BAREKENG: Journal of Mathematics and Its Applications
June 2023 Volume 17 Issue 2 Page 0717–0726
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol17iss2pp0717-0726

COMPARISON OF K-MEANS AND GAUSSIAN MIXTURE MODEL IN PROFILING AREAS BY POVERTY INDICATORS

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

[1,2]*Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Islam Indonesia*
Kaliurang Street Km 14.5, Sleman, Yogyakarta, 55584, Indonesia

*Corresponding author's e-mail: * **dina.t.utari@uii.ac.id***

Article History:

Received: 26th November 2022
Revised: 8th April 2023
Accepted: 12th April 2023

Keywords:

Clustering index;
Gaussian Mixture Model;
K-Means;
Poverty.

ABSTRACT

The Covid-19 pandemic has led to income degradation of the Indonesian population, which potentially triggers poverty. According to the Indonesian Central Statistics Agency, the Province of Central Java is one of the areas that is most affected by Covid-19, especially on the economic aspect. In 2020, the percentage of poor people has increased by 0.6% from 2019. If this condition is ignored for the long term, it will have a negative impact on hampering national development. As a first step in designing a strategy for mitigating the impact of poverty, it is necessary to carry out an appropriate profiling of the areas affected by the economic aspect based on poverty indicators. This study compares the K-Means Clustering and Gaussian Mixture Model (GMM) in providing the best data grouping based on clustering indices, including connectivity, Dunn, and silhouette. GMM is a generalization of K-Means clustering to include information about the covariance structure of the data as well as latent Gaussian centers. We used poverty indicators data from the Central Statistics Agency of Central Java, such as poverty line, percentage of poor population, poverty depth index, and poverty severity index. The results obtained from this study indicate that the GMM gives the best results with the 3 clusters, with the number of members for the first, second, and third being 10, 19, and 6, respectively.

How to cite this article:

Z. Wahidah and D. T. Utari., "COMPARISON OF K-MEANS AND GAUSSIAN MIXTURE MODEL IN PROFILING AREAS BY POVERTY INDICATORS," *BAREKENG: J. Math. & App.*, vol. 17, iss. 2, pp. 0717-0726, June, 2023.

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

Various indicators influence poverty in different areas. Appropriate measurement of poverty can help in knowing the number of poor people, distribution, and conditions of poverty. The method of calculating the poor population carried out by BPS uses the basic needs approach. Based on this approach, poverty is seen as an inability from an economic standpoint, so poverty status is measured according to the poverty line. According to the basic needs approach, the Head Count Index (HCI) indicator is used. Apart from the headcount index (P0), other indicators are used to measure poverty levels, namely the poverty gap index or P1 and the distributionally sensitive index or P2 formulated by Foster-Greer-Thorbecke [1]. This method is the basis for calculating the percentage of poor people in all districts or cities.

When viewed by island, the percentage of poor people on the island of Java in 2020 the number of poor people in Java is 14.05 million people. This number shows that over half of Indonesia's poor population is on Java. The increase in the number of poor people in Java is because the area of Java has found many cases of Covid-19 compared to other islands in Indonesia [2]. Central Java is one of the provinces in Java that has been most affected by Covid-19. In 2020 the percentage of poor people was 11.41%. This percentage had increased from the previous 2019 when the percentage of poor people was 10.80%. According to the Socio-Demographic Survey on the impact of Covid-19, almost 50% of respondents in the low-income group (<1.8 million) said they had experienced a decline in income. This decrease causes poverty to increase because more and more people have an average expenditure below the poverty line. This condition will undoubtedly be a big challenge for the Central Java Provincial government to overcome the increasing poverty rate.

The poverty that occurs in a region in the long term will have an impact on hampering national development. The government needs to get an overview of the poverty of each district/city in Central Java to adopt poverty alleviation policies. In order to support the successful implementation of development programs to reduce poverty in Central Java Province, a study is needed to classify districts/cities in Central Java with almost the same or homogeneous characteristics or characteristics of poverty. As a solution to dig up poverty description information, one of the methods that can be used is clustering. Clustering aims to group data with the same characteristics into one group. With this grouping, the position of data distribution in actual conditions and finding a solution to a problem. One clustering method is the K-Means and the Gaussian Mixture Model. K-Means is a non-hierarchical cluster analysis that seeks to divide data with the same characteristics into one cluster. The K-means algorithm is performed by minimizing the sum squares distance between the data of each cluster center (centroid-based). Meanwhile, the Gaussian Mixture Model is a method that assumes that each Gaussian distribution number represents a cluster. A combination of means and variance will represent each Gaussian.

Several previous studies on poverty, as in [3] – [5], used K-Means and Average Linkage to map the characteristics of each group formed based on the value of each poverty indicator. Meanwhile, research that compares the performance of K-Means and GMM can be seen in [6]. The results of this study indicate that the GMM algorithm is superior to the K-Means algorithm based on the accuracy and speed of computation.

Based on the results of previous studies, researchers want to use the K-Means and GMM algorithms for grouping poverty data. The use of the GMM algorithm is relatively new for poverty indicator data. This study aims to classify poverty based on districts/cities in Central Java Province in 2020 using the K-Means algorithm and the Gaussian Mixture Model (GMM). Furthermore, profiling of the cluster results was carried out to map poverty in Central Java.

2. RESEARCH METHODS

2.1 Cluster Assumption

There are two assumptions that must be fulfilled in cluster analysis [7], which are as follows.

1) Representative of the Sample
   A representative sample is a sample with the same characteristics as the population. Using a representative sample will provide maximum results and be under the conditions of the existing population. If the research uses population data, it can be concluded that representative assumptions are met [7].

Another way to see whether a sample is representative is to use the Kaiser-Meyer Olkin (KMO) test. The KMO is conducted to see whether the sample is representative of the existing population so that the clustering or grouping process can be carried out correctly. This KMO test measures sample adequacy for each indicator. The KMO has a value of 0 to 1. If the KMO value is more than 0.5, the sample can be said to represent the population or a representative sample [8]. The following equation describes the KMO test [7].

$$KMO = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2}{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2 + \sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}^2} \quad (1)$$

where: p *is* the number of variables, r_{ij} is correlation between variables i and j, and a_{ij} is partial correlation between variables i and j.

2) Impact of Multicollinearity
The assumption in clusters is that there is no multicollinearity between variables. One way to determine the presence of multicollinearity is to look at the VIF value,

$$VIF = \frac{1}{(1 - R_i^2)} \quad (2)$$

where R_i^2 is coefficient of determination If the VIF value exceeds 10, it can be concluded that there is multicollinearity among variables [9].

2.2 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a solution in cluster analysis if multicollinearity occurs in the data. PCA aims to reduce variables to fewer than the previous number of variables, where the number of new variables will be less than the original variables. The principal component (PC) is a linear combination of the original variables. The formation of the principal component is based on two methods, namely the covariance matrix and the correlation matrix [10]. The stages of PCA are as follows.

1) Create an M matrix that contains data from variable X that has been standardized.
2) Make a correlation matrix from M, namely $M'M$. Principal component reduction begins by finding the eigenvalues obtained from the equation:

$$|M'M - \lambda I| = 0 \quad (3)$$

The number of selected principal components is based on the eigenvalue (λ). The number of principal components selected is the value of $\lambda > 1$ [11].

2.3 Determination of the Optimum Number of Clusters

There are several approaches to determining the optimum number of clusters: the connectivity index, Dunn index, and silhouette index. The formula for each of these indices is as follows [12].

1) Connectivity Index

$$Conn\ (C) = \sum_{i=1}^{N}\sum_{j=1}^{L} X_{i,nni(j)} \quad (4)$$

where $nni(j)$ is the closest neighbor observation, i to j, and L is a parameter that determines the number of neighbors contributing to connectivity measurements.

2) Dunn Index
Dunn index is the ratio of the smallest distance between observations in different clusters with the most significant distance in each data cluster.

$$D = \frac{min_{1\le i<j\le n} d(i,j)}{max_{1\le k\le n} d'(k)} \quad (5)$$

where $\boldsymbol{i}$, $\boldsymbol{j}$, and $\boldsymbol{k}$ are indices for each cluster, $\boldsymbol{d}$ measures the distance between clusters, and $\boldsymbol{d'}$ measures the differences between clusters.

3) Silhouette Index
Silhouette index used to measure the confidence level in the clustering process. The clustering results are said to be good if the index value is close to 1 and vice versa if the index value is close to -1.

$$(i) = \frac{bi-ai}{\max(bi,ai)} \quad (6)$$

where ***ai*** is the average distance between ***i*** and other data in the same cluster, and ***bi*** is the average distance between ***i*** and other data in different clusters.

2.4 K-Means Algorithm

The K-Means algorithm was first proposed by McQueen (1967) [13] and developed by Hartigan and Wong in 1979 [14], which aims to divide M data points in N dimensions into several k clusters. The clustering steps using the K-Means algorithm are as follows [15].

1) Determine the number of clusters;
2) Randomly allocate the initial and centroid of the cluster;
3) Find the distance for each centroid using the Euclidean distance with the formula

$$d_{ij} = \sqrt{\sum_{i=1}^{n}(x_{ik} - c_{jk})^2}; \quad (7)$$

4) Calculate the new centroid of the average data in each cluster

$$C_{kj} = \frac{x_{1j}+x_{2j}+x_{nj}}{n}; \quad (8)$$

5) Allocate each data to the nearest centroid,

$$a_{ij} = \begin{cases} 1, & s = \min\{d(x_i, C_{kj})\}; \\ 0, & other. \end{cases} \quad (9)$$

6) If data is still moving clusters, return to step 3.

2.5 Gaussian Mixture Model Algorithm

McLachlan and Basford (1989) provide an approach by paying attention to data distribution, namely model-based analysis [16]. The model-based clustering method is a cluster group algorithm using statistical analysis to analyze group results. The model-based clustering method assumes that the data is generated by a mix of probability distributions, with each component representing a different cluster. If the model is a mixture of G Gaussian components, it is called the Gaussian Mixture Model. The Gaussian mixture model is a method that assumes that each Gaussian distribution number represents a cluster. A combination of means and variance will represent each Gaussian. The purpose of grouping using the Gaussian mixture model is to find the model parameters (mean and covariance matrices of each distribution and weight) so that the resulting model best fits the data.

Fraley and Raftery (2003) identified several models used to group data with various geometric properties obtained through Gaussian components with different parameters [17], as seen in **Table 1.** Characteristics of geometric distribution (orientation, volume, and shape) are obtained from various shapes groups or limited to the same group. The variance matrix for all components can be equal or variance.

Table 1. Covariance matrix and geometric interpretation of MCLUST in the multivariate Gaussian mixture Model

Symbol	Model	Volume	Geometry shape	Orientation	Shape
EII	λI	Same	Same	-	Spherical
VII	$\lambda_k I$	Different	Same	-	Spherical
EEI	λA	Same	Same	Coordinate axes	Diagonal
VEI	$\lambda_k A$	Different	Same	Coordinate axes	Diagonal
EVI	λA_k	Same	Different	Coordinate axes	Diagonal
VVI	$\lambda_k A_k$	Different	Different	Coordinate axes	Diagonal
EEE	λDAD^T	Same	Same	Identity	Ellipsoidal
EEV	$\lambda D_k A D_k^T$	Same	Same	Different	Ellipsoidal
VEV	$\lambda_k D_k A D_k^T$	Different	Same	Different	Ellipsoidal
VVV	$\lambda_k D_k A_k D_k^T$	Different	Different	Different	Ellipsoidal

In one dimension, only two models are available: E for the same variance and V for different variances. For more than one dimension, the geometric characteristics of the model are identified. For example, in the EVI model, where the volume of all clusters is the same (E), the shape of the clusters varies (V), and the orientation is identity (I). clusters with the EVI model have a diagonal covariance and orientation parallel to the coordinate axes.

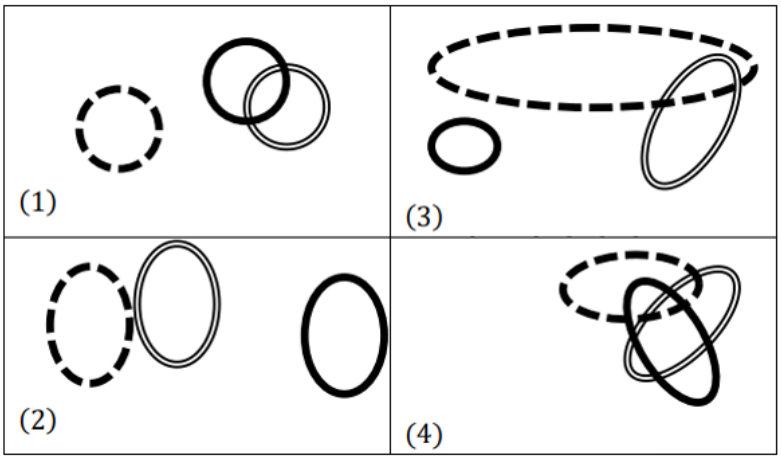

Figure 1. Illustration of cluster forms based on the variance covariance matrix in the MCLUST package [18]

The explanation of **Figure 1** is as follows.

1) $\sum_k = \sum = \lambda I$ (EII) produces all spherical clusters with the same volume between clusters.
2) $\sum_k = \sum = \lambda DAD^T$ (EEE) produces clusters with the same shape, volume, and orientation.
3) $\sum_k = \sum = \lambda_k D_k A_k D_k^T$ (VEV) produces clusters that differ in shape, volume, and orientation.
4) $\sum_k = \sum = \lambda_k D_k A D_k^T$ (VVV) produces clusters that differ only in orientation.

To get the best results, the thing that must be done is to maximize the possibilities of the data from the GMM model. This can be achieved using the expectation maximization (EM) algorithm. In each iteration using the EM algorithm, there are two stages: the expectation stage (E-Step) and the maximization stage (M-Step).

The clustering steps using the EM algorithm in the Gaussian Mixture Model are as follows.

1) Initialize $\boldsymbol{\mu_k}$, $\boldsymbol{\sigma_k}$, and $\boldsymbol{\pi_k}$ values randomly for all clusters. $\boldsymbol{\mu}$ is the mean, $\boldsymbol{\sigma}$ is the variance, $\boldsymbol{\pi}$ is the coefficient of the mixture, $\boldsymbol{k}$ is the number that refers to a mixture in the Gaussian distribution, and the equivalent $\boldsymbol{k}$ is the value that refers to a cluster.
2) E-Step: Evaluate the log-likelihood results using the parameters $\boldsymbol{\mu_k}$, $\boldsymbol{\sigma_k}$, and $\boldsymbol{\pi_k}$. Suppose cluster $\boldsymbol{C_k}$ is represented by a Gaussian distribution ($\boldsymbol{\mu_k}$, $\boldsymbol{\sigma_k}$), then the probability of $\boldsymbol{X_i}$ belonging to cluster $\boldsymbol{C_k}$ is calculated from the equation:

$$z_{ik} \;/\; \rho(C_k|x_i) = \frac{\rho(x_i|C_k)\,\rho(C_k)}{\rho(x_i)} \tag{10}$$

Then calculate the likelihood value and evidence:

$$(x_i|C_k) = \frac{1}{\sqrt{2\pi_k\sigma}} \exp\left(\frac{-(x_i-\mu k)^2}{2\sigma^2}\right) \tag{11}$$

$$\rho(x_i) = \sum_k \rho(x_i|C_k)\,\rho(C_k) \tag{12}$$

3) M-Step: Change the value of $\boldsymbol{\mu_k}$, $\boldsymbol{\sigma_k}$, $\boldsymbol{\rho(C_k)}$ by calculating with the following equations:

$$\mu_k = \frac{\sum_i (C_k|x_i)\,x_i}{\sum_i \rho(C_k|x_i)} \tag{13}$$

$$\sigma_k = \frac{\sum_i (C_k|x_i)\,(x_i-\mu_k)^2}{\sum_i \rho(C_k|x_i)} \tag{14}$$

$$\pi_k = \frac{\sum_i (C_k|x_i)}{n} \tag{15}$$

4) Repeat steps 2 and 3 until the convergence criteria are met. For convergence, determine specific threshold values for changes in means and variance in successive iterations, so that cluster members can be grouped using the Maximum a Posteriori (MAP) classification method with the following conditions:

$$MAP\ \{\hat{z}_{ik}\} = \begin{cases} 1 & if\ max\{\hat{z}_{ik}\}\ \epsilon\ ke\ k \\ 0 & other \end{cases} \quad (16)$$

The selection of the best model in the Gaussian Mixture Model (GMM) method uses a commonly used approach, Bayes Information Criterion (BIC). Fraley and Raftery (1998) took a mixed model approach through the Bayes factor (BIC) with a systematic selection for model parameterization and the number of groups [19]. Generally, the greater the BIC value, the more substantial the evidence for the best model and number of clusters. The equation can obtain the value for BIC:

$$2\log P(y|M_k)\ \approx 2\log P\left(y\middle|\hat{\theta}_k, M_k\right) - V_k \log(n) = BIC_k \quad (17)$$

where:

$P(y|M_k)$: integration of likelihoods for M_k model,

$P\left(y\middle|\hat{\theta}_k, M_k\right)$: integrated maximum mixed likelihood for M_k model,

V_k : the number of independent parameters estimated in the M_k model.

3. RESULTS AND DISCUSSION

3.1. Materials

The data source in this study came from the Central Bureau of Statistics (BPS) official website of Central Java [20]. At the same time, the type of data is secondary data from BPS, which refers to concepts in the Handbook on Poverty and Inequality published by the World Bank. We calculate the poor population using the basic needs approach by the BPS.

This study uses poverty indicators which consist of 4 variables, namely poverty line (GK), percentage of poor population (P0), poverty depth index (P1), and poverty severity index (P2). Since each variable has a different unit, the data is standardized using the Z-score method before carrying out the cluster analysis.

3.2. Cluster Assumptions

Based on the results of the KMO test, a KMO value of 0.53 was obtained, which exceeded the threshold. It means that the sample represents the population or a representative sample so that the analysis can proceed to the next stage. Next is the multicollinearity test, and we used VIF to evaluate each variable.

Table 2. VIF value

Variables	VIF
GK	1.783
P0	14.845
P1	85.700
P2	45.352

Based on the VIF value of each variable in Table 2, there is a VIF value of more than 10. Thus, there is an indication of multicollinearity in the independent variables in the data used. Therefore, it is necessary to do PCA before cluster analysis to overcome this condition.

3.3. Overcoming Multicollinearity with PCA

PCA aims to reduce variables to fewer than the previous number of variables, where the number of new variables will be less than the old variables. The principal component (PC) is a linear combination of the original variables. The number of selected principal components is seen based on the eigenvalue (λ) obtained from Equation (3), and the resulting eigenvalues are 2.820, 1.041, 0.132, and 0.007. Based on those results, eigenvalues of more than one are found in factors 1 to 2, which means that the number of factors to be formed is two factors (PC1 and PC2).

The next step is to retest the cluster assumption using the result of PCA, and it was found that all cluster assumptions had been fulfilled.

3.4. Determination of the Optimum Number of Clusters

Several approaches to a clustering algorithm intend to find the best number of clusters. This study used the connectivity, Dunn, and silhouette indices, which were calculated based on equations 4-6. The best number of clusters is determined based on the smallest connectivity index value, the most extensive Dunn index, and the silhouette index close to 1. **Table 3** shows the results of several index values obtained from the district/city poverty indicators in Central Java in 2020.

Table 3. Index value comparison

Algorithm	Index	Number of Cluster		
		3	4	5
K-Means	Connectivity	**13.346**	16.461	22.259
	Dunn	0.181	**0.282**	0.216
	Silhouette	**0.485**	0.458	0.393
GMM	Connectivity	**12.348**	19.260	27.868
	Dunn	**0.243**	0.208	0.176
	Silhouette	**0.497**	0.425	0.333

We compare the number of clusters based on the respective index criteria applied to both algorithms. It was found that almost all indices were fulfilled for the best number of clusters of three.

3.5. K-Means Clustering

The K-means algorithm starts with choosing k, where k is the number of clusters that want to form. Then set the value of k; temporarily, the value becomes the center of the cluster or can be referred to as the centroid. The value of k is 3, and the initial centroid value was selected, as shown in **Table 4** below.

Table 4. Initial centroid

Cluster	Centroid PC1	Centroid PC2
1	-3.320	-1.789
2	0.436	-0.164
3	2.792	1.936

After selecting the initial centroid randomly, the Euclidean distance is calculated with **Equation (7)**, and after getting the cluster results in the first iteration, proceed with the second iteration's calculation with a new centroid derived from the average value of the data in each cluster. In this study, five iterations were carried out until there were no more data-moving clusters.

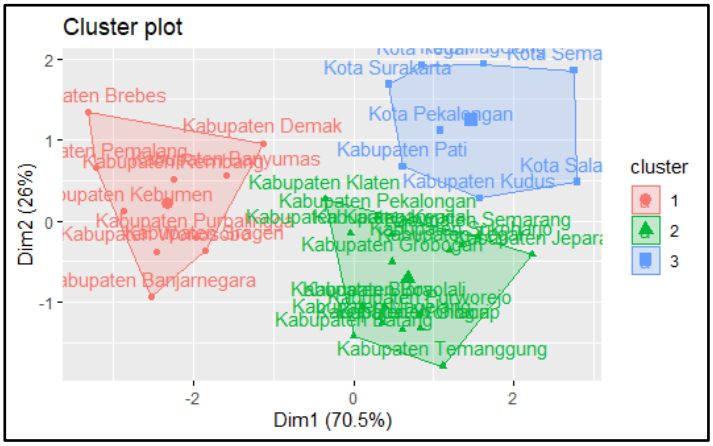

Figure 2. K-Means result plot

Based on the above results using the K-means algorithm, it is found that Cluster 1 includes ten districts/cities, Cluster 2 includes 17 districts/cities, and Cluster 3 includes eight districts/cities, each of which can be seen in **Figure 2** above.

3.6. GMM Clustering

In data clustering using the GMM, nine models were identified to group data with various geometric properties, which can be seen in **Table 1**. The best model can be determined based on the BIC. This study uses R with the help of the MCLUST package, which provides nine models with several components from 1 to 9. The best model is EII, with an optimal number of components of 3, as shown in **Figure 3** below.

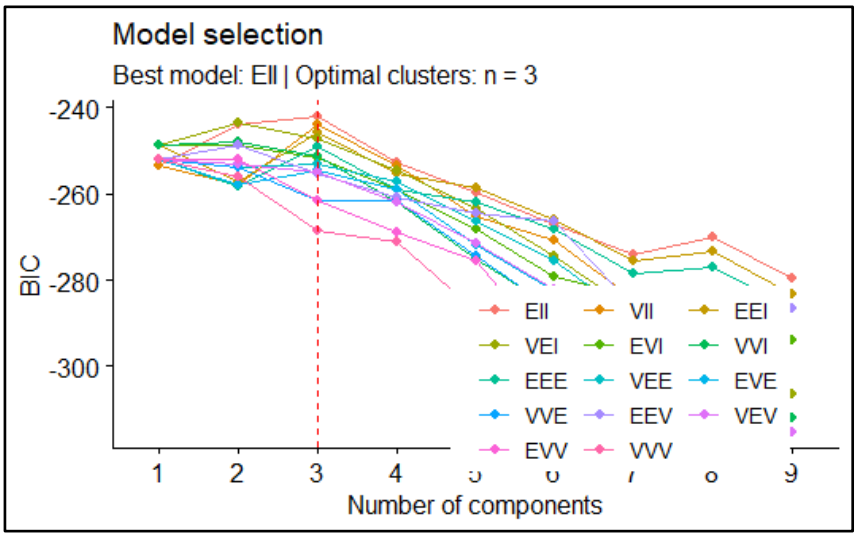

Figure 3. BIC value of GMM result

The GMM in R moves based on the Expectation Maximization (EM) algorithm. The final result is a mixing proportion, means vector, and covariance matrix.

Mixing Proportion:

Cluster 1	Cluster 2	Cluster 3
0.287	0.517	0.196

Means Vector:

Cluster 1

$$\begin{matrix} PC1 \\ PC2 \end{matrix} = \begin{bmatrix} -2.309 \\ 0.225 \end{bmatrix}$$

Cluster 2

$$\begin{matrix} PC1 \\ PC2 \end{matrix} = \begin{bmatrix} 0.701 \\ -0.624 \end{bmatrix}$$

Cluster 3

$$\begin{matrix} PC1 \\ PC2 \end{matrix} = \begin{bmatrix} 1.529 \\ 1.316 \end{bmatrix}$$

Covariance Matrix:

Cluster 1

$$\begin{matrix} \\ PC1 \\ PC2 \end{matrix} \begin{matrix} \begin{matrix} PC1 & PC2 \end{matrix} \\ \begin{bmatrix} 0.477 & 0 \\ 0 & 0.477 \end{bmatrix} \end{matrix}$$

Cluster 2

$$\begin{matrix} \\ PC1 \\ PC2 \end{matrix} \begin{matrix} \begin{matrix} PC1 & PC2 \end{matrix} \\ \begin{bmatrix} 0.477 & 0 \\ 0 & 0.477 \end{bmatrix} \end{matrix}$$

Cluster 3

$$\begin{matrix} \\ PC1 \\ PC2 \end{matrix} \begin{matrix} \begin{matrix} PC1 & PC2 \end{matrix} \\ \begin{bmatrix} 0.477 & 0 \\ 0 & 0.477 \end{bmatrix} \end{matrix}$$

Then we also get the probability data for each cluster, and we get Cluster 1 covering 10 districts/cities, Cluster 2 covering 19 districts/cities, and Cluster 3 covering 6 districts/cities, which can be seen in **Figure 4** below.

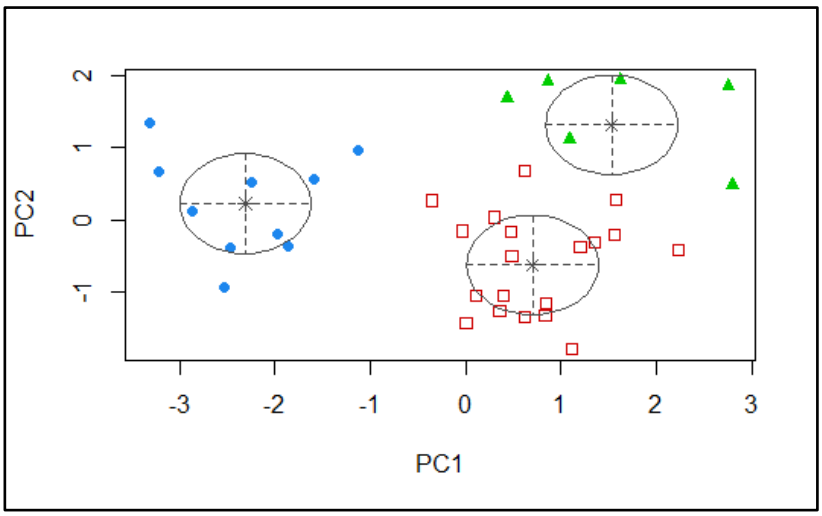

Figure 4. GMM result plot

In **Figure 4**, members of Cluster 1 are indicated by blue dots, Cluster 2 is colored red, and Cluster 3 is colored green.

After carrying out the clustering analysis, it is continued by looking at the best method of cluster analysis using the K-Means algorithm and the Gaussian Mixture Model. This study uses three indices, connectivity, Dunn, and silhouette, as shown in **Table 3** above. With a total of 3 clusters, all indices show that the GMM algorithm produces better cluster analysis results than the K-Means algorithm.

3.7. Cluster Outcome Profiling

After performing cluster analysis using the K-Means and GMM algorithms, the best results were obtained using GMM. Furthermore, cluster profiling is carried out for the best results by looking at the average of each cluster.

Table 5. Average of each variable

Cluster	GK	P0	P1	P2
1	384,948.400	15.432	2.453	0.537
2	379,241.895	10.016	1.178	0.206
3	491,534.833	6.810	1.082	0.252

Based on **Table 6**, the characteristics of each cluster are different. From the smallest unit value to the most significant unit value for each variable, the smallest average unit value is obtained successively as low, medium, and high values for the largest average unit. Green indicates low, yellow indicates medium, and red indicates high.

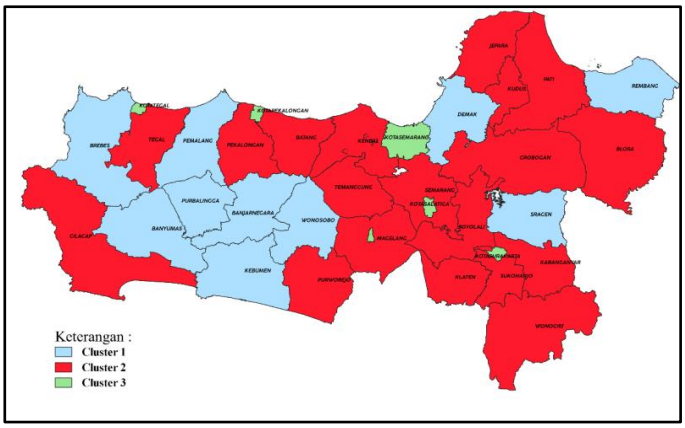

Figure 5. Cluster result mapping of province of Central Java using Gaussian Mixture Model

In general, the districts/cities included in Cluster 1 are groups with a moderate poverty line, a high percentage of poor people, a poverty depth index, and a poverty severity index. The distribution of Cluster 1 can be seen on the map in **Figure 5**, which is shown in blue. Districts included in Cluster 2 are groups with a low poverty line and poverty severity index, an average percentage of poor people, and a poverty depth index—the distribution of Cluster 2 groups is shown in red. Furthermore, the districts included in Cluster 3 are groups with a high poverty line, a low percentage of poor people, a poverty depth index, and a moderate poverty severity index. The distribution of Cluster 3 groups is shown in green.

4. CONCLUSIONS

The fact that GMM provides estimates of the likelihood that each data point belongs to each cluster is one of their key advantages. Compared to the solo cluster assignment that most other clustering algorithms offer, this offers much more contextual information. The advantage of GMM models over others, such as K-Means clustering, is that they do not presuppose all clusters have sphere-like shapes. Instead, clusters with different shapes can be accommodated using GMM.

Based on a comparison between the K-Means and GMM, all clustering indices (connectivity, Dunn, silhouette) show the best clustering results with GMM, with the number of clusters being 3.

REFERENCES

[1] T. Tambunan, *Perekonomian Indonesia (Teori dan Temuan Empiris)*. Jakarta: Ghalia Indonesia, 2001.
[2] F. A. Hafiez, "Ini 5 Provinsi Penyumbang Kasus Covid-19 Terbanyak," *medcom.id*, Mar. 02, 2022.
[3] N. I. Febianto and N. D. Palasara, "Analisis Clustering K-Means pada Data Informasi Kemiskinan Di Jawa Barat Tahun 2018," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 8, no. 2, pp. 130–140, 2019.
[4] K. Aprilia and F. Sembiring, "Analisis Garis Kemiskinan Makanan Menggunakan Metode Algoritma K-Means Clustering," in *Seminar Nasional Sistem Informasi dan Manajemen Informatika*, 2021, pp. 1–10.
[5] D. Widyadhan, R. B. Hastuti, I. Kharisudin, and F. Fauzi, "Perbandingan Analisis Klaster K-Means dan Average Linkage untuk Pengklasteran Kemiskinan di Provinsi Jawa Tengah," in *PRISMA: Prosiding Seminar Nasional Matematika*, 2021, pp. 584–594.
[6] S. A. Prabawa, "Perbandingan Algoritma K-Means dan Gaussian Mixture Model untuk Pengelompokan Berita pada Kompas.com," Universitas Multimedia Nusantara, Tangerang, 2021.
[7] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis*, 7th Edition. New York City: Pearson Education Limited, 2013.
[8] S. Yamin, L. A. Rachmach, and H. Kurniawan, *Regresi dan Korelasi dalam genggaman Anda: Aplikasi dengan Software SPSS, Eviews, MINITAB, dan STATGRAPHICS*. Jakarta: Salemba Empat, 2011.
[9] J. I. Daoud, "Multicollinearity and Regression Analysis," *Journal of Physics: Conference Series 949*, pp. 1–6, 2017.
[10] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis 6th edition*. United States of America: Pearson Education Inc., 2007.
[11] J. F. J. Hair, W. C. Black, B. J. Babin, R. E. Anderson, and R. L. Tatham, *Multivariate Data Analysis 6th edition*. New Jersey: Pearson Education, 2006.
[12] E. Irwansyah and M. Faisal, *Advanced Clustering: Teori dan Aplikasi*. Yogyakarta: Deepublish, 2015.
[13] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1967, pp. 281–297.
[14] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," *J R Stat Soc Ser C Appl Stat*, vol. 28, no. 1, pp. 100–108, 1979.
[15] M. Wahyudi, Masitha, R. Saragih, and Solikhun, *Data Mining: Penerapan Algoritma K-Means Clustering dan K-Medoids Clustering*. Medan: Yayasan Kita Menulis, 2020.
[16] G. L. McLachlan, K. E. Basford, and M. Dekker, "Mixture Models: Inference and Applications to Clustering," *J Am Stat Assoc*, vol. 84, no. 405, pp. 337–338, 1989.
[17] L. Scrucca, "Identifying connected components in Gaussian finite mixture models for clustering," *Comput Stat Data Anal*, vol. 93, pp. 5–17, 2016.
[18] E. Genge, "Analysis of Massive Emigration from Poland: The Model-Based Clustering Approach," *Argumenta Oeconomica Cracoviensia*, vol. 16, pp. 37–49, 2017.
[19] N. Shen and B. Gonz´alez, "Bayesian Information Criterion for Linear Mixed-effects Models," 2021.
[20] BPS Provinsi Jawa Tengah, "Kemiskinan dan Ketimpangan," *BPS Provinsi Jawa Tengah*.