ENTHUSIASTIC
INTERNATIONAL JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 2, Issue 1, April 2022, pp. 37-46

Implementation of K-Means Algorithm to Group Provinces by Factors Influenced Criminal Act in Indonesia in 2019

Zumrotul Wahidah [a,1], Dina Tri Utari [b,2,*]

[a, b] Departement of Statistics,Universitas Islam Indonesia , Jl Kaliurang KM 14.5, Yogyakarta 55584, Indonesia
[1] 18611094@students.uii.ac.id; [2] dina.t.utari@uii.ac.id*
* corresponding author

ARTICLE INFO

Article history
Received August 30, 2021
Revised April 25, 2022
Accepted April 30, 2022

Keywords
K-means
Poverty
Unemployment

ABSTRACT

A *criminal act* is prohibited by criminal law accompanied by a sanction in the form of a particular crime for whoever violates the prohibition. Criminal action as a social phenomenon is more influenced by various aspects of life in society, including poverty and unemployment factors. Grouping the factors that influence a crime is necessary to find the most recent information that was previously unknown. This research uses the K-Means method, a non-hierarchical cluster analysis that seeks to partition data with the same characteristics into one cluster. The results showed that 3 clusters formed, with cluster 1 covering 18 provinces being the areas with the characteristics of the lowest percentage of poverty and the highest percentage of unemployment. Then cluster group 2 includes 13 regions with the characteristics of moderate poverty and the lowest unemployment rate. Meanwhile, cluster group 3 consists of 3 provinces with the characteristics of the highest percentage of poverty and average percentage of unemployment.

1. Introduction

According to Molejatno, a criminal act is an act prohibited by a criminal law accompanied by sanctions in the form of a specific crime for whoever violates the prohibition [1]. Based on the National Police's Bareskrim Pusiknas, in 2019, there were 83,705 criminal cases from all groups and dominated by conventional crime with a total of 64,583 cases or 77% of all crime groups. In contrast, the highest rank of conventional crime is theft, with 9,988 cases or 15% of conventional crime.

The criminal act is a social phenomenon that never ends to be studied, and it is considering the growing number of criminal cases and the development of human life. Criminal acts as a social phenomenon are more influenced by various aspects of life in society such as politics, economy, social culture, and other matters related to state defense and security. Research conducted by Ria Pasiza et al. (2008) found that crime in Indonesia is influenced by population density, the open unemployment rate, and the percentage of the poor [2]. Another research conducted by Kosmaryati et al., which purposed to determine the factors that influenced crime in Indonesia in 2011-2016, found that the number of unemployed, drug abuse cases, domestic violence cases, embezzlement cases, and fraud cases influenced the crime in Indonesia [3].

p-ISSN 2798-253X
e-ISSN 2798-3153

Some previous research used the K-Means method as a problem-solving solution. First, research conducted by Usap Tatang Suryadi and Yana Supriatna (2019) purposed to create an application to classify the level of theft crimes in districts/cities in West Java [4]. Then, research conducted by Widi Astuti and Djoko Adi Widodo (2016) mapped the city of Semarang. That research purposed to classify street crimes in Semarang City for 2014 [5]. Another research was conducted by Jajang Jaya Purnama et al. (2019), which also used the K-Means Cluster method to group illegal fishing crimes [6].

The rise of criminal acts makes everyone restless; the government needs to get a picture to deal with criminal acts. Therefore, based on the factors that influence the occurrence of these crimes, a study is needed to classify provinces in Indonesia that have almost the same characteristics. As a solution to find information about a crime, one of the methods that can be used is clustering. One type of clustering method is the K-Means method. K-Means is a non-hierarchical cluster analysis that tries to divide data with the same characteristics into one cluster.

2. Method

2.1. Data and Source Data

The research obtained secondary data taken from Badan Pusat Statistik Indonesia [7]. The data used in this research are the factors that influenced criminal acts in Indonesia in 2019.

2.2. Research Variable

The data used in this study are the factors that influenced criminal acts in Indonesia in 2019, including variables are:

1. Percentage of poverty
   This variable describes the percentage (%) of poverty by the province in Indonesia in 2019.
2. Percentage of unemployment
   This variable describes the percentage (%) of unemployment percentage by the province in Indonesia in 2019.

2.3. Research Stage

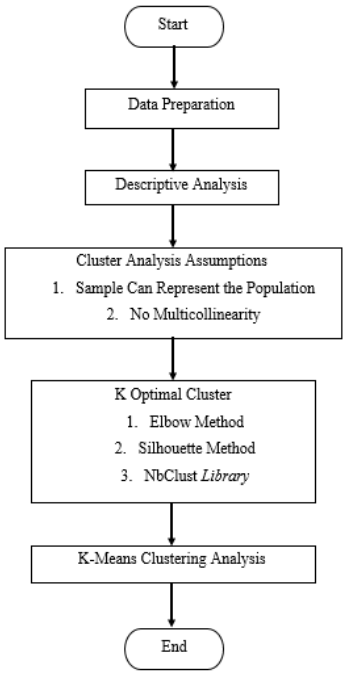

Fig. 1. Research Stage

p-ISSN 2798-253X
e-ISSN 2798-3153

This research uses the K-Means method, non-hierarchical cluster analysis with the help of R studio software. This research has the purpose of grouping provinces by factors that influenced criminal acts in Indonesia in 2019.

2.4. Cluster Analysis Assumption

2.4.1. KMO (Kaiser-Mayer Olkin)

The Kaiser-Mayer Olkin (KMO) test is to see whether the sample is sufficiently representative of the existing population or not so that the clustering can be processed correctly. This KMO test measures the adequacy of the sample for each indicator. The Kaiser-Mayer Olkin (KMO) test has a value of 0 to 1. If the KMO value ranges from 0.5 to 1, the sample can represent the population or a representative sample. KMO test is described by the following formula [9],

$$\boldsymbol{KMO} = \frac{\sum_{j=1}^{p}\sum_{k=1 k\neq j}^{p} r_{x_j x_k}^2}{\sum_{j=1}^{p}\sum_{k\neq j}^{p} r_{x_j x_k}^2 + \sum_{j=1}^{p}\sum_{k\neq j}^{p} p_{x_j x_k, x_1}^2} \quad (4)$$

where,

p = number of variables
$r_{X_j X_k}$ = correlation between X_j and X_k
X_j = average X_j
X_k = average X_k
n = number of observations
$r_{X_j X_k, X_i}$ = partial correlation between X_j, X_k and X_1

2.4.2. No Multicollinearity

Multicollinearity is a perfect or definite linear relationship between several or more variables. Multicollinearity refers to the presence of more than one definite linear relationship. To find out the existence of multicollinearity, one of them is to look at the magnitude of the correlation between the independent variables with the formula [10],

$$r = \frac{n\left(\sum_{j=1}^{p} x_i y_j\right) - \left(\sum_{j=1}^{p} x_i \sum_{j=1}^{p} y_i\right)}{\sqrt{\left(n\sum_{j=1}^{p} {x_i}^2 - \left(\sum_{j=1}^{p} x_i\right)^2\right)\left(n\sum_{j=1}^{p} {y_i}^2 - \left(\sum_{j=1}^{p} y_i\right)^2\right)}} \quad (5)$$

where,

r = correlation coefficient
x_i = first variable value
y_i = second variable value
n = number of data

2.5. K Optimal Cluster

2.5.1. Elbow Method

The elbow method is one of the methods used to select the optimal number of clusters or groups [11]. The elbow algorithm is based on the sum of square error (SSE),

$$SSE = \sum_{k-1}^{k} \sum_{x_i \epsilon S_k} \|X_i - C_k\|^2 \quad (6)$$

where k is the number of groups used in the K-means algorithm, X_i is the number of data, and C_k is the number of clusters in the *k* cluster.

2.5.2. Silhouette Method

Silhouette coefficient is used to see the quality and strength of the cluster, how well an object is placed in a cluster. This method is a combination of cohesion and separation methods. Silhouette coefficient steps are,

1. Calculate the average distance from a document

$$a(i) = \frac{1}{|A|-1} \sum j \epsilon_{A, j\neq i} d(i,j) \quad (7)$$

   where j is another document from a cluster A and d (i,j) is the distance between document I and j
2. Calculate the average distance between document i and all documents in other clusters, and take the smallest value,

p-ISSN 2798-253X
e-ISSN 2798-3153

$$d(i,C) = \frac{1}{|A|}\sum j \, \epsilon d(i,j) \quad (8)$$

where d(I,C) is the average distance of document I with all objects in other clusters C with A≠C.

$$b(i) = \min C \ \neq A \ d(i,C) \quad (9)$$

3. Silhouette coefficient value is:

$$s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \quad (10)$$

2.5.3. NBClust Library

NbClust is one of the libraries in R that helps to find the best number of clusters. The NbClust library provides 30 indexes to determine the number of clusters. To find the number of clusters with the NbClust library, the following syntax is:

```
library(NbClust)
set.seed(123)
nc <- NbClust(data, min.nc=, max.nc=, method="kmeans")
```

where min.nc and max.nc are the minimum and maximum number of desired clusters [12].

2.6. K-Means Cluster

Steps of the K-Means Cluster Analysis method are [8]**,**

1. Determine the number of clusters,
2. Randomly allocate data into the cluster,
3. Calculate the average centroid of the data in each cluster with an equation:

$$C_{kj} = \frac{x_{1j}+x_{2j}+x_{nj}}{n} \quad (1)$$

where

C_{kj} is the center of the k cluster on the j (j=1,2,…,p)

n is quantity cluster k

4. Find the distance of each centroid using Euclidean distance,

$$d(X_i,X_g) = \sqrt{\sum_{j=1}^{p}(X_{ij}-X_{gj})^2} \quad (2)$$

5. Allocate each data to the nearest centroid/average,

$$a_{ij} = \begin{cases} 1, & s = \min\{d(x_i,C_{kj})\} \\ 0, & lainnya \end{cases} \quad (3)$$

a_{ij} = value x_i to cluster center C_{kj},

s = short distance from data x_i to the cluster centerC_{kj} after comparison.

6. If there is still data moving into another cluster, return to step 3.

3. Results and Discussion

3.1. Descriptive Analysis

Before conducting the analysis, descriptive analysis is needed to find out the description of the existing data. The result of a descriptive analysis of the factors that influenced criminal acts in Indonesia in 2019 is in **Table 1**. The descriptive analysis obtained includes the minimum value, quartile 1, median, mean, quartile 3, and maximum value.

p-ISSN 2798-253X
e-ISSN 2798-3153

Table 1. Descriptive Analyze

	Variable	
	Percentage of poverty	Percentage of unemployment
Minimum	3.470%	1.220%
Quartile 1	6.665%	3.228%
Median	9.600%	4.020%
Mean	10.875%	4.411%
Quartile 3	14.290%	5.495%
Maximum	27.530%	7.780%

Based on **Table 1.**, it was found that the highest percentage of poverty reached 27.530% which occurred in Papua, it can be seen from the bar chart **Fig 2.**, while the lowest percentage of poverty was 3.470% which occurred in the capital city of Indonesia, namely Jakarta. The average percentage of poverty in Indonesia is 10.875%. Meanwhile, the highest percentage of unemployment rate is 7.780% in West Java, it can be seen in **Fig 3.** and the lowest is 1.220% in Bali.

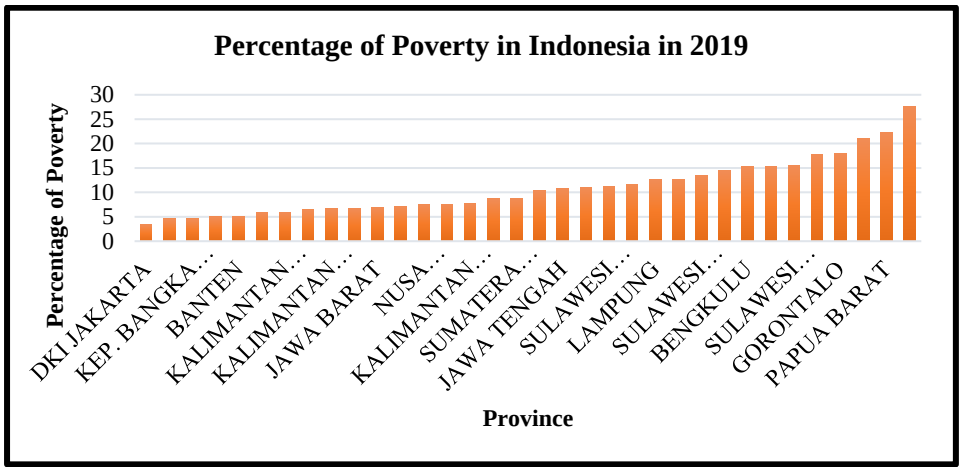

Fig. 2. Percentage of Poverty in Indonesia in 2019.

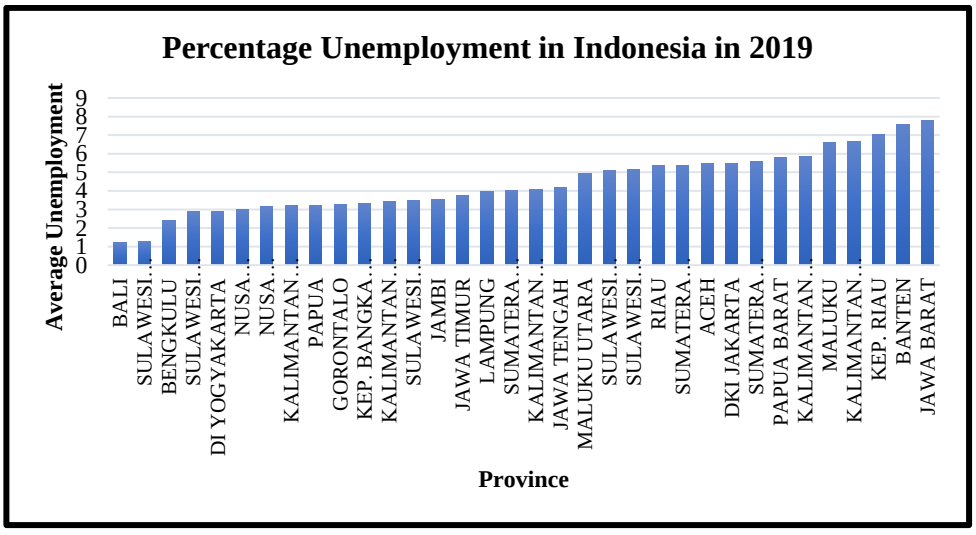

Fig. 3. Percentage of Unemployment in Indonesia in 2019.

p-ISSN 2798-253X
e-ISSN 2798-3153

3.2. Cluster Analysis Assumptions

Before conducting cluster analysis, it is necessary to fulfill two assumptions, namely the sample that can represent the population and there is no multicollinearity. The assumption that the sample can represent the population is obtained from the result of KMO, while the assumption of no multicollinearity is obtained from the correlation coefficient in **Table 2.**

Based on the results of the KMO, it is found that the KMO value is 0.5. This value ranges from 0.5 to 1, which means that the sample can represent the population or a representative sample.

Table 2. Correlation Coefficient

	Percentage of poverty	Percentage of unemployment
Percentage of poverty	1	-0.203
Percentage of unemployment	-0.203	1

Based on **Table 2.**, it is found that the correlation value between the percentage of poverty and the percentage of unemployment is -0.203, which shows that the correlation value is not close to 1 or -1, which means there is no correlation or there is no multicollinearity between the two variables.

From the two assumptions of the test results above, it can be concluded that if both assumptions have been met, the sample is sufficient to represent the population, and there is no multicollinearity.

3.3. *K* Optimal Cluster

One of the most popular methods for determining the optimal number of *k* is using the Elbow method. The Elbow method uses the value of WSS (Within Sum Square).

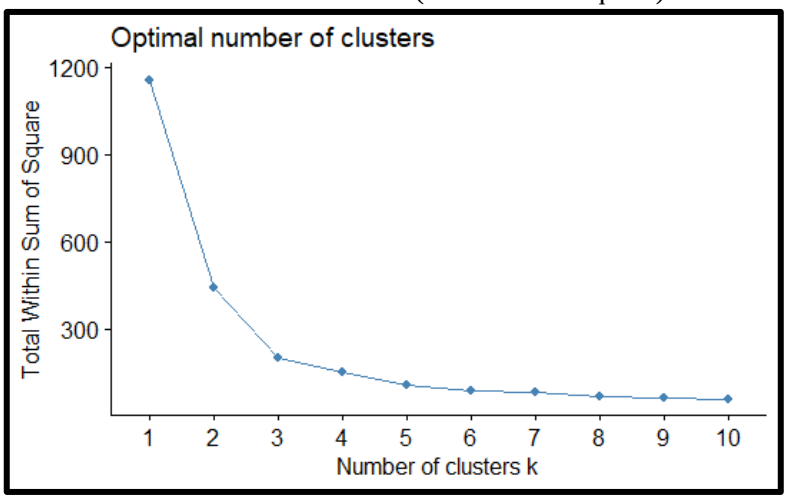

Fig. 4. K Optimal With Elbow Method.

Based on the image above, it is found that the line has a fracture that forms the elbow at $k = 3$. Therefore, it can be concluded that using the Elbow method, k is optimal at $k = 3$.

p-ISSN 2798-253X
e-ISSN 2798-3153

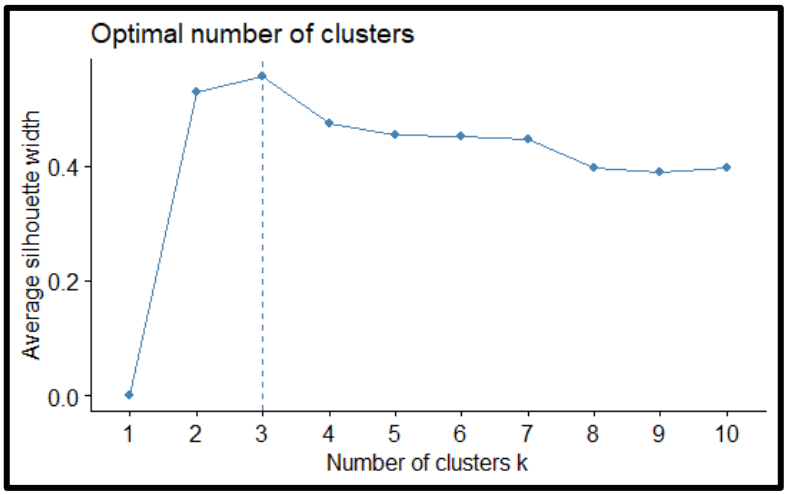

Fig. 5. K Optimal With Silhouette Method.

Meanwhile, using the Silhouette method, the optimal number of the cluster formed is also $k = 3$. Because the average silhouette value at $k = 3$ is the highest of the others.

In addition to using the Elbow and Silhouette methods, there is an R library, namely the NbClust library which is used to determine the number of relevant clusters. The number of cluster results with the help of the NbClust library can be seen in the conclusion of **Fig 6** output.

```
*** : The Hubert index is a graphical method of determining the number of clusters.
                In the plot of Hubert index, we seek a significant knee that corresponds to a
                significant increase of the value of the measure i.e the significant peak in
Hubert
                index second differences plot.

*** : The D index is a graphical method of determining the number of clusters.
                In the plot of D index, we seek a significant knee (the significant peak in
Dindex
                second differences plot) that corresponds to a significant increase of the
value of
                the measure.

*******************************************************************
* Among all indices:
* 4 proposed 2 as the best number of clusters
* 12 proposed 3 as the best number of clusters
* 2 proposed 5 as the best number of clusters
* 1 proposed 9 as the best number of clusters
* 2 proposed 14 as the best number of clusters
* 2 proposed 15 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  3


*******************************************************************
```

Fig. 6. Output Number of Cluster with NbClust Library

Based on the results of looking for k in **Fig 6**, using the NbClust library obtained from a total of all existing indexes, 4 indexes propose 2 as the best number of clusters, 12 indexes propose 3 as the best number of clusters, 2 index proposes 5 as the best number of clusters, 1 index proposes 9 as the best number of clusters, 2 index proposes 14 as the best number of clusters, and 2 indexes proposes 15 as the best number of clusters. According to the majority of the index, it can be concluded that the best number of clusters is when $k = 3$.

Based on the three methods, the authors conclude that k is optimal when $k = 3$.

3.4. K-Means Clustering Analysis

Based on the results of clustering using the K-Means method, cluster group 1 includes 18 provinces, cluster group 2 includes 13 provinces, and cluster group 3 includes 3 provinces, each of which can be seen in **Figure 7.** and **Table 3.** below:

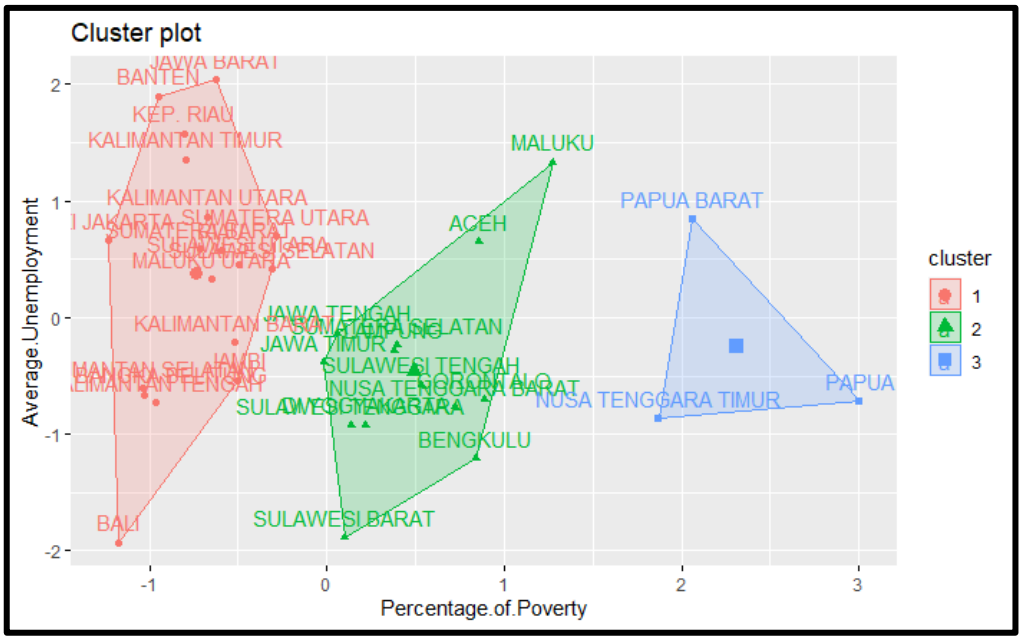

Fig. 7. Cluster Plot

Table 3. Cluster Member

Cluster	Province
1	Sumatera Utara, Sumatera Barat, Riau, Jambi, Kepulauan Bangka Belitung, Kepulauan Riau, DKI Jakarta, Jawa Barat, Banten, Bali, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Selatan, and Maluku Utara.
2	Aceh, Sumatera Selatan, Bengkulu, Lampung, Jawa Tengah, DI Yogyakarta, Jawa Timur, Nusa Tenggara Barat, Sulawesi Tengah, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, and Maluku.
3	Nusa Tenggara Timur, Papua Barat, and Papua.

Furthermore, the interpretation of the cluster profile includes the study of the centroid, namely the average value of objects in the cluster obtained for each variable [13]. The interpretation of the profile is obtained from calculating the average value of the object against the three clusters formed by analyzing the variables that differentiate between the three clusters. The average results of each cluster are in **Table 4**, following:

Table 4. Average of K-Means Cluster

Cluster	Average	
	Percentage of Poverty	Percentage of unemployment
1	6.246	5.034
2	13.251	3.642
3	23.597	4.003

The characteristics of each cluster are different, it can be seen from the average results in **Table 4.** From the smallest unit value to the largest unit value of each variable, the smallest average unit is obtained as low, medium, and high values for the largest average unit.

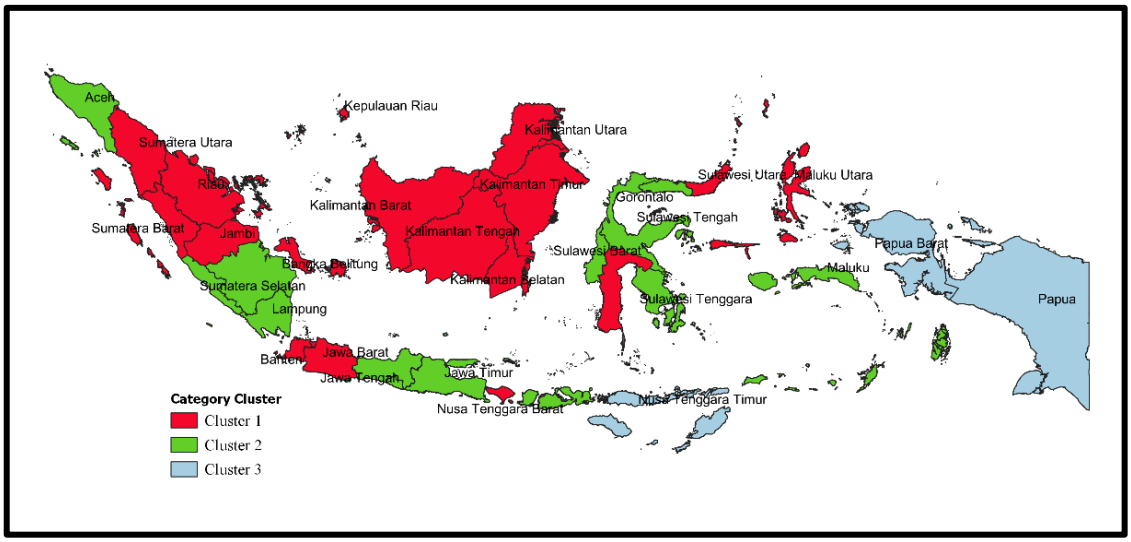

Fig. 8. Cluster Distribution Maps

In general, provinces that are included in cluster 1 (Sumatera Utara, Sumatera Barat, Riau, Jambi, Kepulauan Bangka Belitung, Kepulauan Riau, DKI Jakarta, Jawa Barat, Banten, Bali, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Selatan, and Maluku Utara) are areas with the characteristics of the lowest percentage of poverty and the highest Percentage of unemployment. The distribution of cluster 1 can be seen on the map in **Fig 8**. which is shown in red.

Provinces that are included in cluster 2 (Aceh, Sumatera Selatan, Bengkulu, Lampung, Jawa Tengah, DI Yogyakarta, Jawa Timur, Nusa Tenggara Barat, Sulawesi Tengah, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, and Maluku) are areas with the characteristics of the percentage of moderate poverty and the lowest Percentage of unemployment. The distribution of cluster 1 can be seen on the map in **Fig 8**. which is shown in green.

Furthermore, provinces included in cluster 3 (Nusa Tenggara Timur, Papua Barat, and Papua) are areas with the characteristics of the highest percentage of poverty and moderate unemployment. The distribution of cluster 2 can be seen on the map in **Fig 8**. which is shown in blue.

4. Conclusion

Based on the research about grouping provinces by factors that influenced criminal acts in Indonesia in 2019, the conclusion is:

1. Based on the results of clustering using the K-Means method obtained three clusters which are cluster group 1 includes 18 provinces, cluster group 2 includes 13 provinces, and cluster group 3 includes 3 provinces.
2. Cluster 1 is areas with the characteristics of the lowest percentage of poverty and the highest percentage of unemployment. Cluster 2 is areas with the characteristics of the moderate poverty percentage and the lowest percentage of unemployment. While Cluster 3 is areas with the characteristics of the highest percentage of poverty and moderate percentage of unemployment.

References

[1] A. Chazawi, Pelajaran Hukum Pidana I, Jakarta: PT Raja Girafindo, 2005.

[2] R. Pasiza, S. Nugroho and F. Faisal, "Analisis Jalur Faktor-Faktor Penyebab Kriminalitas di Indonesia," 2008. [Online]. Available: http://sigitnugroho.id/e-Skripsi/2015/08/Analisis%20Jalur%20Faktor-faktor%20Penyebab%20Kriminalitas%20di%20Indonesia.pdf.

[3] K. C. A. Handayani, R. N. Isfahani and E. Widodo, "Faktor-Faktor yang Mempengaruhi Kriminalitas di Indonesia Tahun 2011-2016 dengan Regresi Data Panel," *Indonesian Journal of Applied Statistics*, vol. 2, 2019.

[4] U. T. Suryadi and Y. Supriatna, "Sistem Clustering Tindak Kejahatan Pencurian di Wilayah Jawa Barat Menggunakan Algoritma K-Means," *Jurnal Teknologi Informasi dan Komunikasi*, 2019.

[5] W. Astuti and D. A. Widodo, "Pemetaan Tindak Kejahatan Jalanan di Kota Semarang Menggunakan Algoritma K-Means Clustering," *Jurnal Teknik Elektro, Universitas Negeri Semarang*, vol. 8, 2016.

[6] J. J. Purnama, R. Nurfalah, S. Rahayu and H. B. Novitasari, "Analisa Algoritma K-Means Clustering Pemetaan Jumlah Tindak Pidana," *Kumpulan Jurnal Ilmu Komputer (KLIK)*, pp. 128-142, 2019.

[7] Badan Pusat Statistik, 2019. [Online]. Available: https://www.bps.go.id/indicator/34/101/1/jumlah-tindak-pidana-menurut-kepolisian-daerah.html. [Accessed 24 April 2021].

[8] E. Prasetyo, Data Mining : Konsep dan Aplikasi Menggunakan METLAB, Yogyakarta: Andi, 2012.

[9] A. widarjono, Analisis Statiistika Multivariat Terapan, Yogyakarta: UPP STIM YKPN, 2010.

[10] Suparto, "Analisis Korelasi Variabel-Variabel yang Mempengaruhi Siswa dalam Memilih Perguruan Tinggi," *Jurnal IPTEK*, vol. 18, 2016.

[11] M. Syakur, B. Khotimah, S. Rochman and B. Satoto, "Integration K-means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluser," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 336, 2018.

[12] M. P. Frushicheva, 11 August 2016. [Online]. Available: https://rstudio-pubs-static.s3.amazonaws.com/201598_e96ae3be88b64ba8baffb2923bfdf5c6.html.

[13] S. Nugroho, Statistika Multivariat Terapan, Pertama ed., UNIB Press, 2008.

p-ISSN 2798-253X
e-ISSN 2798-3153