JURNAL STATISTIKA DAN APLIKASINYA
e-ISSN 2620 – 8369
VOLUME 8 (ISSUE 1) JUNE 2024: 17 - 30
https://doi.org/10.21009/JSA.08102

CHARACTERISTICS OF PROVINCES IN INDONESIA BASED ON JKN INDICATOR OUTCOMES BY GAUSSIAN MIXTURE MODEL WITH EXPECTATION-MAXIMIZATION ALGORITHM AND BIPLOT

Dania Siregar [1*], Widyanti Rahayu[2], Bintang Mahesa Wardana[3], Ketrin Natasya Stefany[4], Bayu Wibisono[5]

[1,2,3,4,5]Statistics Department, Faculty of Mathematics and Natural Sciences, Jakarta State University Rawamangun Subdistrict, Pulogadung District, East Jakarta 12330, Indonesia

*Corresponding author's e-mail: * **dania-siregar@unj.ac.id***

ABSTRACT

Article History:
Received: 27 April 2024
Revised: 7 June 2024
Accepted: 27 June 2024
Published: 30 June 2024
Available online.

Keywords:
Gaussian Mixture Model (GMM), Expectation Maximization (EM), National Health Insurance (JKN), Biplot Analysis, Provincial Clusters.

Indonesia, an archipelago with a population of 257.77 million in 2022, faces significant challenges in enhancing the quality of life to improve human resource productivity. This study aims to identify provincial characteristics in Indonesia based on the outcomes of the Jaminan Kesehatan Nasional (JKN) program from 2019 to 2021. Using a Gaussian Mixture Model (GMM) with the Expectation Maximization (EM) algorithm, we cluster 34 provinces based on 14 health indicators. The data were obtained from the BPJS website and included variables such as access to health services, program effectiveness, and service quality. Our methodology allows for clustering provinces with similar health outcomes and analyzing the unique indicators for each cluster using biplot analysis.

The results indicate significant variation in cluster membership across the years. In 2019, three clusters were identified, with cluster sizes of 16, 12, and 6 provinces. In 2020, the optimum model also had three clusters, but with different member distributions: 24, 7, and 3 provinces. By 2021, four clusters emerged with sizes of 9, 16, 3, and 6 provinces. These findings highlight the dynamic nature of health outcomes across Indonesia's provinces and suggest the need for tailored policy interventions to improve the JKN program's effectiveness.

The study's limitations include the reliance on available BPJS data and the assumption that the selected health indicators comprehensively represent the JKN program's impact. This research's novelty lies in its use of advanced clustering techniques to provide a nuanced understanding of regional health disparities in Indonesia, which can inform more targeted and effective health policies.

How to cite this article:
D. Siregar, W. Rahayu, B.M. Wardana, K.N. Stefany, B. Wibisono, "CHARACTERISTICS OF PROVINCES IN INDONESIA BASED ON JKN INDICATOR OUTCOMES BY GAUSSIAN MIXTURE MODEL WITH EXPECTATION-MAXIMIZATION ALGORITHM AND BIPLOT", Journal Statistika dan Aplikasinya, vol. 8, iss. 1, pp. 17 – 30, June 2024

Journal homepage: *https://journal.unj.ac.id/unj/index.php/statistika*
Journal e-mail: *jsa@unj.ac.id*
Research Article · Open Access

1. INTRODUCTION

Indonesia is an archipelago with a population that reached 257.77 million people in 2022, based on data from Statistics Indonesia. The population is dispersed among the provinces in Indonesia; in 2018, the number of provinces in Indonesia was 34, while in 2022 it had reached 36. This high population, if juxtaposed with the quality of productive human resources, will have great potential for the country's development. In order to improve the productivity and quality of human resources, increasing the quality-of-life expectancy becomes more essential and relevant in terms of economic productivity. Health development is expected to build productive human resources as an asset to the nation.

Various attempts to improve health quality, such as by conducting research, can be used as a basis to develop policies. This includes forming clusters by region in Indonesia from a study of health aspects to obtain certain characteristics as a basis for developing policies. Many studies have been conducted to see the characteristics of provinces in Indonesia in terms of health, including [1] using multiple linear regression models and cluster analysis with the hierarchical clustering method to analyze the characteristics of provinces in Indonesia based on environmental health variables, which produce provincial clusters with health indicators classified as low, medium, and high. Clustered provinces in Indonesia based on the prevalence of infectious and non-infectious diseases using multidimensional scaling (MDS) by [2] and Clustering of sub-districts based on health indicators from the Healthy Indonesia Program with a Family Approach, using the K-Medoids method produces five clusters with diverse characteristics that serve as recommendations for improving health quality by [3]. Meanwhile, clustered districts and cities in Central Java based on ownership of health insurance using the fuzzy c-means algorithm by [4].

According to [4], which used health insurance ownership data, the researchers are interested in knowing more about the characteristics of provinces in Indonesia when measured by the overall outcomes of the Jaminan Kesehatan Nasional (JKN) program. The ownership indicator is not the only indicator used to measure the outcomes of the JKN program; there are other indicators, including access to health services, program effectiveness, and service quality. This research has not been widely performed, even though it is helpful to identify the characteristics of clusters of provinces that require certain attention or improvement to optimize the outcomes of the JKN program. In addition, the formation of clusters from several different years can also provide a bigger picture of the impact of JKN program development in Indonesia. Another interesting point is that the analysis of indicators that are the main characteristics of each cluster has never been done.

Widely used clustering methods are distance-based methods such as k-means. If the clustering forms a circular pattern, then the k-means method works properly, but if the clustering forms a non-circular pattern, an oval, for example, then k-means will not work properly. Alternative clustering methods such as gaussian mixture models with expectation maximization (EM) algorithms can work properly for data with even oval cluster patterns, as they work on distribution-based models [5]. The K-means algorithm is performed by minimizing the sum square 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 [6].

Clustering techniques are categorized into two types, hard-clustering and soft-clustering. In soft-clustering, data points can belong to multiple clusters. An example of soft clustering is the Gaussian Mixture Model (GMM) algorithm. GMM is a probabilistic model that assumes all data points are generated from a mixture of some Gaussian distributions with unknown parameters [7].

The gaussian mixture model approach with the EM algorithm estimator can be used to cluster provinces, and biplots can be used to analyze the outcome indicators that are unique to each cluster. The EM algorithm is an iterative method that uses the Expectation stage (E-step) and the Maximization stage (M-step) to find the best possible local parameter value (maximum likelihood) in clustering to produce clusters with similar objects, whereas the biplot is used to observe the relationship between objects, the relationship between variables and objects, and the relationship between the variables themselves. Using these two statistical methods, it is expected to gain an overview of the features of Indonesian province clusters based on the JKN program development results.

2. METHODS

Material and Data

The data for this study was obtained from the BPJS (Badan Penyelenggara Jaminan Kesehatan) website. The data are indicators of JKN program results presented as standardized values. They objectively monitor JKN users' access to and usage of health services. The dataset consists of 34 provincial observations. The data includes 14 variables, each from 2019 to 2021: 1) information on the growth of participation per province; 2) the development of health service access rates per province, namely RJTP (rawat jalan tingkat pertama), RITP (rawat inap tingkat pertama), RJTL (rawat jalan tingkat lanjut), and RITL (rawat inap tingkat lanjut); 3) program efficacy, i.e., consumption rates that measure the number of visits (RJTP and RJTL), admission rates or people treated (RITP and RITL), and the number of treatment days (RITP, RITL); 4) facilities that cooperate with BPJS, namely: cooperative individual practitioners, cooperative private clinics, and cooperative FKRTL (fasilitas kesehatan tingkat lanjutan); 3) program efficacy, i.e., consumption rates that measure the number of visits (RJTP and RJTL), admission rates or people treated (RITP and RITL), and the number of treatment days (RITP, RITL); and 4) facilities that collaborate with BPJS, namely cooperative individual practitioners, cooperative private clinics, and cooperative FKRTL (fasilitas kesehatan tingkat lanjutan).

Research Method

Gaussian Mixture Model

Stauffer and Grimson [8] first introduced the Gaussian Mixture Model. This model is described as a set of density function components consisting of k components, where $k \in \{1, \dots, K\}$. In other words, the density function for the Gaussian Mixture Model is a gaussian density function that has $k \geq 2$. If there are more variables, it becomes a multivariate Gaussian density function, as defined by Bishop in his book [9].

$$p(\pi, \mu, \Sigma) = \sum_{k=1}^{K} \pi_k N(\mu_k, \Sigma_k), i = 1,2, \dots, n \quad (1)$$

where:

x_i: Point value of i-th object in the dataset

μ_k: Mean vector of the k-th mixture component

Σ_k: Covariance matrix of k-th mixture component

π_k: k-th mixture component

Maximum Likelihood Estimator Method

Bain and Engelhardt [10] described the maximum likelihood method as one of the techniques to evaluate unknown parameters. By estimating the equation of formula (1) using maximum likelihood, we obtain the following formulation:

$$L(\psi) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k N(\mu_k, \Sigma_k) \quad (2)$$

$$ln\ L(\psi) = \sum_{i=1}^{n} ln \sum_{k=1}^{K} \pi_k N(\mu_k, \Sigma_k) \quad (3)$$

Although the equation was derived from ψ, this process can be challenging both analytically and numerically since the likelihood function in the equation is not in closed form. Therefore, to resolve this problem, the expectation maximization algorithm is used.

Expectation Maximization Algorithm for GMM Estimation

According to the book by Ng, Xiang, and Yau [11], the Expectation Maximization Algorithm for estimating Gaussian Mixture Model is described as follows.

At $(k+1)th$ iteration of the EM algorithm, the E-Step computes the Q function, which represents the expectation of the complete log likelihood data conditioned on the observation x given the current fit for the parameter ψ. This can be expressed in the following equation:

$$Q(\psi, \psi^{(k)}) = \sum_{k=1}^{K} \sum_{i=1}^{n} \tau_k(x_i, \psi^{(k)})\{ln\pi_k + lnN(x_i|\mu_k, \Sigma_k)\} \quad (4)$$

Where $E_{\psi(k)}$ represents the expectation using the current vector of parameters$(\psi^{(k)})$. Besides, $\tau_k(x_i, \psi^{(k)})$ explained in the following equation:

$$\tau_k(x_i, \psi^{(k)}) = \frac{{\pi_k}^{(k)} lnN(\mu_k, \Sigma_k)}{\sum_{j=1}^{k} {\pi_j}^{(k)} lnN(\mu_j, \Sigma_j)} \quad (5)$$

$\tau_k(x_i, \psi^{(k)})$ is a posterior probability estimation that the i-th observation (x_i) is assigned to the k-th component of the mixture model based on the parameter $\psi^{(k)}(k = 1, \dots, K; i = 1, \dots, n)$. In the case of mixtures with normal component densities, in computing, it is suggested to perform in E-Step with adequate statistics [12] which formulates the following equation:

$$T_{k1}^{(k)} = \sum_{i=1}^{n} \tau_k(x_i, \psi^{(k)})$$
$$T_{k2}^{(k)} = \sum_{i=1}^{n} \tau_k(x_i, \psi^{(k)})x_i$$
$$T_{k3}^{(k)} = \sum_{i=1}^{n} \tau_k(x_i, \psi^{(k)})x_i x_i^T \quad (6)$$

Furthermore, M-Step will renew ψ estimation to $\psi^{(k+1)}$ that maximize the function Q with respect to ψ over the parameter space. For mixtures with normal component densities, M-Step is in closed form. Since the statistical basis is sufficient, the parameter renewal in M-Step can be expressed in the following equation:

$${\pi_k}^{(k+1)} = T_{k1}^{(k)}/n \quad (7)$$
$${\mu_k}^{(k+1)} = T_{k2}^{(k)}/T_{k1}^{(k)}$$
$${\Sigma_k}^{(k+1)} = \left\{T_{k3}^{(k)} - {T_{k1}^{(k)}}^{-1} T_{k2}^{(k)} {T_{k2}^{(k)}}^{T}\right\}/T_{k1}^{(k)}$$

Optimum Cluster Size Selection

Considering the number of K components in the Gaussian Mixture Model (GMM) that need to be determined, we can use the Bayesian Information Criteria (BIC). The BIC value formula can be formulated as follows:

$$BIC = qln(n) - lnL_c(\psi) \quad (8)$$

Where q is the number of parameters in the model, n is the number of data points, and $L_c(\psi)$ is the likelihood function of the maximum likelihood estimator ψ. The optimum cluster size can be determined by the largest BIC value among the other clusters.

Data Analysis Stages

The application of the EM algorithm for clustering will be performed using the R programming language with the support of the 'mclust' module [13]. In general, the stages of the analysis can be described as follows:

1. Import the 'mclust' module into the R workspace.
2. Import data into the R workspace, which contains 34 observations representing provinces in Indonesia along with 14 JKN outcome indicator variables for 2019–2021.
3. Data partitioning into three separate data sets based on year, i.e., 2019 data, 2020 data, and 2021 data, followed by pre-processing of all three data sets such that the format or form of all data in the R workspace fulfills the criteria required by the 'mclust' module.
4. Applying EM algorithm clustering to the three data years 2019, 2020, and 2021, respectively.
5. Selecting the most optimum cluster size as the basis for modeling by looking at the largest BIC value.
6. Analyzing the cluster results by examining the parameters of the resulting model.

7. Applying biplot analysis of the cluster results of each cluster in each year to obtain the characteristic variables of the clusters formed.
8. Evaluation and comparison of clustering results from 2019 to 2021 and their association with IPKM 2018.
9. Conclusion on the development outcomes of the 2019–2021 JKN program.

3. RESULTS

Prior to analyzing and evaluating the model, it is required to select the most optimum cluster size to use as the modeling basis. This is accomplished by selecting the outcomes of data modeling based on the highest BIC value. Figure 1 and Figure 2 show the estimated BIC values and their plots for the modeling results of the 2019 data at the numbers of clusters 1, 2, 3, 4, and 5. Similar results were found for 2020 and 2021; however, they are not presented in this paper.

```
Bayesian Information Criterion (BIC):
        EII       VII       EEI       VEI       EVI       VVI       EEE
1 -15340.07 -15340.07 -7542.924 -7542.924 -7542.924 -7542.924 -6809.221
2 -14544.05 -14207.05 -7455.875 -7401.930 -7305.283 -7348.000 -6833.658
3 -14176.30 -13812.98 -7439.871 -7348.814 -7311.124 -7347.675 -6849.525
4 -14074.54 -13604.62 -7392.658 -7283.192 -7377.187 -7314.333 -6871.367
5 -14023.72 -13377.99 -7253.537 -7216.082 -7314.452 -7249.265 -6838.967
        VEE       EVE       VVE       EEV       VEV       EVV       VVV
1 -6809.221 -6809.221 -6809.221 -6809.221 -6809.221 -6809.221 -6809.221
2        NA        NA        NA -7053.013 -6879.374        NA        NA
3        NA        NA        NA -7190.227 -6798.786        NA        NA
4        NA        NA        NA        NA        NA        NA        NA
5        NA        NA        NA        NA        NA        NA        NA

Top 3 models based on the BIC criterion:
    VEV,3     EEE,1     EEV,1
-6798.786 -6809.221 -6809.221
```

Figure 1. Estimated BIC value for 2019 data

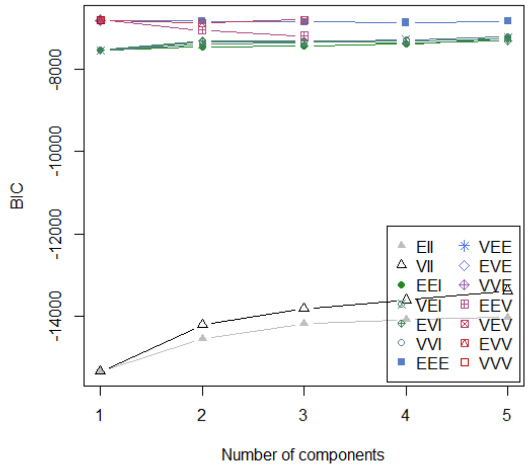


Figure 2. Plot of estimated BIC value for 2019 data

The best cluster size and distribution model shape for each year is the VEV (ellipsoidal, equal shape) model with 3 clusters for 2019 as shown in Figure 1, the VEV model shows the highest value, which is -6798,786. The visualization of the estimated BIC values is shown in Figure 2, where the VEV model as the best distribution model is symbolized by a red box containing an X symbol.

The best cluster size and distribution model shape for 2020 is the EEE (ellipsoidal, equal volume, shape, and orientation) model with 3 clusters and the VEV (ellipsoidal, equal shape) model with 4 clusters for 2021.

The cluster results for each year are shown in Table 1. According to the table, the placement of cluster members in each province varies from year to year. In 2019, clusters 1, 2, and 3 consisted of 16, 12, and 6 cluster members, respectively. In 2020, clusters 1, 2, and 3 consisted of 24, 7, and 3 cluster members, respectively. While in 2021, clusters 1, 2, 3, and 4 consist of 9, 16, 3, and 6 cluster members, respectively.

Table 1. The clustering results of the provinces for 2019, 2020, and 2021

Year	Cluster	Number of Cluster Member	Cluster Items	Average JKN users	Minimum JKN users	Maximum JKN users
2019	1	16	Aceh, West Sumatra, Jakarta, West Java, Central Java, Yogyakarta, East Java, Bali, West Nusa Tenggara, East Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, Gorontalo, West Sulawesi	9.628.447	642.132	36.557.171
	2	12	North Sulawesi, Riau, Jambi, South Sumatra, Bengkulu, Lampung, Bangka Belitung Islands, Riau Islands, Banten, West Kalimantan, South Kalimantan, Southeast Sulawesi	4.638.211	1.119.794	10.537.442
	3	6	East Nusa Tenggara, Central Kalimantan, Maluku, North Maluku, West Papua, Papua	2.405.891	931.912	4.475.557
2020	1	24	Aceh, North Sumatra, West Sumatra, Riau, South Sumatra, Bengkulu, Lampung, Bangka Belitung Islands, Riau Islands, Jakarta, Yogyakarta, Banten, Bali, West Nusa Tenggara, West Kalimantan, South Kalimantan, East Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi	4.410.847	669.195	11.427.939
	2	7	Jambi, East Nusa Tenggara, Central Kalimantan, Maluku, North Maluku, West Papua, Papua	2.260.205	874.683	4.698.976
	3	3	West Java, Central Java, East Java	33.408.390	30.312.468	39.338.216
2021	1	9	Aceh, West Nusa Tenggara, West Kalimantan, North Kalimantan, Central Sulawesi, South Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi	3.529.433	701.353	8.806.903
	2	16	North Sumatra, West Sumatra, Riau, Jambi, South Sumatra, Bengkulu, Lampung, Bangka Belitung Islands, Riau Islands, Jakarta, Yogyakarta, Banten, Bali, South Kalimantan, East Kalimantan, North Sulawesi	5.244.360	1.266.064	12.057.253

Year	Cluster	Number of Cluster Member	Cluster Items	Average JKN users	Minimum JKN users	Maximum JKN users
	3	3	West Java, Central Java, East Java	34.912.017	31.278.359	40.895.912
	4	6	East Nusa Tenggara, Central Kalimantan, Maluku, North Maluku, West Papua, Papua	2.538.171	1.006.592	4.855.719

Interpretation of the clusters in each model can be carried out through analysis of the parameters for the probability distribution of each cluster. There are 3 common parameters for the j-th cluster, namely the parameters π_j, μ_j and Σ_j. Table 2 presents the $\pi = (\pi_1, \pi_2, \ldots, \pi_m)^T$ parameter for all three years in one table. Table 3, meanwhile, presents the $\mu = (\mu_1, \mu_2, \ldots, \mu_m)^T$ parameter for 2019, 2020, and 2021, respectively. For the $\Sigma = (\Sigma_1, \Sigma_2, \ldots, \Sigma_m)$ parameter is not presented due to its large size for each year, that is, 3 matrices of 14×14 (14 is the number of variables used) in 2019 and 2020 and 4 matrices of 14×14 in 2021.

Table 2. Parameter π

	Probability in 2019	Probability in 2020	Probability in 2021
Cluster 1	0,471	0,712	0,265
Cluster 2	0,353	0,200	0,471
Cluster 3	0,176	0,088	0,088
Cluster 4	NA	NA	0,176

Table 3. Parameter μ for the years 2019, 2020, and 2021

	2019			2020			2021			
	1st Cluster	2nd Cluster	3rd Cluster	1st Cluster	2nd Cluster	3rd Cluster	1st Cluster	2nd Cluster	3rd Cluster	4th Cluster
RITP.Admisi	156,812	112,833	60,500	99,137	54,737	95,333	130,222	48,375	64,667	40,000
RITL.Admisi	639,500	517,750	349,833	445,045	285,662	439,000	402,222	405,250	383,000	239,000
RJTP.Akses Pelayanan	3833,750	3609,833	1546,500	3218,283	1576,417	3402,667	2843,000	3078,438	3143,667	1315,167
RITP.Akses_Pelayanan	141,750	102,917	54,167	91,005	49,567	88,333	118,556	44,562	61,000	36,667
RJTL.Akses_Pelayanan	978,125	791,667	437,000	659,425	355,339	629,333	503,444	693,125	567,000	296,167
RITL.Akses_Pelayanan	516,125	420,000	291,333	357,876	237,288	350,333	327,333	323,500	309,333	198,000
RITP.Hari_Rawat_Pelayanan	274,209	182,500	183,000	339,364	289,358	107,667	417,889	223,500	88,333	238,500
RITL.Hari_Rawat	2409,687	1832,833	1221,667	1743,238	1433,668	416,667	1422,000	1080,625	299,333	996,000
RJTP.Kunjungan_Pelayanan	15096,000	14186,583	4783,500	16411,992	6667,183	17551,667	13328,333	17008,750	16807,333	5446,000
RJTL.Kunjungan	4430,500	3180,833	1303,500	2995,484	1098,896	3523,000	2120,333	3696,188	3440,000	1000,333
BPJS.Dokter_Perorangan	231,187	92,833	52,667	89,703	49,793	729,000	73,444	105,438	739,667	49,333
BPJS.Klinik_Pratama	267,563	175,500	45,333	145,815	47,616	934,667	81,778	191,188	995,667	49,333
BPJS.FKRTL	99,750	55,667	24,667	57,469	25,996	326,333	39,444	73,188	352,333	27,000
Peserta_JKN	9628446,563	4638210,667	2405891,000	4410846,570	2260205,146	33408390,000	3529432,889	5244359,750	34912017,333	2538170,833

Table 4 presents the probability of each province being a member of, or coming from, a particular cluster based on the results of the probability distribution estimated by EM modeling. These probabilities are often called posterior probabilities, and in each different year, they are denoted by τ_m.

Table 4. Estimation results of the probability of the province sample unit in 2019, 2020, and 2021

Province	2019			2020			2021			
	τ_1	τ_2	τ_3	τ_1	τ_2	τ_3	τ_1	τ_2	τ_3	τ_4
Aceh	1	0	0	1	0	0	1	0	0	0
North Sumatera	0	1	0	1	0	0	0	1	0	0
West Sumatera	1	0	0	1	0	0	0	1	0	0
Riau	0	1	0	1	0	0	0	1	0	0
Jambi	0	1	0	0.211	0.789	0	0	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Papua	0	0	1	0	1	0	0	0	0	1

Lastly, Table 5 shows the characteristics of each cluster based on the average values of the 14 variables in each cluster for each year 2019, 2020, and 2021.

Table 5. Average value of each variable in each cluster for 2019, 2020, and 2021 data

	2019			2020			2021			
	1st Cluster	2nd Cluster	3rd Cluster	1st Cluster	2nd Cluster	3rd Cluster	1st Cluster	2nd Cluster	3rd Cluster	4th Cluster
RITP.Admisi	156,812	112,833	60,500	99,250	55,714	95,333	130,222	48,375	64,667	40,000
RITL.Admisi	639,500	517,750	349,833	446,125	286,857	439,000	402,222	405,250	383,000	239,000
RJTP.Akses Pelayanan	3833,750	3609,833	1546,500	3225,458	1602,286	3402,667	2843,000	3078,438	3143,667	1315,167
RITP.Akses_Pelayanan	141,750	102,917	54,167	91,125	50,429	88,333	118,556	44,562	61,000	36,667
RJTL.Akses_Pelayanan	978,125	791,667	437,000	661,167	358,714	629,333	503,444	693,125	567,000	296,167
RITL.Akses_Pelayanan	516,125	420,000	291,333	358,625	238,429	350,333	327,333	323,500	309,333	198,000
RITP.Hari_Rawat_Pelayanan	274,209	182,500	183,000	338,708	293,143	107,667	417,889	223,500	88,333	238,500
RITL.Hari_Rawat	2409,688	1832,833	1221,667	1745,667	1434,857	416,667	1422,000	1080,625	299,333	996,000
RJTP.Kunjungan_Pelayanan	15096,000	14186,583	4783,500	16437,833	6878,143	17551,667	13328,333	17008,750	16807,333	5446,000
RJTL.Kunjungan	4430,500	3180,833	1303,500	3008,625	1112,143	3523,000	2120,333	3696,188	3440,000	1000,333
BPJS.Dokter_Perorangan	231,188	92,833	52,667	89,875	50,429	729,000	73,444	105,438	739,667	49,333
BPJS.Klinik_Pratama	267,562	175,500	45,333	146,500	48,286	934,667	81,778	191,188	995,667	49,333
BPJS.FKRTL	99,750	55,667	24,667	57,708	26,143	326,333	39,444	73,188	352,333	27,000

	2019			2020			2021			
	1st Cluster	2nd Cluster	3rd Cluster	1st Cluster	2nd Cluster	3rd Cluster	1st Cluster	2nd Cluster	3rd Cluster	4th Cluster
Peserta_JKN	962844	463821	240589	442622	227359	334083	352943	524435	349120	253817
	6,562	0,667	1,000	2,375	9,286	90,000	2,889	9,750	17,333	0,833

Table 1 includes the members of the three clusters identified by the 2019 clustering results. The first cluster consists of 16 provinces, with the average of each variable having the highest value compared to the average of the other two clusters, implying that the first cluster comprises provincial members with the highest level of outcome in the BPJS program. The second cluster has 12 members with the second-highest average value of each variable, indicating that it is a cluster of provincial members with the second-best highest program outcome level. In contrast, the last cluster, namely cluster 3, consists of 6 members, namely East Nusa Tenggara, Central Kalimantan, Maluku, North Maluku, West Papua, and Papua is a group of provincial members who have the lowest BPJS program outcome level, as shown by the lowest values for all variables. Furthermore, this cluster consists mainly of Indonesian provinces from the Eastern Indonesia Region, except Central Kalimantan.

Further analysis of the characteristics for the third cluster, based on the Biplot analysis, can be seen in Figure 3, where there are no distinctive variables for the six provinces.

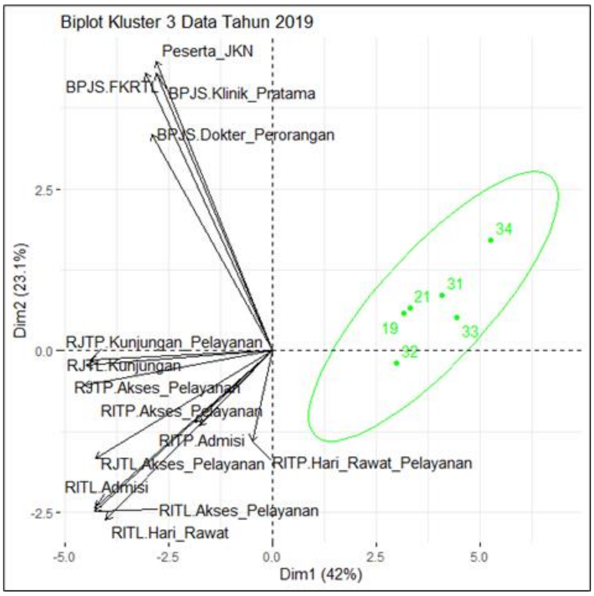
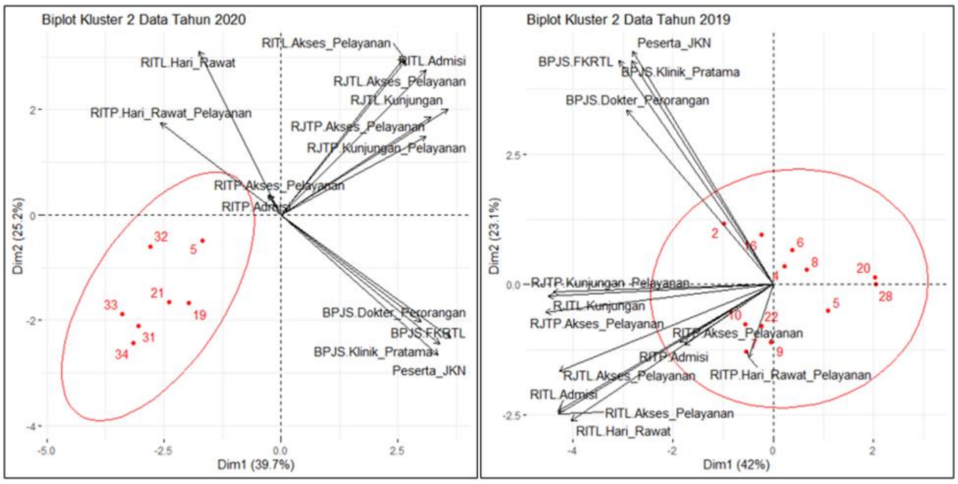


Figure 3. Biplot for cluster 3 in 2019

The clustering results for 2020 show that 3 clusters were formed, with members of clusters 1, 2, and 3 being 24, 7, and 3, respectively. In 2020, the clusters formed will be more diverse in terms of the outcomes for each variable in each cluster. According to Table 5, the average number of JKN users in 2020 reached 33,408. 390 people were in cluster 3, which included West Java, Central Java, and East Java, while cluster 2, which consists of seven provinces, including Jambi, East Nusa Tenggara, Central Kalimantan, Maluku, North Maluku, West Papua, and Papua, had the lowest average number of JKN users. Furthermore, members of cluster 2 in 2020 were also members of cluster 3 in 2019 with the addition of Jambi Province. In 2019, Jambi Province was part of cluster 2, which had the second highest average for each JKN program outcome variable. Thus, it can be said that in 2020, Jambi Province shifted to a cluster with the lowest JKN program outcomes.

Further analysis of the comparison of Jambi's position in 2019 and 2020, which can be seen from the biplot graph in Figure 4, shows that the 5th observation of Jambi in 2020 and 2019 tends not to be characterized by any variables.



(a) (b)

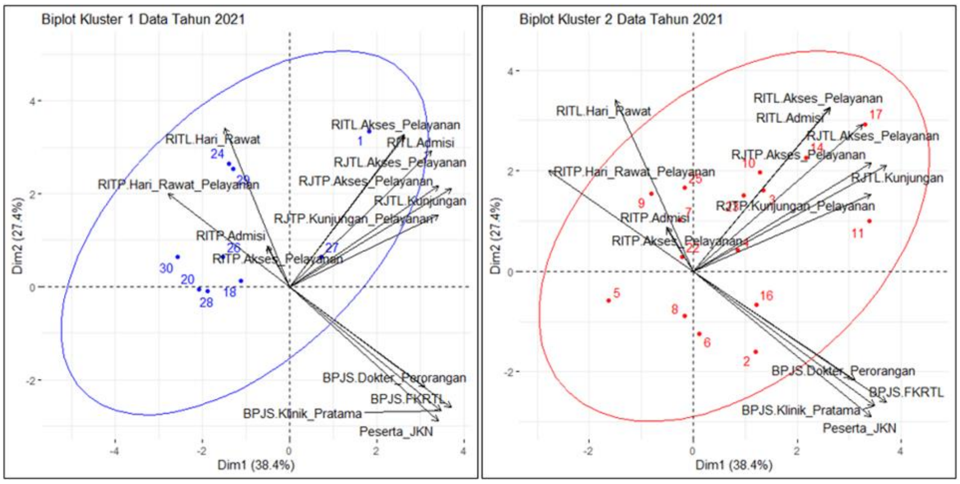
Figure 4. Biplots for 2nd cluster in 2020 (a) and 2019 (b)

Another interesting finding from the clustering results in 2020 is that cluster 3, consisting of Central Java, East Java, and West Java, is the cluster with the highest average number of JKN participants, reaching 33,408,390 individuals. These three provinces formed their own cluster after previously being part of cluster 1 in 2019, which included 16 provinces and had an average number of JKN participants of 9,628,447 individuals.

In Table 1, the number of clusters formed in 2021 is divided into four, with cluster 1 consisting of nine members: one province from Sumatra, Aceh, and five provinces from the island of Sulawesi, Central Sulawesi, South Sulawesi, Southeast Sulawesi, Gorontalo, and West Sulawesi. The remaining three provinces are from West Nusa Tenggara, West Kalimantan, and North Kalimantan. According to Table 5, cluster 1 in 2021 had an average of 3,529,432 third place JKN users. Furthermore, as shown in Biplot graph Figure 5a, the outcomes of RITL days of care and RITP days of care are more characterized by the provinces in cluster 1, in addition to observations 1 and 27, i.e., Aceh and South Sulawesi.

Cluster 2 in 2021 consists of 16 provinces: North Sumatra, West Sumatra, Riau, Jambi, South Sumatra, Bengkulu, Lampung, Bangka Belitung Islands, Riau Islands, Jakarta, Yogyakarta, Banten, Bali, South Kalimantan, East Kalimantan, and North Sulawesi. Based on these findings, it is obvious that cluster 2 is dominated by provinces on the island of Sumatra—up to nine provinces, three provinces on Sumatra, and two provinces on Kalimantan. According to Table 1, cluster 2 has the second-greatest average number of JKN users, 5,244,359 people.

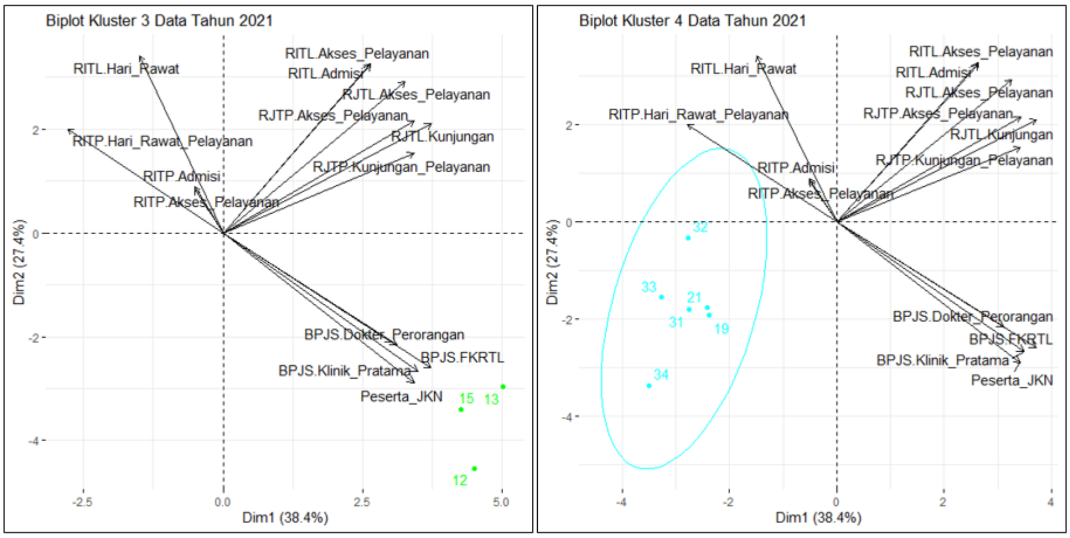
Furthermore, the biplot graph in Figure 5b shows that the fifth and eighth observations, i.e., Jambi and Bengkulu, are not characterized by any variables. While the other variables can be described by variables that point in the same direction as the locations of the observations.



(a) (b)

Figure 5. Biplot for 1st cluster in 2021 (a) and 2nd cluster in 2021 (b)

Based on Table 5, cluster 3 in 2021 has the highest average number of JKN participants, reaching 34,912,017 individuals, with 3 members: West Java, Central Java, and East Java. Further analysis using the Biplot graph for cluster 3 (Figure 6a) reveals that these three provinces are characterized by variables related to participation and health facilities collaborating with BPJS. On the other hand, cluster 4 in 2021 is characterized by the smallest average number of JKN participants, totaling 2,538,170 individuals. The members of cluster 4 in 2021 are the same as those in cluster 3 in 2019, namely NTT, Central Kalimantan, Maluku, North Maluku, West Papua, and Papua. Looking at the Biplot graph (Figure 6b) formed, it is evident that these six provinces are not characterized by any variables.



(a) (b)

Figure 6. Biplot for cluster 3 in 2021 (a) and cluster 4 in 2021 (b)

4. DISCUSSIONS

In the era of increasing awareness about the importance of public health, the role of the Indeks Pembangunan Kesehatan Masyarakat (IPKM) becomes increasingly significant. The Community Health Development Index (IPKM) is an indicator used to measure the level of development in the field of health in a region. IPKM aims to provide an overview of how far a region has succeeded in improving the health of its population and serves as a tool for the government and related agencies to evaluate health policies and programs that have been implemented, including the Jaminan Kesehatan Nasional (JKN). By regularly monitoring IPKM, the government can identify areas where public health development still needs improvement and determine more effective policy directions to enhance the overall well-being of the population.

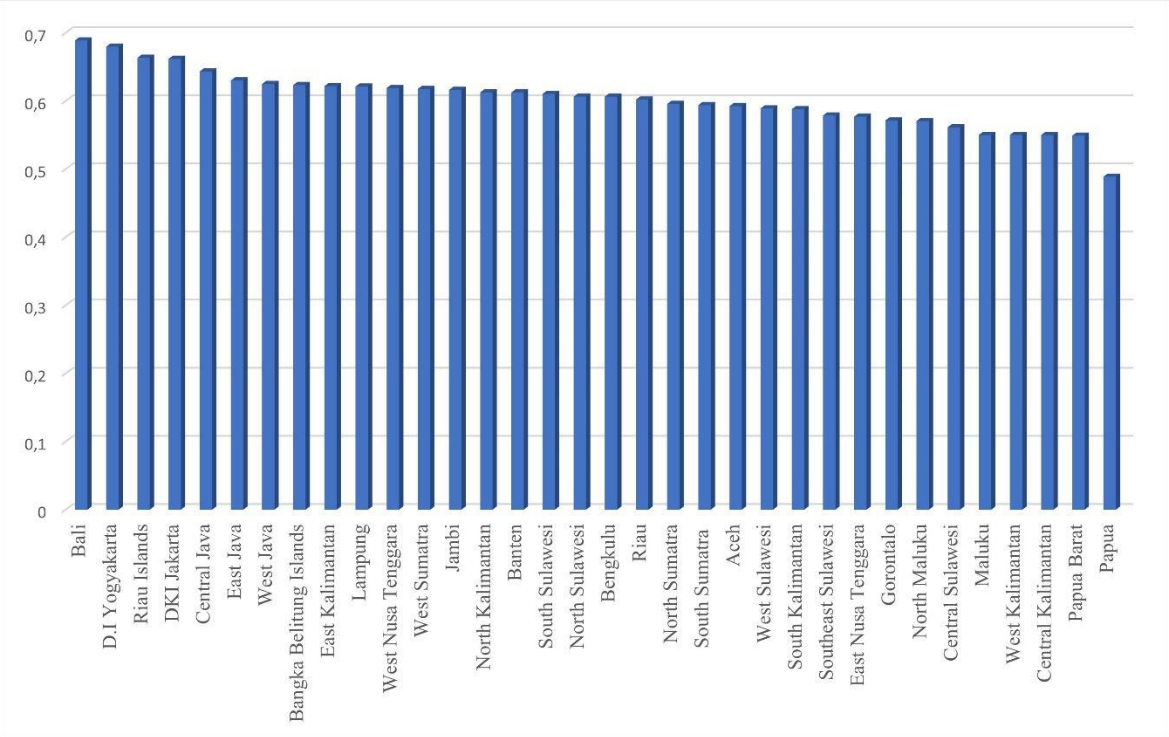
As of the publication of this article, the accessible IPKM data still uses the IPKM from 2018. In this discussion, we will explore the relationship between provincial IPKM, and the clustering results of provinces based on the outcomes of the National Health Insurance Program (JKN) in Indonesia from 2019 to 2021. The clustering of JKN outcomes in 2019 is most closely related to the IPKM data from 2018. Figure 7 presents information on provincial IPKM scores in Indonesia in 2018, which are sorted from the highest score in Bali province (0.6889) to the lowest score in Papua province (0.4888), with an average IPKM score across Indonesia of 0.6020. According to Table 6, when associated with provincial IPKM in 2018, it is observed that IPKM > 0.66 was not found in cluster 3. Cluster 3 is predominantly composed of members with IPKM < 0.56, totaling 4 provinces. Clusters with low JKN program outcomes in 2019 also had low IPKM scores compared to other clusters in 2018.

Table 6. Provincial clustering based on the outcome of the JKN program in 2019 and its relationship with provincial IPKM scores in 2018

Column Header	IPKM < 0,5600	0,5600 < IPKM < 0,6500	IPKM > 0,6600
Cluster 1	-	Aceh, West Sumatra, North Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, Gorontalo, West Sulawesi	Jakarta, West Java, Central Java, Yogyakarta, East Java, Bali, West Nusa Tenggara (NTB), East Kalimantan
Number of cluster members	0	8	8
Cluster 2	West Kalimantan	North Sulawesi, Riau, Jambi, South Sumatra, Bengkulu, Banten, South Kalimantan, Southeast Sulawesi	Lampung, Bangka Belitung Islands, Riau Islands
Number of cluster members	1	8	3
	Central Kalimantan, Maluku, West Papua, Papua	East Nusa Tenggara (NTT), North Maluku	-
Number of cluster members	4	2	0

Based on Figure 7, it can be seen that in 2018, the four provinces with the highest IPKM scores—Bali, Yogyakarta, Riau Islands, and Jakarta—are grouped in cluster 1. Table 5 shows that these four provinces were also members of cluster 1 in 2019, which had the highest JKN indicator outcomes across all variables. Therefore, it can be inferred that high JKN program outcomes are also associated with high IPKM scores. For a more specific relationship, further analysis between IPKM scores and JKN program outcomes is needed.

Based on the biplot results, an interesting characteristic to examine is cluster 3 in 2021, which had the highest average number of JKN participants, reaching 34,912,017 individuals. Members of cluster 3, namely West Java, Central Java, and East Java, are characterized by their enrollment rates and health facilities that collaborate with BPJS. When looking at their IPKM scores as shown in Figure 7, West Java, Central Java, and East Java are also provinces with high IPKM scores, above 0.6 in 2018.



Source: Visualization of data from the 2018 IPKM book, Ministry of Health, Republic of Indonesia.

Figure 7. IPKM scores of provinces in Indonesia in 2018

5. CONCLUSION

The research results show that the clustering outcomes vary each year, reflecting changes in the performance and impact of the JKN program across different provinces. In 2019, clusters were formed based on the initial implementation phase of the program, revealing distinct regional patterns in health service utilization and access. In 2020, the clusters exhibited more diversity, indicating shifts in program outcomes and the impact of regional policies on health service access and efficiency. In 2021, the clustering continued to evolve, with some provinces showing significant improvements in JKN outcomes while others lagged, highlighting areas requiring further policy intervention.

This research confirms that GMM with the EM algorithm effectively identifies clusters with similar characteristics, providing a deep understanding of the regional impact of the JKN program. Additionally, using biplot, it was found that the variables characterizing clusters with the highest JKN outcomes are participation variables and healthcare facilities collaborating with BPJS. In relation to IPKM 2018, low JKN program outcomes are also found in clusters of provinces with low IPKM scores, and vice versa. These findings suggest that continuous monitoring and targeted policies are crucial for optimizing health service delivery and achieving equitable health outcomes across Indonesia.

6. ACKNOWLEDGMENTS

The authors extend their gratitude to the Faculty of Mathematics and Natural Science, Universitas Negeri Jakarta which has funded this research.

7. REFERENCES

[1] T. R. Mayasari, “Pengelompokkan Provinsi Berdasarkan Variabel Kesehatan Lingkungan dan Pengaruhnya terhadap Kemiskinan di Indonesia Tahun 2018,” *J. Siger Mat.*, vol. 1, no. 1, 2020.

[2] H. Maryani, L. Kristiana, A. Paramita, P. Andarwati, and N. Izza, “Pengelompokan Provinsi berdasarkan Penyakit Menular dan Penyakit Tidak Menular untuk Upaya Pengendalian Penyakit dengan Pendekatan Multidimensional Scaling (MDS),” *Bul. Penelit. Sist. Kesehat.*, vol. 23, no. 3, pp. 213–225, 2021.

[3] E. M. P. Hermanto, H. B. Rochmanto, and R. Agustin, “Pemetaan Program Indonesia Sehat dengan Pendekatan Keluarga (PIS PK) di Kabupaten Bondowoso dengan K-Medoids”, STATKOM, vol. 2, no. 2, pp. 83–92, Dec. 2023.

[4] A. K. Islami and E. Widodo, “Pengelompokkan Kepemilikan Jaminan Kesehatan Menggunakan Fuzzy C-Means Algorithm (Studi Kasus: Kabupaten/Kota di Provinsi Jawa Tengah Tahun 2015),” in *SI MaNIs (Seminar Nasional Integrasi Matematika dan Nilai Islami)*, 2017, vol. 1, no. 1, pp. 299–305.

[5] C. Maklin, “Gaussian Mixture Models Clustering Algorithm Explained,” *Medium*, 2019. https://towardsdatascience.com/gaussian-mixture-models-d13a5e915c8e.

[6] Z. Wahidah and D. Utari, “Comparison Of K-Means and Gaussian Mixture Model in Profiling Areas By Poverty Indicators”, *BAREKENG: J. Math. & App.*, vol. 17, no. 2, pp. 0717-0726, Jun. 2023.

[7] A. S. Muthahharah, M. Tiro, and A. Aswi, “Application of Soft-Clustering Analysis Using Expectation Maximization Algorithms on Gaussian Mixture Model”, *Jurnal Varian*, vol. 6, no. 1, pp. 71 - 80, Nov. 2022.

[8] C. Stauffer and W. E. Grimson, “Adaptive Background Mixture Models for Real-time Tracking,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1999, vol. 2, pp. 246–252, doi: 10.1109/CVPR.1999.784637.

[9] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*, 4th ed. New York: Springer, 2006.

[10] L. J. Bain and M. Engelhardt, *Introduction to Probability and Mathematical Statistics*, 2nd ed. California: The Thomson Learning, 1992.

[11] S. K. Ng, L. Xiang, and K. K. W. Yau, *Mixture Modelling for Medical and Health Sciences*, 1st ed. New York: CRC Press, 2019.

[12] S. K. Ng and G. J. McLachlan, “An EM-based Semi-Parametric Mixture Model Approach to the Regression Analysis of Competing-risks Data.,” *Stat. Med.*, vol. 22, no. 7, pp. 1097–1111, 2003, doi: https://doi.org/10.1002/sim.1371.

[13] L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery, “Mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models,” *R J.*, vol. 8, no. 1, pp. 289–317, 2016.