluster is seen from the high Dunn Index (DI) value, which indicates that objects within a cluster are compact and objects between clusters are well separated [17]. The following equation is used in calculating the Dunn Index [18].

1. Calculate the inter-cluster distance and intra-cluster diameter with the following equation:

$$d_{(c_p,c_q)} = \min_{x_i \in c_p, y_i \in c_q} d_{(x_i,y_i)} \quad (7)$$

$$diam(c_r) = \max_{z_i, z_j \in c_r} d_{(z_i,z_j)} \quad (8)$$

where:
$d_{(c_p,c_q)}$ : distance between cluster $p$ and cluster $q$
$diam(c_r)$ : diameter of the cluster $r$
$x_i$ : $i$-th object in cluster $p$
$y_i$ : $i$-th object in cluster $q$
$z_i, z_j$ : $i$-th and $j$-th objects in cluster $r$

2. Calculate the Dunn Index (DI) value using the following equation:

$$\boldsymbol{DI = \frac{\min d_{(c_p,c_q)}}{\max diam(c_r)}} \quad (9)$$

## 3. RESULTS AND DISCUSSION

### 3.1 Overview of Indicators and Factors of Crime in Districts/Cities in West Kalimantan

The following are the descriptive statistics of the indicators and factors of crime in West Kalimantan 2022, presented in Table 2. Based on Table 2, it is known that the data consists of 14 objects, including districts/cities in West Kalimantan.

**Table 2. Descriptive Statistics**

| Variable | | *N* | Minimum | Maximum | Mean |
|---|---|---|---|---|---|
| Crime Total | $X_1$ | 14 | 79.00 | 1078.00 | 292.90 |
| Crime Clearance | $X_2$ | 14 | 69.62 | 116.61 | 84.95 |
| Crime Rate | $X_3$ | 14 | 38.00 | 162.00 | 72.07 |
| Crime Clock | $X_4$ | 14 | 29254.00 | 399189.00 | 171960.00 |
| Population | $X_5$ | 14 | 131104.00 | 669795.00 | 395813.00 |
| The percentage of Poor People | $X_6$ | 14 | 4.12 | 11.44 | 7.07 |
| Open Unemployment Rate | $X_7$ | 14 | 1.33 | 9.92 | 4.62 |
| Gini Ratio | $X_8$ | 14 | 0.26 | 0.36 | 0.30 |
| Average Years of Schooling | $X_9$ | 14 | 6.21 | 10.44 | 7.43 |
| Human Development Index | $X_{10}$ | 14 | 63.81 | 80.48 | 68.48 |

The maximum number of crime cases ($X_1$) reached 1078 cases, which occurred in Kubu Raya Regency, while the district with the minimum number of crimes occurred in Kayong Utara Regency, where there were only 79 cases. The highest crime clearance ($X_2$) occurred in Sambas Regency and reached more than 100%, namely 116.61%. This happened because cases in 2021 were only resolved in 2022, so the completion percentage was more than 100%. Also, for crime rate ($X_3$) it is known that the maximum value is 162, where this number occurs in Kubu Raya, while the minimum occurs in Kapuas Hulu, where there were only 38 cases.

In addition to Kubu Raya Regency contributing the most cases in West Kalimantan, Kubu Raya Regency is also the district with the fastest time interval between cases of criminality, namely 29254 seconds or 8 hours 7 minutes 34 seconds between cases. The highest population ($X_5$), open unemployment rate ($X_7$), average years of schooling ($X_9$), and human development index ($X_{10}$) occurred in Kota Pontianak, while the highest percentage of poor people ($X_6$) and Gini ratio ($X_8$) occurred in Melawi Regency.

### 3.2 Multicollinearity Test

Before conducting the multicollinearity test, data standardization is needed to homogenize data units by converting them to $Z$ Score by using the **Equation (2)**. The standardized data will be the basis for further analysis, namely the multicollinearity test. A multicollinearity test is conducted to determine whether the research variables are correlated. In cluster analysis, it is expected to pass this test. In this study, the multicollinearity test was carried out by calculating the VIF value with **Equation (1)**. If the VIF value was $\geq 10$, multicollinearity occurred. **Table 3** below is a VIF value of each variable.

**Table 3. VIF Value**

| Variable | | VIF Value |
|---|---|---|
| Crime Total | $X_1$ | 7.46 |
| Crime Clearance | $X_2$ | 6.53 |
| Crime Rate | $X_3$ | 8.09 |
| Crime Clock | $X_4$ | 6.01 |
| Population | $X_5$ | 8.12 |
| The percentage of Poor People | $X_6$ | 1.94 |
| Open Unemployment Rate | $X_7$ | 4.22 |
| Gini Ratio | $X_8$ | 6.03 |
| **Average Years of Schooling** | $\boldsymbol{X_9}$ | **30.96** |
| **Human Development Index** | $\boldsymbol{X_{10}}$ | **14.99** |

The VIF value of $X_9$ and $X_{10}$ are 30.96 and 14.99, it means that there is a correlation between variables that need to be overcome by removing one of the variables, which is $X_9$ because it has the largest VIF value. After removing that variable then checking the VIF value again.

**Table 4. VIF Value After Removing Variable $X_9$**

| Variable | | VIF Value |
|---|---|---|
| Crime Total | $X_1$ | 7.09 |
| Crime Clearence | $X_2$ | 2.27 |
| Crime Rate | $X_3$ | 6.49 |
| Crime Clock | $X_4$ | 5.13 |
| Population | $X_5$ | 4.43 |
| The percentage of Poor People | $X_6$ | 1.93 |
| Open Unemployment Rate | $X_7$ | 3.67 |
| Gini Ratio | $X_8$ | 3.26 |
| Human Development Index | $X_{10}$ | 7.72 |

After removing that variable, it is found that multicollinearity symptoms have been resolved because all new VIF values do not exceed 10, so the correlation between variables can be overcome, and all variables have passed the multicollinearity test. After the multicollinearity test, the distance between objects is calculated with the Euclidean distance using the **Equation (3)** to measure the similarity between objects.

### 3.3 Grouping Districts/Cities in West Kalimantan

The grouping of districts/cities in West Kalimantan in determining the level of vulnerability to crime was carried out using the K-Means++ method, where only the first centroid was randomly selected, while the selection of further initial centroids was carried out through calculation. In this study, the number of clusters used is between 2 to 4 clusters. **Table 5** below shows the results of clustering districts/cities in West Kalimantan using the K-means++.

**Table 5. Clusters of District/Cities**

| $k$ | Cluster | Number of Members | Districts/Cities | Initial Centroid |
|---|---|---|---|---|
| $k = 2$ | 1 | 11 | Sambas, Bengkayang, Landak, Mempawah, Sanggau, Sintang, Kapuas Hulu, Sekadau, Melawi, Kayong Utara, Kota Singkawang | Kayong Utara and Kubu Raya |
| | 2 | 3 | Ketapang, Kota Pontianak, Kubu Raya | |
| $k = 3$ | 1 | 10 | Bengkayang, Landak, Mempawah, Sanggau, Sintang, Kapuas Hulu, Sekadau, Melawi, Kayong Utara, Kota Singkawang | Kapuas Hulu, Kubu Raya, and Kota Pontianak |
| | 2 | 3 | Sambas, Ketapang, Kubu Raya | |
| | 3 | 1 | Kota Pontianak | |
| $k = 4$ | 1 | 4 | Kapuas Hulu, Sekadau, Melawi, Kayong Utara | Melawi, Kubu Raya, Kota Pontianak, and Kota Singkawang |
| | 2 | 2 | Ketapang, Kubu Raya | |
| | 3 | 1 | Kota Pontianak | |
| | 4 | 7 | Sambas, Bengkayang, Landak, Mempawah, Sanggau, Sintang, Kota Singkawang | |

After grouping the districts/cities with the K-Means++ method, the clusters will be evaluated using the Dunn Index validation to determine the optimal number of clusters. The higher the DI value, the better the number of clusters formed. **Table 6** below is the result of the Dunn Index value for $k = 2, 3,$ and 4.

**Table 6. Dunn Index Value**

| Number of Clusters | Dunn Index (DI) Value |
|---|---|
| 2 | 0.50 |
| 3 | 0.55 |
| 4 | 0.41 |

The largest Dunn Index value is 0.55, which is obtained when the number of clusters $k = 3$ so that the optimal number of clusters in grouping districts/cities in West Kalimantan based on indicators and crime factors is three clusters. Then, the clustering results can be interpreted by looking at the average value of each variable in each cluster.

**Table 7. Average Variables of Each Cluster**

| Variable | | Average | | |
|---|---|---|---|---|
| | | 1 | 2 | 3 |
| Crime Total | $(X_1)$ | 188.50 | 652.00 | 259.00 |
| Crime Clearance | $(X_2)$ | 82.07 | 98.13 | 74.13 |

| Variable | | Average | | |
|---|---|---|---|---|
| | | 1 | 2 | 3 |
| Crime Rate | $(X_3)$ | 57.80 | 107.33 | 109.00 |
| Crime Clock | $(X_4)$ | 209198.50 | 64563.00 | 121760.00 |
| Population | $(X_5)$ | 300960.30 | 620659.33 | 669795.00 |
| The percentage of Poor People | $(X_6)$ | 7.40 | 6.81 | 4.46 |
| Open Unemployment Rate | $(X_7)$ | 3.60 | 6.22 | 9.92 |
| Gini Ratio | $(X_8)$ | 0.30 | 0.27 | 0.36 |
| Human Development Index | $(X_{10})$ | 67.34 | 68.26 | 80.48 |

Based on **Table 7** in the first cluster, the values of crime total $(X_1)$, crime rate $(X_3)$, population $(X_1)$, and open unemployment rate $(X_7)$ are the lowest among the other clusters. Also, the crime clock $(X_4)$ in this cluster is the largest, it means that the distance between crimes is very far or crimes rarely occur. This indicates that the first cluster is a cluster of districts/cities in West Kalimantan that are not vulnerable to crime.

The second cluster has the highest average total crime $(X_1)$ value of the other clusters. Crime clearance $(X_2)$ in this cluster is the largest, but has the lowest average crime clock $(X_4)$. This means that crime cases occur very quickly between cases. However, the average values of crime rate $(X_3)$, population $(X_5)$, percentage of poor people $(X_6)$, open unemployment rate $(X_7)$, and HDI $(X_{10})$ in the second cluster are between the first cluster and the third cluster, so it can be said that the second cluster is a cluster consisting of districts/cities with vulnerable levels of crime.

The third cluster has the highest average population value of the other clusters, followed by the highest open unemployment rate and Gini ratio. In addition, the third cluster has the smallest percentage of crime clearance, only 74.13%, with a population risk of 109 people affected by crime. This shows that the third cluster is a cluster that consists of areas with a very high level of vulnerability to crime compared to other clusters.

Based on the analysis of the average value of variables in each cluster, **Table 8** shows which groups districts/cities in West Kalimantan based on the crime that occurred.

**Table 8. District/City Vulnerability Level in West Kalimantan**

| Cluster | Number of Members | District/City | Vulnerability Level of Crime |
|---|---|---|---|
| 1 | 10 | Bengkayang, Landak, Mempawah, Sanggau, Sintang, Kapuas Hulu, Sekadau, Melawi, Kayong Utara, Kota Singkawang | Not Vulnerable |
| 2 | 3 | Sambas, Ketapang, Kubu Raya | Vulnerable |
| 3 | 1 | Kota Pontianak | Very Vulnerable |

The district/city included in the first cluster with a cluster category that is not vulnerable to crime is Kayong Utara Regency. One of the efforts of the Kayong Utara Police in handling cases in the area is to hold discussions with the local community to accommodate security and public order issues. In addition, Kayong Utara also has a small population, and this creates an opportunity for criminals. This is evidenced by the data on the crime total in Kayong Utara, which is the lowest.

The district/city included in the second cluster with the category of areas vulnerable to crime is Kubu Raya Regency. Kubu Raya police handle the most reported cases among other regions, but the community considers that Kubu Raya police are not responsive enough to community reports, so Kubu Raya is a vulnerable area.

Kota Pontianak is the district/city in the third cluster, which includes areas very vulnerable to crime. This occurs because Kota Pontianak is the capital of West Kalimantan, which indicates that the city is the

center of population activity. The dense life of the capital city can lead to limited employment opportunities, resulting in poverty and economic inequality in the area. This can encourage people to commit crimes to fulfill their needs.

## 4. CONCLUSIONS

Based on the results of the analysis of the application of the K-Means++ method with validation of the Dunn Index in grouping districts/cities in West Kalimantan based on the level of vulnerability to crime. The following conclusions were obtained:

1. The optimal cluster formed using K-Means++ with Dunn Index validation is three clusters with a Dunn Index value of 0.55.
2. The first cluster with members Bengkayang, Landak, Mempawah, Sanggau, Sintang, Kapuas Hulu, Sekadau, Melawi, Kayong Utara, and Kota Singkawang is a cluster categorized as not vulnerable to crime. The second cluster, with members Sambas, Ketapang, and Kubu Raya, is a cluster that is vulnerable to crime. Also, the third cluster with one member, Kota Pontianak, is a cluster that is very vulnerable to crime.

## REFERENCES

[1] R. E. Fajri and C. Z. Rizki, "Pengaruh Pertumbuhan Ekonomi, Kepadatan Penduduk, dan Pengangguran Terhadap Kriminalitas Perkotaan Aceh," J. Ilm. Mhs., vol. 4, no. 3, pp. 255–263, 2019.
[2] Pusiknas, "Statistik Kejahatan," Pusat Informasi Kriminal Nasional. https://pusiknas.polri.go.id/data_kejahatan (accessed Sep. 18, 2023).
[3] F. P. Marpaung, P. S. Ilham, Y. Chenata, and R. B. Nugroho, Provinsi Kalimantan Barat Dalam Angka 2023. Pontianak: BPS Provinsi Kalimantan Barat, 2023.
[4] F. K. Gulagiz and S. Suhap, "Comparison of Hierarchical and Non-Hierarchical Clustering Algorithms," Int. J. Comput. Eng. Inf. Technol., vol. 9, no. 1, pp. 6–14, 2017, [Online]. Available: www.ijceit.org
[5] C. A. S. Fastaf and Y. Yamasari, "Analisa Pemetaan Kriminalitas Kabupaten Bangkalan Menggunakan Metode K-Means dan K-Means++," J. Informatics Comput. Sci., vol. 3, no. 04, pp. 534–546, 2022.
[6] H. Malikhatin, A. Rusgiyono, and D. A. I. Maruddani, "Penerapan K-Modes Clustering dengan Validasi Dunn Index Pada Pengelompokan Karakteristik Calon TKI Menggunakan R-GUI," J. Gaussian, vol. 10, no. 3, pp. 359–366, 2021.
[7] N. Nugroho and F. D. Adhinata, "Penggunaan Metode K-Means dan K-Means ++ Sebagai Clustering Data Covid-19 di Pulau Jawa," TEKNIKA, vol. 11, no. 3, pp. 170–179, 2022.
[8] N. Afira and A. W. Wijayanto, "Analisis Cluster Kemiskinan Provinsi di Indonesia Tahun 2019 dengan Metode Partitioning dan Hierarki," Komputika J. Sist. Komput., vol. 10, no. 2, pp. 101–109, 2021.
[9] S. Hanada and T. S. Yanti, "Penggunaan Analisis Cluster dalam Pengelompokan Kecamatan di Kabupaten Karawang Berdasarkan Metode Kontrasepsi Peserta KB Aktif," Pros. Stat., pp. 42–49, 2021.
[10] N. Ulinnuh and R. Veriani, "Analisis Cluster dalam Pengelompokan Provinsi di Indonesia Berdasarkan Variabel Penyakit Menular Menggunakan Metode Complete Linkage , Average Linkage dan Ward," InfoTekJar J. Nas. Inform. dan Teknol. Jar., vol. 5, no. 1, pp. 101–108, 2020.
[11] N. Thamrin and A. W. Wijayanto, "Comparison of Soft and Hard Clustering: A Case Study on Welfare Level in Cities on Java Island," Indones. J. Stat. Its Appl., vol. 5, no. 1, pp. 141–160, 2021.
[12] S. M. Miraftabzadeh, C. G. Colombo, M. Longo, and F. Foiadelli, "K-Means and Alternative Clustering Methods in Modern Power Systems," IEEE Access, vol. 11, no. September, pp. 119596–119633, 2023, doi: 10.1109/ACCESS.2023.3327640.
[13] A. S. Ahmar, D. Napitupulu, R. Rahim, R. Hidayat, Y. Sonatha, and M. Azmi, "Using K-Means Clustering to Cluster Provinces in Indonesia," J. Phys. Conf. Ser., vol. 1028, no. 1, 2018, doi: 10.1088/1742-6596/1028/1/012006.
[14] X. Linyao and W. Jianguo, "Improved K-means Algorithm Based on Optimizing Initial Cluster Centers and Its Application," Atl. Press, vol. 79, pp. 5–10, 2018.
[15] D. A. Pramudita and B. Sumargo, "Pengelompokan Pengguna Internet dengan Metode K-Means Clustering," J. Stat. dan Apl., vol. 3, no. 1, pp. 1–12, 2019, doi: 10.21009/jsa.03101.
[16] K. P. Sinaga and M. S. Yang, "Unsupervised K-means clustering algorithm," IEEE Access, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.
[17] M. S. Yang and I. Hussain, "Unsupervised Multi-View K-Means Clustering Algorithm," IEEE Access, vol. 11, no. January, pp. 13574–13593, 2023, doi: 10.1109/ACCESS.2023.3243133.
[18] H. Mahmood, T. Mehmood, and L. A. Al-Essa, "Optimizing Clustering Algorithms for Anti-Microbial Evaluation Data: A Majority Score-Based Evaluation of K-Means, Gaussian Mixture Model, and Multivariate T-Distribution Mixtures," IEEE Access, vol. 11, no. May, pp. 79793–79800, 2023, doi: 10.1109/ACCESS.2023.3288344.