The Application of the K-Medoid Classification Method for Analyzing Crime Rates in South Sulawesi

Suwardi Annas[1], Aswi Aswi[2]*, Irwan[3]
[1,2] Statistics Department, Universitas Negeri Makassar, Makassar, Indonesia
[3] Mathematics Study Program, Universitas Negeri Makassar, Makassar, Indonesia
*Corresponding author: aswi@unm.ac.id

Received: 8 August 2024
Revised: 29 Oktober 2025
Accepted: 30 Oktober 2025

ABSTRACT – **This study employs the *k*-medoid clustering method to analyze districts and cities in South Sulawesi based on their crime rates. As the population increases, employment opportunities may decline, potentially elevating stress levels and, consequently, the likelihood of criminal behavior. To evaluate the distribution of criminal incidents across South Sulawesi, the *k*-medoid method is used to classify regions into clusters. Unlike other clustering methods, *k*-medoid utilizes the median as the cluster center (medoid), making it more robust to outliers. Specifically, the Partitioning Around Medoids (PAM) algorithm is applied, in which initial objects are randomly selected to represent clusters. If the error value is high, the cluster centers are iteratively adjusted until the error is minimized. The dataset consists of crime incidence data for South Sulawesi in 2020, encompassing various types of crimes. Based on the Silhouette coefficient, the optimal number of clusters was determined to be three: Cluster 1 comprises 11 regions, Cluster 2 includes 8 regions, and Cluster 3 contains 5 regions. These clusters provide a comprehensive overview of the crime patterns across different regions within the province.**

Keywords – **Cluster, k-medoid, crime, South Sulawesi.**

1. INTRODUCTION

Object clustering is a versatile analytical method applicable across various domains. One significant application is in the spatial mapping of regions, allowing for the grouping of areas based on shared characteristics. The use of object clustering can provide valuable insights for local government policymakers by identifying regions with specific potentials or deficiencies. To effectively perform such clustering, cluster analysis is employed, which provides a robust statistical framework for grouping objects based on their similarities.

Cluster analysis is a statistical technique used to group objects or variables into clusters, where each object or variable shares similar properties and characteristics [1], [2], [3]. Hierarchical methods involve a sequence of clustering decisions that form a hierarchical or tree-like structure. These methods combine objects into clusters based on their similarities or divide large clusters into smaller ones. An example of a hierarchical method is the Ward method [2]. In contrast, non-hierarchical methods do not utilize a tree-like structure. Instead, they assign objects to clusters once the number of clusters is predetermined. This process begins by determining the cluster centers and assigning objects to the nearest center [1]. Examples of non-hierarchical clustering methods include the k-medoid method and Clustering Large Applications (CLARA) [4].

The k-medoid method uses the median as the cluster center, or medoid, making it robust to outliers. This method, which employs the Partitioning Around Medoids (PAM) algorithm, randomly selects objects to represent the clusters until all objects are selected. If the error value is large, the cluster center is replaced with another point until the error is minimized. *K*-medoids perform well on small-scale data [5].

Some studies have used the *k*-medoid method [6], [7]. A study on clustering provinces in Indonesia based on sustainable development indicators using three methods: Ward's method, *k*-means, and *k*-medoids has been conducted [6]. The results showed that the seven *k*-medoid clusters were the most effective, as they had a very small diversity ratio and distinct characteristics between clusters, providing a clearer view of sustainable development conditions in Indonesia. A study focused on outlier detection in hotspot data using the *k*-medoid and CLARA methods [7]. The results of these methods were utilized for clustering hotspots and detecting outliers in hotspot data.

This research employs the *k*-medoid method to cluster districts and cities in a case study of the crime rate in South Sulawesi. Crime occurs in nearly all regions, affecting both rural and especially urban areas. As the population increases, employment opportunities become more limited, leading to heightened stress levels. These adverse conditions ultimately contribute to higher crime rates, including offenses such as theft, extortion, murder, and harassment. Such criminal activities significantly disrupt the local community. To mitigate crime, it is essential to provide recommendations for policymakers to implement effective measures. Therefore, it is necessary to observe and map crime rates within neighborhoods.

2. LITERATURE REVIEW

2.1. *Cluster Analysis*

Cluster analysis is a technique used to group objects based on similar characteristics [2]. Objects within a cluster tend to be more similar to each other than to objects in other clusters, resulting in minimal diversity within clusters and maximal diversity between clusters. To perform cluster analysis, a measure is needed to determine the similarity or dissimilarity between objects. The most common approach is to use the distance between pairs of objects. Objects with shorter distances between them are more similar and are grouped into the same cluster, whereas objects with longer distances are placed in different clusters. The Euclidean distance is the most commonly used measure in cluster analysis because it is straightforward to apply to more than two variables. For a dataset X with p variables across m observations, the Euclidean distance formula is as follows:

$$d_{ij} = \left\{\left(x_{i1} - x_{j1}\right)^2 + \left(x_{i2} - x_{j2}\right)^2 + \cdots + \left(x_{im} - x_{jm}\right)^2\right\}^{\frac{1}{2}} \tag{1}$$

The value of d_{ij} represents the distance between the data of the i-th object and the j-th object [5].

2.2. K-Medoid

K-medoid is a non-hierarchical clustering technique similar to *k*-means, but there is a fundamental difference between the two: *k*-means uses the average as the centroid, while *k*-medoid uses the median as the center of the cluster [5]. The algorithm commonly used in *k*-medoid clustering is PAM. This method uses the median of the data as the cluster center, making it more robust to outliers than the *k*-means method [4]. PAM minimizes the sum of the distances between each object and the medoid point. The absolute error criterion used is given by the formula:

$$E = \sum_{g=1}^{k} \sum_{O j \in k} \left|O_j - O_i\right| \tag{2}$$

where

E is the sum of absolute errors for all objects in the dataset,

O_j represents non-medoid objects in the g -th cluster,

O_i is the object selected as the medoid in the g -th cluster.

The iteration of the algorithm continues until all objects have been considered as medoids and the sum of absolute errors is minimized. The initial medoids are chosen randomly. The process of swapping medoids with non-medoid objects continues as long as the quality of the clustering results improves. The quality is assessed using a cost function that calculates the average distance between an object and its medoid within a cluster.

Determining whether to swap O_{Random} (unselected objects) with objects selected as *medoids* (O_i) depends on the cost function calculated during each iteration [5]. The PAM algorithm for clustering a dataset with n objects is described as follows:

1) Determine the number of clusters k (initial cluster selection).
2) Randomly select k objects in the dataset as initial or temporary medoids (O_i).
3) Randomly select non-medoid objects (O_{Random}) in the dataset.
4) Calculate the distance between the non-medoid object (O_{Random}) and the initial or temporary medoid (O_i) using Equation (1).
5) Calculate the total proximity cost T (average of the smallest or minimum distances) from the objects to the medoids.
6) Assign cluster members to the temporary medoids.
7) Perform medoid iteration by swapping medoid O_i with O_{Random}.
8) Calculate the total distance difference ($S_{\text{total distance}}$) as the difference between the new medoid cost T and the old medoid cost T.
9) If $S_{\text{total distance}} < 0$, swap O_i with O_{Random} to form a new set of k-medoids. If $S_{\text{total distance}} > 0$, stop the iteration, and the cluster result from the previous iteration is retained.
10) Repeat steps 2 to 9 until there is no change in the clustering result.

2.3. Cluster Evaluation

Cluster evaluation is crucial for assessing the representativeness of clusters in relation to the population, determining their generalizability to other objects, and ensuring their stability over time. One method for evaluating clusters is the silhouette coefficient method. This method combines the concepts of cohesion and separation to assess the quality of clusters. Cohesion measures the closeness of objects within a cluster, as shown in Figure 1(a), while separation assesses how distinct a cluster is from other clusters, as depicted in Figure 1(b). The Figure 1 illustrates these concepts of cohesion and separation.

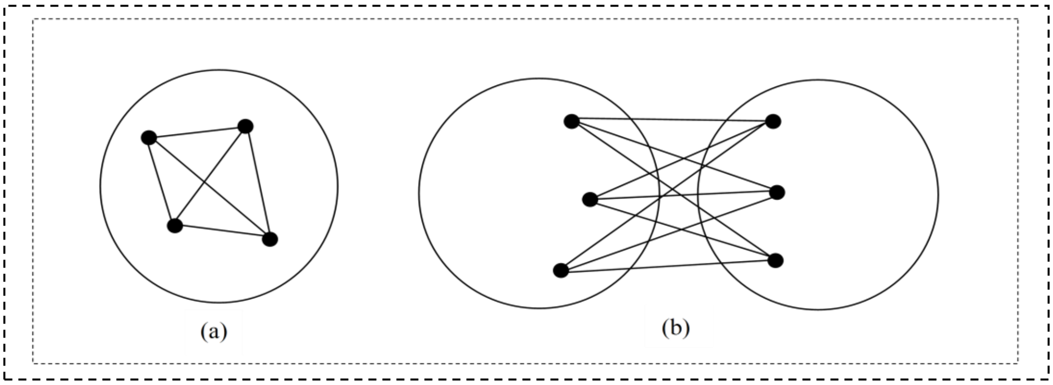

Figure 1 Illustration of silhouette method, distance within clusters (a) and distance between nearest clusters (b)

The formula for calculating the silhouette value of a point is provided in Equation (3) [4], [8].

$$s(i) = \frac{b(i) - a(i)}{\max[a(i), b(i)]} \tag{3}$$

where

$s(i)$ denote the silhouette value of the i -th object in a cluster

$a(i)$ the average distance between the i -th object and all other objects within the same cluster

$b(i)$ the average distance between the i -th object and objects in the nearest neighboring cluster.

The term max[$a(i)$, b(i)] represents the larger of these two values.

The average silhouette value of all points in the dataset, *S'(k)* is used to identify the optimal number of clusters. The cluster *k* with the largest *S'(k)* value is selected. The silhouette coefficient (SC) is defined in Equation (4)

$$SC = \max\{S'(k)\} \tag{4}$$

Where

SC : Overall silhouette coefficient of the cluster

S'(k) : average silhouette value in the *k*-th cluster

max *{S'(k)}* : the largest value of *S'(k).*

The silhouette value ranges from -1 to 1. A value of 1 indicates that all objects or points are correctly clustered. A value of 0 means that objects or points are on the boundary between two clusters, while a value of -1 suggests that objects or points are incorrectly clustered [8],[4].

2.4. Criminality

The term "criminality" refers to acts of wrongdoing, which have been prevalent throughout human civilization. These deviant behaviours are a persistent feature of societal dynamics. Crimes encompass all actions that violate established laws and norms. Common examples of criminals include thieves, terrorists, and murderers, although terrorists differ from other criminals as their acts are often motivated by political values or beliefs [9]. The juridical and sociological meanings of crime differ. Juridically, crime is defined as actions that violate laws or regulations established by authorities. Sociologically, crime refers to actions that harm society and elicit negative reactions from it [9].

Law enforcement officials often face challenges in studying and analysing criminal acts. The development of Geographic Information System (GIS) technology and crime analysis significantly enhances the ability to map crime effectively and efficiently, helping to identify safe areas. Crime mapping involves controlling and processing spatial crime data, producing visualizations that help identify locations and gather information about the number of reported crimes [9]. Crime analysis is a process applied to identify patterns in criminal activity. Administrative and operational personnel can utilize the results of crime analysis to focus on investigating and preventing reported crimes [9].

In sociologically-based criminology, criminality is viewed as a pattern of behaviour that harms society (i.e., there are victims) and elicits social reactions from the community. These social reactions can be formal, informal, or non-formal. Sociologically, the definition of crime as an element of criminality includes two key components:

1) Crime is an act that causes economic and psychological harm.

2) Crime offends the moral sensibilities of a group of people who have the right to express disapproval.

Recent developments in crime analytics have also leveraged artificial intelligence and machine learning approaches for spatial and behavioral crime prediction. [10] compared machine learning-based crime hotspot prediction with police district boundaries, showing that model-based hotspot mapping can significantly enhance forecasting accuracy. Meanwhile, [11] introduced a convolutional neural network (CNN) framework for detecting criminal emotions, highlighting the growing role of deep learning in understanding and preventing crime-related behaviors. These studies emphasize that the integration of clustering methods, spatial analytics, and AI-based modeling can provide more comprehensive insights for crime prevention and policy formulation.

In addition to previous clustering applications, k-means and Self-Organizing Maps (SOM) have been combined for spatial clustering of air pollution in Makassar City, Indonesia [12]. Their study demonstrated that integrating machine learning-based clustering with Geographic Information Systems (GIS) visualization effectively maps environmental phenomena across urban areas. This approach aligns with the use of clustering in regional analysis, providing further evidence of the robustness of unsupervised learning techniques for multidimensional spatial data.

3. METHODOLOGY

3.1. *Object of Study*

The focus of this research is on district and city areas within South Sulawesi Province, specifically analyzing crime data in 2020**.** This data was obtained from the Central Statistics Agency (BPS) of South Sulawesi Province. Districts and cities, along with their respective sectors, **are** presented in Table 1.

Table 1 Districts and cities, along with their respective sectors, within South Sulawesi Province.

No	District/City	No	District/City	No	District/City
1	Selayar	9	Maros	17	Luwu
2	Bulukumba	10	Pangkep	18	Tana Toraja
3	Bantaeng	11	Barru	19	Luwu Utara
4	Jeneponto	12	Soppeng	20	Luwu Timur
5	Takalar	13	Wajo	21	Toraja Utara
6	Gowa	14	Sidrap	22	Makassar
7	Sinjai	15	Pinrang	23	Parepare
8	Bone	16	Enrekang	24	Palopo

The variables utilized in this study include the following types of crime:

X_1 : Number of cases involving crimes against life
X_2 : Number of cases involving crimes against physical integrity or the body
X_3 : Number of cases involving crimes against decency
X_4 : Number of cases involving crimes against personal freedom
X_5 : Number of cases involving crimes against property or goods
X_6 : Number of drug-related crimes
X_7 : Number of cases involving fraud, embezzlement, and corruption
X_8 : Number of cases involving crimes against public order

The research methodology includes the following steps:

1) Detecting outliers for each variable.
2) Preprocessing data, which involves data cleaning and normalization.
3) Conducting a multicollinearity test using a correlation matrix and the Variance Inflation Factor (VIF).
4) Performing *k*-medoids clustering with k = 2, 3, 4, 5, and 6, where k represents the number of clusters.
5) Identifying the optimal cluster configuration based on the highest silhouette coefficient value.
6) Mapping the results using QGIS software and interpreting the outcomes of the *k*-medoids clustering method.

IV. RESULTS AND DISCUSSIONS

4.1. *Data Description*

The description of South Sulawesi crime data is presented in Table 2.

Table 2 Description of South Sulawesi Crime Data

	Minimum	Maximum	Total	Mean	Standard Deviation	Variance
X_1	1.00	11.00	99.00	4.1250	2.89396	8.375
X_2	1.00	30.00	261.00	10.8750	8.84805	78.288
X_3	.00	6.00	47.00	1.9583	1.70623	2.911
X_4	.00	2.00	4.00	.1667	.48154	.232
X_5	12.00	137.00	1288.00	53.6667	34.64813	1200.493
X_6	.00	28.00	128.00	5.3333	6.16911	38.058
X_7	1.00	32.00	305.00	12.7083	8.58451	73.694
X_8	3.00	47.00	377.00	15.7083	11.60764	134.737

Table 2 illustrates that the variables with the fewest recorded criminal incidents are X_3, X_4, and X_6. Conversely, the variable X_5 exhibits the highest incidence rate, with an average of 53.67 occurrences. The number of criminal incidents occurring in South Sulawesi is given in Figure 2 and the distribution of crime rates based on variables is given in Figure 3.

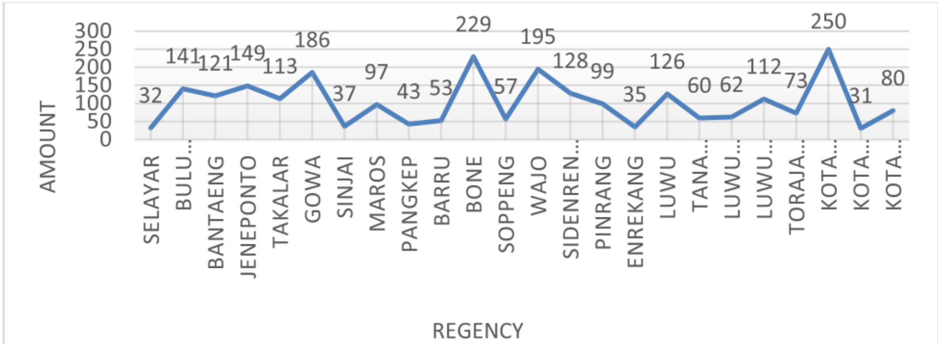

Figure 2 Number of incidents of criminal acts in South Sulawesi

Figure 2 illustrates that Makassar City reports the highest number of criminal incidents, with a total of 250 cases. Conversely, Pare-pare City exhibits the lowest criminal rate, recording only 31 cases.

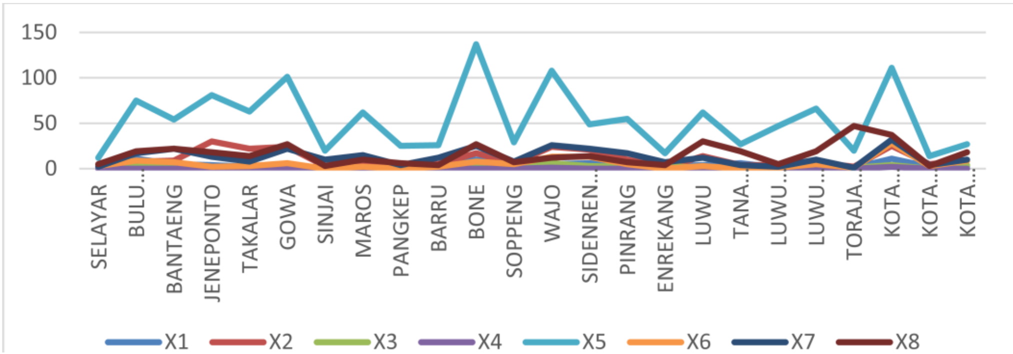

Figure 3 Distribution of crime rates based on variables

Figure 3 demonstrates that variable X_5 is the most frequently observed across all regions in South Sulawesi. Bone records the highest incidence of X_5, followed by Makassar City. Conversely, Selayar shows the lowest occurrence of X_5. Additionally, in the North Toraja region, the frequency of X_8 exceeds that of X_5.

4.2. Clustering Results Using K-Medoid

Clustering using the *K*-Medoids method aims to group data with similar characteristics into a single cluster while separating data with differing characteristics into distinct clusters. For this study, the 2020 criminality data, which includes 8 variables across 24 districts and cities in South Sulawesi Province, were utilized. The clustering results were assessed using the silhouette coefficient, with the optimal clustering identified as the one with the highest silhouette coefficient value. As depicted in Figure 4, the clustering configuration with $k = 3$ (where k represents the number of clusters) achieved the highest silhouette coefficient, which is 0.50.

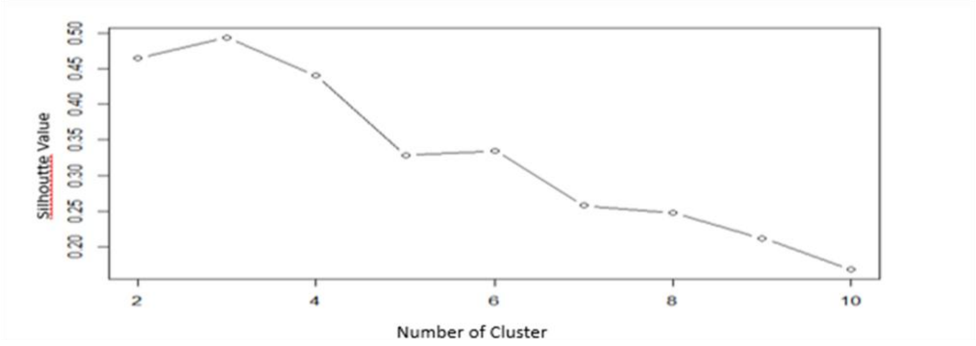

Figure 4 Silhouette value of clusters formed with k - medoid

Table 3 presents the results of the K-Medoids clustering analysis, which identified three distinct clusters.

Table 3 *K*-Medoid method clustering results

Object	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	clustering
SELAYAR	4	2	4	0	12	3	2	5	1
SINJAI	1	3	0	0	20	0	10	3	1
PANGKEP	3	5	0	0	25	0	4	6	1
BARRU	3	5	1	0	26	2	12	4	1
SOPPENG	2	6	0	0	29	5	8	7	1
ENREKANG	3	1	3	0	17	0	7	4	1
TANA TORAJA	6	4	0	0	27	0	4	19	1
LUWU UTARA	1	4	1	1	47	1	2	5	1
TORAJA UTARA	1	2	1	0	20	1	1	47	1
KOTA PARE-PARE	2	5	0	0	14	3	4	3	1
KOTA PALOPO	3	9	4	0	27	9	10	18	1
BULUKUMBA	11	8	2	0	75	9	17	19	2
BANTAENG	5	9	2	0	54	7	22	22	2
TAKALAR	2	22	1	0	63	3	8	14	2
MAROS	3	4	1	0	62	2	15	10	2
SIDRAP	6	21	3	0	49	13	22	14	2
PINRANG	3	10	2	0	55	5	17	7	2
LUWU	4	14	1	0	62	3	12	30	2
LUWU TIMUR	2	7	3	0	66	5	10	19	2
JENEPONTO	4	30	1	0	81	2	13	18	3
GOWA	4	24	2	0	101	6	22	27	3
BONE	10	17	5	0	137	8	25	27	3
WAJO	5	24	6	1	108	13	26	12	3
KOTA MAKASSAR	11	25	4	2	111	28	32	37	3

Table 3 illustrates the composition of the clusters derived from the *K*-Medoids method: Cluster 1 comprises 11 regions, Cluster 2 includes 8 regions, and Cluster 3 encompasses 5 regions. The detailed results of this clustering analysis are presented in Table 4, while the characteristics of each cluster is given in Table 5.

Table 4 Members of each Cluster formed

Cluster	Number of Members	Members
1	11	Selayar, Sinjai, Pangkep, Barru, Soppeng, Enrekang, Tana Toraja, Toraja Utara, Luwu Utara, Pare-Pare, dan Palopo.
2	8	Bulukumba, Bantaeng, Takalar, Maros, Sindenreng Rappang, Pinrang, Luwu, Luwu Timur.
3	5	Jeneponto, Gowa, Bone, Wajo, dan Kota Makassar

Table 5 The characteristics of each cluster

Clusters	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1	29	46	14	1	264	24	64	121
2	36	95	15	0	486	47	123	135
3	34	120	18	3	538	57	118	121

Table 5 presents the characteristics of the three clusters formed. Cluster 1 exhibits the highest number of incidents for variable X_5, with 264 cases, followed by variable X_8 with 121 cases. The lowest number of cases in Cluster 1 is observed for variable X_4, with just 1 case. Given the relatively low number of incidents, this cluster is categorized as a relatively safe area. In Cluster 2, the highest incidence is recorded for variable X_5, with 486 cases, followed by X_8 with 135 cases and X_7 with 123 cases. Notably, there are no recorded cases for variable X_4 in this cluster. Cluster 3 also shows the highest number of incidents for X_5, with 538 cases, making it the highest among all clusters. Despite this, the overall crime rate in Cluster 3 is relatively low, categorizing it as a safe area. However, X_4 has the lowest value in this cluster with 3 cases. Given the high number of incidents for X_5, Cluster 3 is considered a vulnerable area due to its elevated crime rate.

4.3. Clustering Mapping Results Using K-medoid

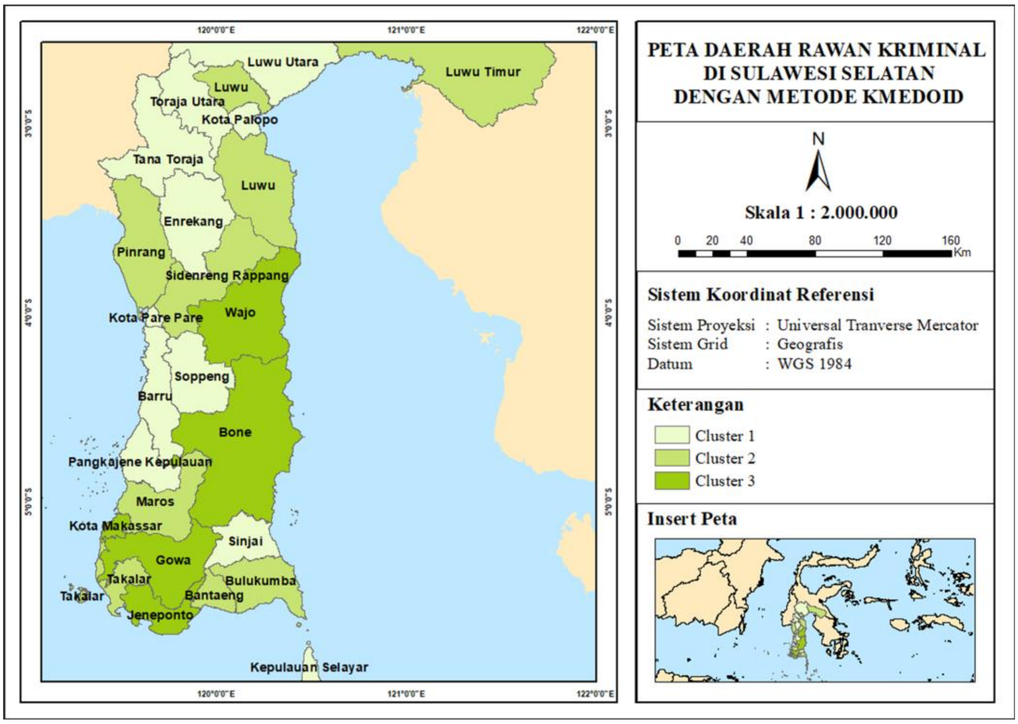

Figure 6 Map of criminal-prone areas in South Sulawesi

Figure 6 illustrates the regional distribution based on the clusters formed. Areas classified in Cluster 1 are predominantly highland regions within South Sulawesi. Similarly, Cluster 2 also encompasses highland areas. In contrast, Cluster 3 includes regions characterized by extensive areas and provincial capitals with high populations.

5. CONCLUSIONS AND SUGGESTIONS

Based on the results of the study, it can be concluded that the cluster analysis of 24 cities/districts, categorized by types of crimes (including crimes against life, physical/body, decency, freedom, property/goods, drugs, fraud, embezzlement, corruption, and public order), results in the formation of three distinct clusters. The first cluster consists of regions with relatively low crime rates, comprising 11 areas; the second cluster consists of regions that are not prone to crime, with 8 areas; and the third cluster consists of regions with relatively high crime rates, comprising 5 areas.

Future studies could extend this analysis by incorporating spatio-temporal dimensions to capture the dynamics of crime rates over multiple years. Moreover, integrating geospatial machine learning techniques such as DBSCAN, SOM, or hybrid K-Medoids–PCA models may enhance clustering accuracy. Further research may also examine socio-economic and demographic covariates to provide deeper insights into the underlying factors contributing to regional variations in crime.

ACKNOWLEDGEMENT

The authors would like to acknowledge Universitas Negeri Makassar for its financial support. This research was funded by the DIPA of Universitas Negeri Makassar, under the grant number SP DIPA-023.17.2.677523/2023, dated February 15, 2023, in accordance with the Rector's Decree of Universitas Negeri Makassar Number 304/UN36/HK/2023, dated March 15, 2023.

REFERENCES

[1] J. F. Hair Jr, W. C. Black Barry J, and B. R. E. Anderson, Multivariate Data Analysis Seventh Edition. New Jersey (US): Pearson, 2010. doi: 10.4324/9781351269360.

[2] R. A. Johnson and D. W. Wichern, Applied Multivariate Statistical Analysis 6th edition.pdf. New Jersey: Prentice Hall, 2007.

[3] J. Supranto, Analisis multivariat : Arti dan Interpretasi. Jakarta: Jakarta: Rineka Cipta, 2004.

[4] L. Kaufman and P. J. Rousseeuw, Finding Group in Data. Kanada: John Willey & Sons, Inc, 1990.

[5] J. Han, M. Kamber, and J. Pei, Data mining: Data mining concepts and techniques. Elsevier: Morgan Kaufmann publications, 2006. doi: 10.1109/ICMIRA.2013.45.

[6] I. M. Ridho, Penggerombolan Provinsi-Provinsi di Indonesia Berdasarkan Indikator Pembangunan Berkelanjutan. Bogor: Skripsi: Institut Pertanian Bogor, 2018.

[7] M. B. Cahyadahrena and I. S. Sitanggang, "Deteksi Pencilan Pada Data Titik Panas Menggunakan Clustering Berbasis Medoids," IPB, 2015. [Online]. Available: http://repository.ipb.ac.id/handle/123456789/76080

[8] D. A. I. C. Dewi and D. A. K. Pramita, "Analisis Perbandingan Metode Elbow dan Silhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali," Matrix J. Manaj. Teknol. dan Inform., vol. 9, no. 3, pp. 102–109, 2019, doi: 10.31940/matrix.v9i3.1662.

[9] D. Dwirohayati, "Pemetaan daerah rawan kriminalitas menggunakan k-means clustering Polrestabes Bandar Lampung," Informatic Institute Business Darmajaya, Bandar Lampung., 2020.

[10] A. Basso, et al., "Comparing Machine Learning-Based Crime Hotspots Versus Police Districts: What's the Best Approach for Crime Forecasting?," Crime Science, Springer Nature, 2025.

[11] J. Raval, N. K. Jadav, S. Tanwar, G. Pau, F. Alqahtani, and A. Tolba, "Criminal Emotion Detection Framework Using Convolutional Neural Network for Public Safety," Scientific Reports, vol. 15, no. 15279, 2025.

[12] S. Annas, U. Uca, I. Irwan, R. H. Safei, and Z. Rais, "Using k-Means and Self Organizing Maps in Clustering Air Pollution Distribution in Makassar City, Indonesia," *Jambura Journal of Mathematics*, vol. 4, no. 1, pp. 167–176, 2022, doi: https://doi.org/10.34312/jjom.v4i1.11883