BAREKENG: Journal of Mathematics and Its Applications
March 2026 Volume 20 Issue 1 Page 0439–0452
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol20iss1pp0439-0452

CLUSTER ANALYSIS OF MULTIVARIATE PANEL DATA ON DATA CONTAINING OUTLIERS

Kristuisno Martsuyanto Kapiluka[1*], **Hari Wijayanto**[2], **Anwar Fitrianto**[3]

[1,2,3]*Statistics and Data Science Study Program, School of Data Science, Mathematics, and Informatics, IPB University*
Jln. Meranti, Wing 22 Level 4 IPB University, Dramaga, Bogor, Jawa Barat, 16680, Indonesia

*Corresponding author's e-mail: * **Kris009kapiluka@apps.ipb.ac.id***

Article Info

Article History:

Received: 7th April 2025
Revised: 7th May 2025
Accepted:24th July 2025
Available online: 24th November 2025

Keywords:

Calinski-Harabasz;
Clustering;
Outlier;
Panel data;
Trajectory.

ABSTRACT

One clustering method for panel data is K-Means Longitudinal (KML), which considers only a single trajectory per subject over time. To address this limitation, KML was extended into K-Means Longitudinal 3D (KML3D), which enables clustering of joint or multivariate longitudinal data by considering multiple trajectories measured simultaneously for each subject. Both KML and KML3D provide new insights into clustering panel data using a non-hierarchical K-means approach. Hereinafter, this method is referred to as KML3D K-Means. KML3D K-Means implements the K-Means algorithm, specifically designed to cluster trajectories in panel data, and uses the mean as the clustering centroid. In practice, the K-Means algorithm is less effective in clustering data with outliers. This issue can be overcome by KML3D K-Medoids, a method based on KML3D that uses the median as the centroid. This study aims to determine cluster analysis for multivariate panel data on data containing outliers with KML3D K-Means and KML3D K-Medoids. Both methods are applied to panel data of social and welfare statistical data from 34 provinces observed for 8 years (2016 – 2023). The comparison of methods is based on the Calinski–Harabasz index. The results of the study show that KML3D K-Medoids has a Calinski-Harabasz index that is higher than KML3D K-Means in clustering multivariate panel data with outliers. The analysis identified three optimal clusters (k = 3) based on the Calinski–Harabasz (CH) index, which can be categorized as the "more prosperous", "moderately prosperous", and "less prosperous" groups. The growth rate analysis reveals disparities in development trajectories across clusters, with cluster 3 showing the most consistent improvements, cluster 1 moderate progress, and cluster 2 lagging in key social and welfare indicators.

How to cite this article:

K. M. Kapiluka, H. Wijayanto, and A. Fitrianto, "CLUSTER ANALYSIS OF MULTIVARIATE PANEL DATA ON DATA CONTAINING OUTLIERS", *BAREKENG: J. Math. & App.*, vol. 20, no. 1, pp. 0439-0452, Mar, 2026.

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

Cluster analysis is a statistical method that specifically groups data into several clusters in such a way that the level of similarity among objects within the same cluster is maximized, while the similarity among clusters is minimized [1]. Based on its working mechanism, cluster analysis is divided into two main groups, namely hierarchical clustering and non-hierarchical clustering. Hierarchical clustering works by grouping objects in a structured manner based on their similarity, whereas non-hierarchical clustering assigns objects into predefined clusters [2]. One of the most popular and widely used methods in non-hierarchical clustering for research related to cluster analysis is the K-Means method. Because, in terms of its algorithm, it can be implemented simply and has powerful functionality [3].

In general, the K-Means method can only be applied to cross-sectional data, which refers to data observed or measured at a specific point in time. Based on the observation period, there are three types of data: cross-sectional data, time series data, and panel data. Panel data is a data structure consisting of three dimensions: objects, indices, and time, which can be viewed as an extension of cross-sectional data in the time dimension [4]. Panel data, also known as longitudinal data, has a unique characteristic of repeated measurements of the same objects [5]. The study by [6] shows that the K-Means clustering method can also be applied to panel data, known as the K-Means longitudinal (KML) method.

KML is a panel data clustering method that specifically uses the K-Means algorithm to work on panel data trajectories. This method uses a naïve-based approach to objects, where the relationship between measurements over time is not modeled [7]. The KML approach clusters univariate longitudinal data, considering only one trajectory per subject over time. Consequently, [8] developed KML3D to cluster joint or multivariate longitudinal data, where multiple trajectories (variables) are measured simultaneously per subject. This method is essentially the same as KML in its core clustering mechanism, differing only in the preprocessing stage to handle multiple trajectories and in the output interpretation. The development of KML and KML3D offers new perspectives on clustering panel or longitudinal data through a non-hierarchical clustering approach, specifically utilizing K-means.

Several studies that have used the KML and KML3D methods in analytical research include: the study conducted by [9] to cluster opioids based on their clinic types. Similarly, [10] used the KML method in their research to cluster longitudinal antibody data. Furthermore, [11] applied the KML3D method in their research to characterize plasma metabolic responses for longitudinal data. Another study that used this method is [12], which is clustering Indonesian's provinces based on the Human Development Index (HDI), from 2010-2019.

In cluster analysis, the presence of outliers in the data negatively affects the performance of the clustering algorithm used [13]. Both the KML and KML3D methods adopt the K-Means algorithm, whose performance can be affected by the presence of outliers. According to [14], the K-Means algorithm has limitations when dealing with data that contains outliers. The study conducted by [15] comparing the performance of K-Means and K-Medoids, and Hierarchical Clustering shows that K-Medoids are more robust to outliers in the dataset. Likewise, [16] in their research compared the performance of K-Means and K-Medoids on data containing outliers. The results showed that the K-Medoids algorithm performed better than K-Means. Based on these studies, K-Medoids is considered more robust to the presence of outliers. Therefore, the K-Medoids is proposed within the KML3D framework as KML3D K-Medoids to address outliers in panel data clustering in this study.

Based on the explanation above, this study aims to compare the KML3D K-Means and KML3D K-Medoids methods for clustering multivariate panel data with the presence of outliers. The clustering is performed on social and welfare statistical panel data from 34 provinces in Indonesia, measured over eight years (2016-2023). The comparison between the two methods and the evaluation of the optimal number of clusters are conducted using the Calinski-Harabasz index [17]. The method that yields the highest Calinski-Harabasz value is considered the most optimal in this study.

2. RESEARCH METHODS

2.1 Data

Data in this study consists of national social and welfare statistics from 34 provinces, sourced from the annual publication of the Central Bureau of Statistics (BPS) [18]. This is panel data observed over eight years (2016–2023) with ten (10) numerical variables**.** The following table presents the variables used in this study.

Table 1. Research Variables

Variables	Explanation
X_1	The ratio of elementary school facilities under the Ministry of Education and Culture per 1,000 residents
X_2	The ratio of higher education facilities under the Ministry of Education, Culture, Research, and Technology per 1,000 residents
X_3	The ratio of general hospital facilities per 1,000 residents
X_4	The percentage of toddlers who have received complete immunization
X_5	The percentage of women using family planning
X_6	The percentage of households with access to proper sanitation
X_7	The percentage of households with access to drinking water sources
X_8	The crime rate ratio per 1,000 residents
X_9	The percentage of the population living in poverty
X_{10}	Human Development Index

2.2 Exploratory and Pre-Processing

Exploration aims to show a general overview of the data through visualization using graphs, while pre-processing is carried out as an initial step in clustering analysis. Visualization is performed using boxplots to observe the distribution of outliers for each variable. Additionally, a time series plot is used to identify data patterns over time. In the pre-processing stage, standardization is also conducted. Standardization is performed to reduce bias in clustering analysis caused by significant differences in scale among the variables used [19]. The standardization process for panel data follows the approach outlined in [8].

The next pre-processing stage is to transform the multivariate panel data format into a two-dimensional table. This table represents the row elements as the observed objects, and the column elements as the time length, where each time point has a certain number of variables used. The following is the form of the pre-processing stages in panel data introduced by [20].

Table 2. Pre-Processing Data

Objects	**Time 1st**				**…**	**Time *t*th**			
	X_1	X_2	…	X_p	…	X_1	X_2	…	X_p
1	x_{11}	x_{12}	…	x_{1p}	…	x_{11}	x_{12}	…	x_{1p}
2	x_{21}	x_{22}	…	x_{2p}	…	x_{21}	x_{22}	…	x_{2p}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	x_{n1}	x_{n2}	…	x_{np}	…	x_{n1}	x_{n2}	…	x_{np}

After the pre-processing steps described in Table 2, the next step is to create a joint trajectories matrix [8]. This joint trajectories matrix consists of a collection of trajectory variables that include observation units, time, and variables. The structure of this joint trajectories matrix can be expressed using Eq. (1).

$$\boldsymbol{X}_{i..} = \begin{pmatrix} x_{i1X_1} & x_{i2X_1} & \cdots & x_{itX_1} \\ x_{i1X_2} & x_{i2X_2} & \cdots & x_{itX_2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{i1X_p} & x_{i2X_p} & \cdots & x_{itX_p} \end{pmatrix}. \tag{1}$$

2.3 Panel Data Clustering with KML3D

The clustering of panel data in this study refers to trajectory clustering, as stated by [8], and is referred to as k-means longitudinal 3D (KML3D). This method operates by designing the K-Means algorithm to specifically work on joint trajectories. In principle, the working mechanism of the algorithm in this method

is similar to the K-Means algorithm in cluster analysis of cross-sectional data. The only difference lies in the specially designed distance measure, which is the Euclidean distance. This distance measures two objects in the form of joint trajectory matrices, as shown in Eq. (1). Suppose that there are two objects in the form of trajectory matrices, namely, $\boldsymbol{X_{1..}}$ and $\boldsymbol{X_{2..}}$, then the distance between these two objects is defined as the norm of the combination of distance vectors measured using the Euclidean distance based on their time points. The distance equation is written as follows:

$$d(\boldsymbol{X_{1..}}, \boldsymbol{X_{2..}}) = \left\| \left(d_{1.}(x_{11.}, x_{21.}), d_{2.}(x_{12.}, x_{22.}), \dots, d_{t.}(x_{1t.}, x_{2t.}) \right) \right\|, \quad (2)$$

where:

$(\boldsymbol{X_{1..}}, \boldsymbol{X_{2..}})$: Object 1 and object 2.
$d_{t.}$: The Euclidean distance between values in $X_{1..}$ and $X_{2..}$ at the same time t for each variable.
$x_{1t.}$: The value of the first element in $X_{1..}$ at time t for each variable.
$x_{2t.}$: The value of the first element in $X_{2..}$ at time t for each variable.

The distance in Eq. (2) is then applied to the K-Means and K-Medoids algorithms to measure the distance between two objects in trajectory clustering or KML3D. In its implementation, KML3D K-Means uses the mean as its centroid, whereas KML3D K-Medoids uses the median as its centroid.

The procedure for KML3D trajectory clustering in this study is as follows:

1. Preparing the panel data.
2. Conducting descriptive statistical exploration.
3. Converting the panel data format into a long cluster data format.
4. Setting the range of the optimal number of clusters ($k = 2$ to $k = 10$).
5. Performing KML3D clustering using the mean as the centroid for KML3D K-Means and the median as the centroid for KML3D K-Medoids.
6. Calculating the Calinski-Harabasz (CH) index for the clustering results of KML3D K-Means and KML3D K-Medoids.
7. Selecting the optimal number of clusters k based on the highest CH index from both methods.
8. Obtaining the final clustering results using the optimal k.
9. Interpreting the clustering results.

2.3.1 K-Means Clustering Algorithm

K-Means is a distance-based clustering algorithm that works by dividing data into a number of clusters or groups [21]. K-Means is a part of non-hierarchical methods that operate by partitioning existing data into one or more groups [22]. The stages of the K-Means algorithm in this study follow the general steps as described by [23].

2.3.2 K-Medoids Clustering Algorithm

K-Medoids is a clustering method that uses the median within a cluster, with medoids serving as representative objects of the cluster [24]. According to [25], K-Medoids can overcome the weakness of K-Means, which tends to be sensitive to outliers that may deviate from the data distribution. The stages of the K-Medoids algorithm in this study follow the general steps as implemented by [26].

2.4 Cluster Validation

Cluster evaluation is conducted to determine the best method based on the comparison of the average Calinski-Harabasz (CH) index values from both methods, while the optimal number of clusters is also determined based on the Calinski-Harabasz index. Calinski-Harabasz [17] is an index that calculates the ratio between the Sum of Squares Between clusters (SSB) as separation and the Sum of Squares Within clusters (SSW) as compactness, multiplied by a normalization factor, which is the difference between the number of data points and the number of clusters divided by the number of clusters (g) minus one. A higher CH value indicates a better clustering solution [27]. The CH index is formulated by:

$$CH = \frac{trace\ SSB}{trace\ (SSW)} \times \frac{N - g}{g - 1}. \quad (3)$$

3. RESULTS AND DISCUSSION

Before performing cluster analysis, data exploration and pre-processing are conducted as initial steps in the panel data clustering analysis.

3.1 Exploratory and Pre-Processing

The results of data exploration, specifically the time series plot of each variable over time, are shown in Fig. 1. From Fig. 1, it can be observed that each variable exhibits a different time series pattern.

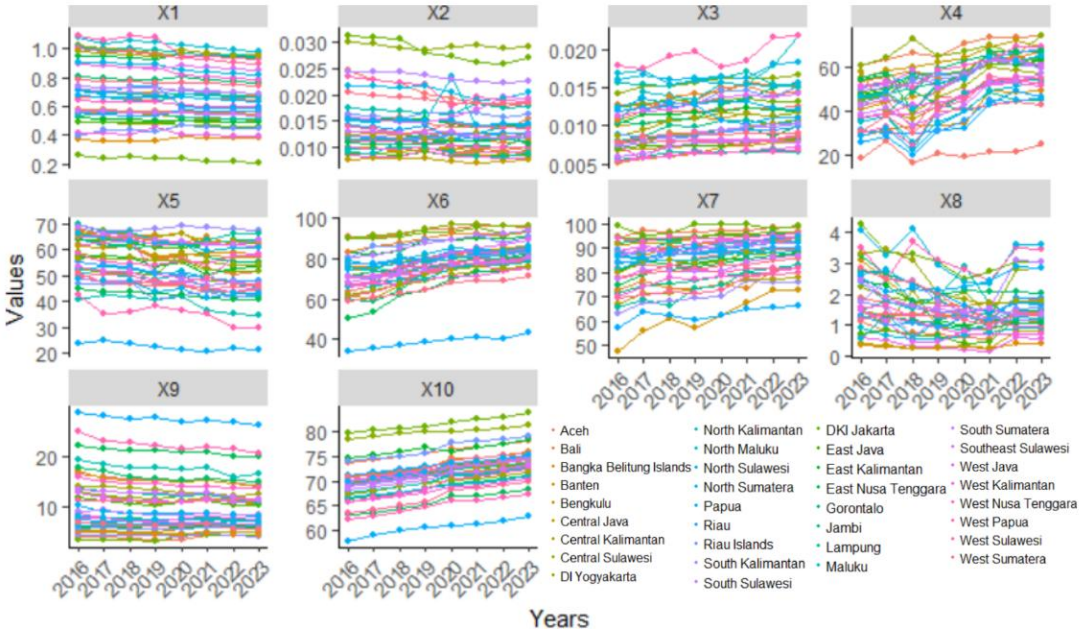

Figure 1. Plot Description of Time Series of Variables
(Source: ggplot RStudio)

Based on the trajectory plots of ten social and welfare indicators (X_1 to X_{10}) observed across 34 Indonesian provinces from 2016 to 2023, notable temporal and regional disparities are evident. Several variables, such as X_6 (access to proper sanitation) and X_{10} (Human Development Index), exhibit consistent upward trends across most provinces, indicating national progress in infrastructure and development outcomes. In contrast, X_5 (family planning participation) generally declines, and X_8 (crime rate) shows irregular patterns, reflecting region-specific socio-political dynamics. While variables like X_1 (elementary education) and X_2 (higher education) remain relatively stable, disparities persist, particularly in underdeveloped regions such as Papua, which records consistently low values in X_5, X_6, X_7, and X_{10}, and high poverty levels (X_9). Conversely, provinces like DKI Jakarta, DI Yogyakarta, and Bali demonstrate strong and stable trajectories in education, health, and human development indicators (e.g., X_2, X_4, X_6, X_7, X_8, X_{10}). Meanwhile, West Nusa Tenggara, West Java, and Banten fall behind in healthcare and safety indicators, and Aceh and Central Java show limited progress in immunization and higher education.

Furthermore, the plots in Fig. 1 reveal differences in measurement scales across the variables. Additionally, some provinces exhibit significant variation in certain variables compared to others. Therefore, standardization is necessary to address these differences, and an outlier plot is required to identify outliers in each variable. The distribution of outliers for each standardized variable is illustrated in a boxplot in Fig. 2.

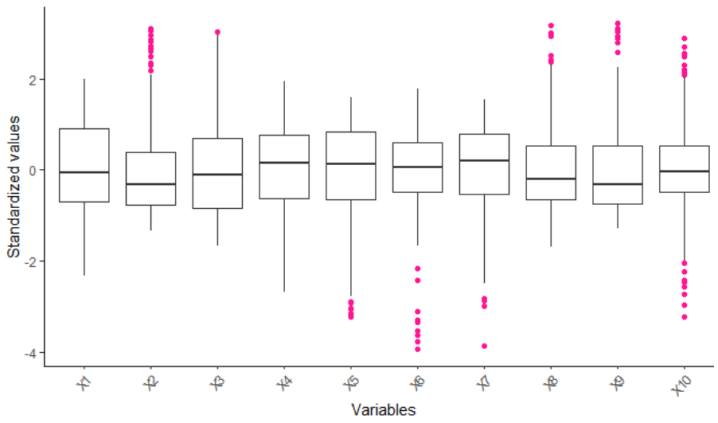

Figure 2. Plot of the Distribution of Outlier Values for Each Variable
(Source: ggplot RStudio)

Based on Fig. 2, it can be seen that almost all variables (except X_1 and X_4) have values that deviate significantly (red points) from the other observed data. This is referred to as outliers. According to [28], the presence of outliers affects the formation of coherent clusters.

3.2 Optimal Cluster

In the analysis of non-hierarchical clustering, such as K-Means and K-Medoids, the determination of the optimal number of clusters is made beforehand, prior to running the clustering algorithm. One of the methods for evaluating the optimal number of clusters is the Calinski-Harabasz index. Table 3 presents the determination of the optimal number of clusters for the KML3D K-Means and KML3D K-Medoids methods in panel data clustering.

Table 3. CH Index of Cluster Optimal

***k* cluster**	**Calinski – Harabasz Index**	
	KML3D K-Means	**KML3D K-Medoids**
2	5.69483	5.69483
3	**5.57220**	**6.93331**
4	5.97505	6.48833
5	5.46124	5.43313
6	5.75985	5.87772
7	4.99992	5.78954
8	4.82053	4.27558
9	5.19941	4.93439
10	5.24943	4.93162

These results suggest that KML3D K-Medoids yields a more optimal number of clusters compared to KML3D K-Means, based on the Calinski-Harabasz Index.

3.3 Cluster Results

Based on the selection of the optimal number of clusters in Table 3, it was found that the KML3D K-Medoids method yielded the highest Calinski-Harabasz (CH) index value at $k = 3$. Therefore, the KML3D K-Medoids method is selected to cluster the social and welfare statistics panel data from 34 provinces measured over eight years (2016–2023). The clustering results using KML3D K-Medoids are presented in Table 4.

Table 4. Cluster Results

Cluster	Number of cluster members	Cluster member
1	13	Riau, Riau Islands, DKI Jakarta, West Java, Central Java, DI Yogyakarta, East Java, Banten, Bali, West Nusa Tenggara, East Kalimantan, North Kalimantan, Southeast Sulawesi.
2	12	Aceh, North Sumatera, West Sumatera, East Nusa Tenggara, Central Sulawesi, South Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua, Papua
3	9	Jambi, South Sumatera, Bengkulu, Lampung, Bangka Belitung Islands, West Kalimantan, Central Kalimantan, South Kalimantan, North Sulawesi,

Based on the clustering results presented in Table 4, provinces grouped within the same cluster share similar characteristics according to the panel data on social and welfare statistics used in this study. Cluster 1 comprises the largest number of provinces, totaling 13, followed by Cluster 2 with 12 provinces, and Cluster 3 with 9 provinces. In terms of geographical distribution, Cluster 1 is dominated by all provinces on the island of Java, along with two provinces from Sumatra (Riau and Riau Islands), Bali, West Nusa Tenggara, two provinces from Kalimantan (East Kalimantan and North Kalimantan), and Southeast Sulawesi from the Sulawesi region. Cluster 2 is primarily composed of provinces from the eastern part of Indonesia, including five provinces from Sulawesi, East Nusa Tenggara, Maluku, North Maluku, Papua, and West Papua, as well as three provinces from Sumatra (Aceh, North Sumatra, and West Sumatra). Lastly, Cluster 3 consists of six provinces from Sumatra and its surrounding areas, three from Kalimantan (West Kalimantan, Central Kalimantan, and South Kalimantan), and one from Sulawesi (North Sulawesi).

3.4 Cluster Medoids Distribution

After obtaining the clustering results as presented in Table 4, the next step is to visualize the distribution patterns in order to gain insights into the spread of values within each formed cluster for every variable. This can be illustrated through the following Fig. 3.

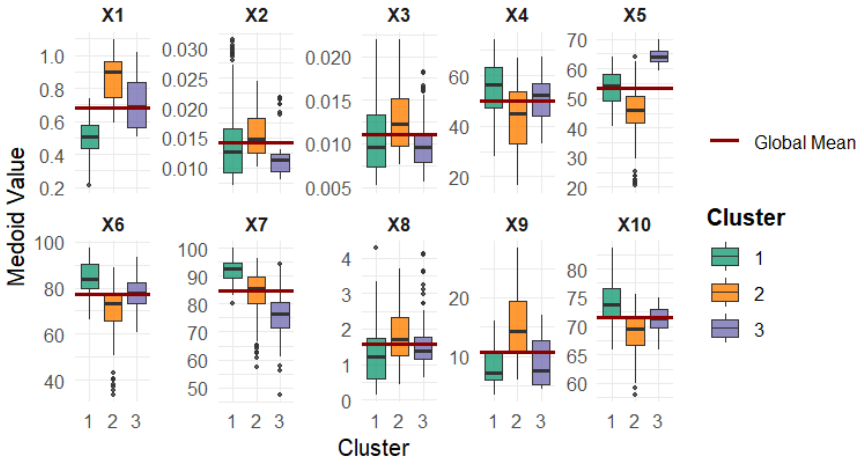

Figure 3. Cluster Modeoids Distribution
(Source: ggplot RStudio)

The distribution of medoid values for each variable across the three identified clusters is illustrated in Fig. 3. The red horizontal line within each facet represents the global mean of the corresponding variable, calculated across all clusters. For variable X_1, the distribution of Cluster 2 lies substantially above the global mean, indicating a higher overall value of this variable within this cluster. Cluster 3 centers around the global mean, while Cluster 1 consistently falls below it, suggesting lower values of X_1 in this group. For X_2, Cluster 2 again shows higher values above the global mean, while Clusters 1 and 3 fall below it, with Cluster 3 having the lowest median values. This pattern points to Cluster 2 being associated with superior outcomes in the dimension represented by X_2. In the case of X_3, all three clusters are more evenly distributed around the global mean, although Cluster 2 maintains a slight advantage with medoid values that tend to be higher. For X_4, Cluster 1 is situated above the global average, Cluster 3 is aligned closely with it, and Cluster 2 is

positioned notably below. This reflects relatively better performance of Cluster 1 in the area captured by X_4. In X_5, Cluster 3 exhibits a clear elevation above the global mean, Cluster 1 aligns with the average, and Cluster 2 is lower, suggesting Cluster 3 performs best on this indicator. A similar trend appears in X_6, where Clusters 1 and 3 are clustered around the global mean, but Cluster 2 stands out with a distinctly lower distribution, highlighting a potential area of deficiency. In X_7, Cluster 1 is well above the global mean, whereas Cluster 3 is centered around it, and Cluster 2 lies below, indicating a potential advantage of Cluster 1. For X_8, all three clusters exhibit distributions near the global mean, though Cluster 2 tends to show slightly higher values, whereas Cluster 1 maintains lower medoid values. In X_9, the lowest values are concentrated in Cluster 3, while Clusters 1 and 2 are somewhat higher but still distributed below the global average, pointing to underperformance across all clusters in this variable. Finally, X_{10} reveals that Cluster 1 consistently lies above the global mean, Cluster 3 is close to the mean, and Cluster 2 falls below, suggesting Cluster 1 is relatively more advanced in the development dimension represented by X_{10}.

3.5 Cluster Trajectories

In the context of panel data clustering, it is also important to examine the clustering results in the form of time series patterns. Cluster trajectories plots are used to illustrate the changes in cluster averages over time for each variable.

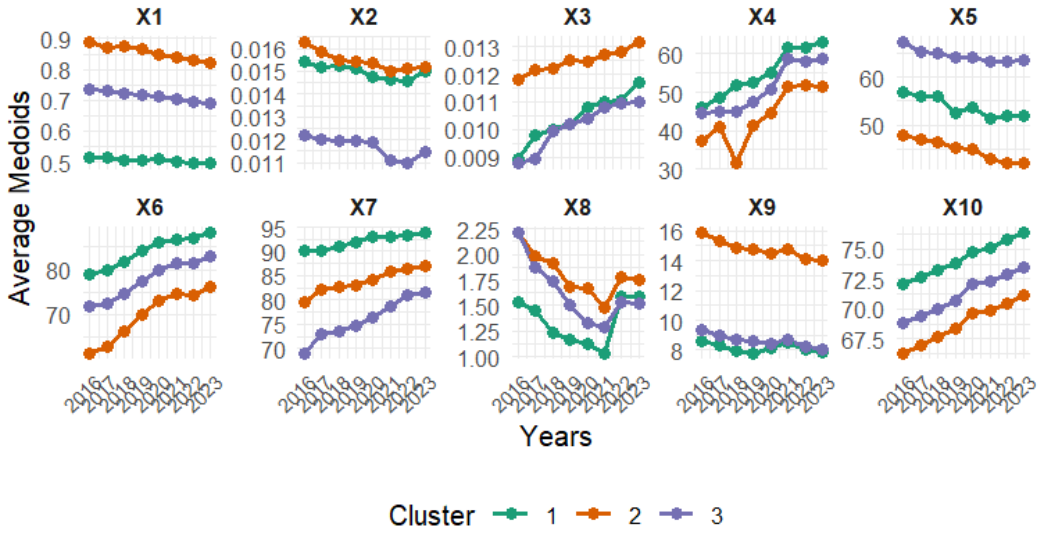

Figure 4. Plot of Cluster Trajectories
(Source: ggplot RStudio)

The temporal trajectories of average medoids for each cluster across ten observed variables (X_1 to X_{10}) over the period from 2016 to 2023 are presented in Fig. 4. These trajectories provide insights into the dynamic patterns of social and welfare indicators among the clusters formed in the study. For variable X_1, the trajectories show that Clusters 1 and 2 maintain relatively higher and stable values throughout the period, while Cluster 3 consistently lags behind, indicating that provinces in Cluster 3 tend to have less favorable conditions related to this indicator. A similar pattern appears in X_2, where Clusters 1 and 2 again dominate with higher average values, whereas Cluster 3 remains at a lower level across the years, suggesting structural disadvantages in that cluster for the corresponding aspect. The trends in X_3 are more stratified, with Cluster 2 showing the most prominent increase over time, followed by Cluster 1 and then Cluster 3, which displays slower growth. This pattern implies differentiated progress in the indicator represented by X_3, where Cluster 2 demonstrates relatively stronger performance. In contrast, X_4 exhibits an interesting dynamic: although all clusters experience an increasing trend, Cluster 3—initially at the lowest point—shows a sharp rise starting around 2016, eventually converging with the other clusters. This indicates notable improvement in the provinces within Cluster 3, possibly as a result of targeted interventions or policy reforms. A reversal in pattern is observed in X_5, where Cluster 3 consistently maintains the highest average values, and Clusters 1 and 2 remain below throughout the period. This suggests that Cluster 3 holds comparative advantages in this particular dimension. The pattern continues in X_6 and X_7, where Cluster 3 outperforms the other clusters with higher values across time, indicating that the provinces in this cluster may excel in specific social or welfare-

related aspects covered by these variables. In X_8, the trajectories reveal a different development. All clusters experience a decline followed by a recovery, with Cluster 1 showing the most pronounced dip before returning to its previous level. This fluctuation may reflect temporary shocks or disruptions affecting the provinces in Cluster 1. The pattern in X_9 is relatively stable across clusters, though Cluster 2 tends to slightly outperform the others, suggesting marginal yet consistent differences. Finally, X_{10} shows a steady increase across all clusters, but Cluster 3 maintains a consistently higher trajectory, indicating continuous improvement and sustained advantage in the dimension represented by this variable. Overall, the trajectory patterns demonstrate that the clusters are not only distinct in terms of their baseline characteristics but also in their developmental progress over time.

3.6 Cluster Characteristics

Characteristics of the cluster obtained using quartile calculations. If the value is less than or equal to Q_1, it is classified as "low". If the value is greater than Q_1 and less than or equal to Q_2, it is classified as "medium". Finally, if the value is greater than Q_2 and less than or equal to Q_3, it is classified as "high". The results of the categorization of the characteristic variables for each cluster are presented in Table 5.

Table 5. Characteristic Variable Categories

Cluster	Categories		
	Low	**Medium**	**High**
1	$X_1, X_3, \boldsymbol{X_8}, \boldsymbol{X_9}$	X_5	$X_2, X_4, X_6, X_7, X_{10}$
2	X_4, X_5, X_6, X_{10}	X_7	$X_1, X_2, X_3, \boldsymbol{X_8}, \boldsymbol{X_9}$
3	$X_2, X_3, X_7, \boldsymbol{X_9}$	$X_1, X_4, X_6, \boldsymbol{X_8}, X_{10}$	X_5

Cluster 1 is characterized by relatively strong performance in infrastructure and basic service access. Provinces in this cluster exhibit high values in variables such as the ratio of higher education facilities (X_2), the percentage of toddlers receiving complete immunization (X_4), access to proper sanitation (X_6), access to drinking water sources (X_7), and the Human Development Index (X_{10}). These indicators suggest that the regions in this cluster are well-equipped in terms of health services and educational facilities, and show notable achievements in key aspects of human development. Variable X_5—representing the percentage of women using family planning—falls into the medium category, implying a moderate level of reproductive health outreach. However, the provinces in Cluster 1 record low values in variables such as the availability of elementary schools (X_1), general hospital facilities (X_3), the crime rate (X_8), and the poverty rate (X_9). The low crime rate (X_8) may reflect a positive condition, but low values in education and hospital infrastructure suggest potential areas for investment to complement otherwise strong social indicators.

Cluster 2 demonstrates a profile of mixed development, with both high-performing and lagging indicators. High scores are observed in variables such as access to elementary and higher education facilities (X_1 and X_2), general hospital availability (X_3), lower crime rates (X_8), and lower poverty levels (X_9), reflecting a favorable combination of education, health, safety, and economic well-being. The percentage of households with access to drinking water (X_7) is categorized as medium, suggesting ongoing efforts in infrastructure provision. Nonetheless, this cluster shows low values in variables such as immunization coverage (X_4), family planning participation (X_5), access to sanitation (X_6), and the Human Development Index (X_{10}). These deficiencies highlight the need for social protection programs, improved hygiene infrastructure, and targeted interventions in maternal and child health services.

Cluster 3, on the other hand, reflects transitional or intermediate development conditions. Most of the variables—including X_1, X_4, X_6, X_8, and X_{10}—fall into the medium category, suggesting a balanced level of development without pronounced strengths or weaknesses. Provinces in this cluster exhibit high values in the percentage of women using family planning (X_5), indicating relative success in reproductive health services. However, this cluster records low levels in key infrastructure and social protection variables such as access to higher education (X_2), hospital availability (X_3), drinking water access (X_7), and a relatively high crime rate (X_9). These conditions point to challenges in both health service capacity and public safety, while also suggesting opportunities for targeted policy interventions to enhance access and resilience.

Since this is panel data involving a time dimension—specifically, years—it is necessary to examine the growth rate of each variable within each cluster. The growth rate reflects the annual percentage increase

or decrease in the value of each variable for a given cluster [12]. A high growth rate indicates that the variable is developing more rapidly, whereas a low growth rate suggests slower progress in that indicator.

Table 6. Growth Rate of Variables

Cluster	1	2	3
X_1	-0.57	-1.09	-0.95
X_2	-0.41	-1.01	-0.89
X_3	3.93	1.59	3.32
X_4	4.61	4.60	3.94
X_5	-1.22	-1.85	-0.79
X_6	1.57	3.15	2.07
X_7	0.59	1.30	2.45
X_8	0.54	-3.30	-5.21
X_9	-1.30	-1.75	-2.04
X_{10}	0.83	1.02	0.95

Based on Table 6, cluster 1 demonstrates a mixed trajectory in terms of annual growth across social and welfare indicators. Educational infrastructure indicators show declining trends, with the ratio of elementary school facilities (X_1) decreasing by −0.57% annually, and higher education facilities (X_2) declining by −0.41%. In contrast, general hospital facilities (X_3) exhibit a strong annual growth of 3.93%, signaling a positive shift in health service capacity. Immunization coverage (X_4) grows steadily at a rate of 4.61%, reflecting consistent public health initiatives. However, participation in family planning (X_5) shows a negative growth rate of −1.22%, suggesting decreasing engagement. Sanitation access (X_6) and drinking water access (X_7) grow at moderate annual rates of 1.57% and 0.59%, respectively, indicating incremental improvements in basic services. The crime rate (X_8) experiences a slight annual increase of 0.54%, while poverty (X_9) declines at a rate of −1.30%, highlighting gradual socio-economic advancement. The Human Development Index (X_{10}) increases steadily at 0.83% per year, suggesting sustained improvements in overall well-being. Cluster 2 shows more pronounced negative growth patterns, particularly in educational and reproductive health indicators. The elementary school facility ratio (X_1) has the sharpest decline among clusters at −1.09%, while higher education facilities (X_2) decrease by −1.01% per year. General hospital facilities (X_3) grow modestly at 1.59%, slower than other clusters, possibly limiting service expansion. Immunization coverage (X_4) improves consistently with an annual increase of 4.60%. Family planning participation (X_5) experiences the steepest annual decline (−1.85%), reflecting significant barriers or decreasing public engagement. Sanitation (X_6) improves at 3.15% per year, one of the higher rates across clusters, while access to drinking water (X_7) increases at a slower rate of 1.30%. The crime rate (X_8) drops considerably by −3.30% annually, suggesting a notable improvement in public safety conditions. Poverty (X_9) decreases moderately by −1.75% per year, and the Human Development Index (X_{10}) exhibits the highest annual growth rate among clusters at 1.02%, indicating proactive development interventions despite low starting conditions. Cluster 3 reflects relatively favorable dynamics in growth trends across several key dimensions. The family planning indicator (X_5) shows a mild annual decline of −0.79%, indicating a slight reduction in participation. Immunization (X_4) and sanitation (X_6) grow at 3.94% and 2.07% per year, respectively, suggesting improved healthcare access. Drinking water access (X_7) demonstrates the highest annual increase (2.45%) among clusters, highlighting infrastructure gains. Despite moderate baseline levels, elementary (X_1) and higher education facilities (X_2) are both declining annually by −0.95% and −0.89%, respectively, which may warrant further policy attention. Notably, the crime rate (X_8) exhibits the sharpest decline (−5.21%) across all clusters, indicating substantial gains in safety and security. Poverty (X_9) declines at −2.04% per year, representing the most rapid reduction among clusters. The Human Development Index (X_{10}) increases consistently at 0.95% annually, reflecting continued improvements in multidimensional aspects of development.

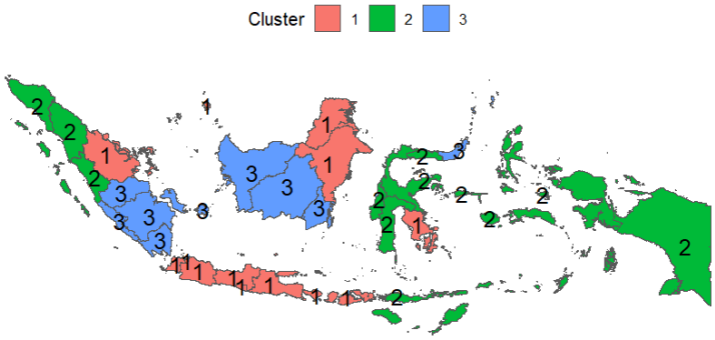

Figure 5. Cluster Map
(Source: ggplot RStudio)

Based on the cluster map in Fig. 5, Cluster 1, representing the more prosperous group, is predominantly located in Java and its surrounding areas, including Bali, Riau, and parts of Kalimantan and Sulawesi, characterized by strong access to infrastructure and higher social development. Cluster 2, classified as the less prosperous group, encompasses most provinces in the eastern region, including Papua, Maluku, East Nusa Tenggara, and several provinces in Sulawesi and Sumatra, where access to essential services and development indicators remains low. Cluster 3, the moderately prosperous group, consists primarily of central provinces in Sumatra and Kalimantan, reflecting an intermediate level of social welfare performance.

4. CONCLUSION

Based on the results of the panel data clustering analysis for data containing outliers, with a case study on the panel data of social and welfare statistics from 34 provinces in Indonesia, measured over 8 years (2016–2013), the following conclusions were obtained:

1. The K-Medoids method implemented through the KML3D algorithm (KML3D K-Medoids) demonstrated superior performance in clustering panel data containing outliers, as indicated by a higher Calinski–Harabasz index compared to the K-Means method (KML3D K-Means). This confirms its robustness and suitability for analyzing multidimensional social and welfare data over time.
2. The clustering analysis identified three distinct clusters of provinces :
   Cluster 1 includes provinces such as those in Java, Riau, Riau Islands, Bali, and parts of Kalimantan and Sulawesi, characterized by strong social and welfare conditions, including high HDI, better infrastructure, and access to services. **Cluster 2** consists mostly of eastern provinces, including Papua, Maluku, Nusa Tenggara, and several from Sulawesi and Sumatra, which exhibit the lowest performance in most welfare indicators. **Cluster 3** represents moderately performing provinces located primarily in central Sumatra and Kalimantan, serving as a transitional group between the two extremes.
3. The growth rate analysis across all clusters highlights significant disparities in the pace of social and welfare development among Indonesian provinces. While **Cluster 1** demonstrates moderate and consistent improvement across key health and infrastructure indicators, **Cluster 2** reflects slower progress and even regression in several areas, particularly education and reproductive health. Conversely, **Cluster 3** exhibits relatively balanced and accelerating growth, especially in sanitation, crime reduction, and poverty alleviation. These findings underscore the importance of targeted regional policies that address both the speed and direction of development trajectories to promote equitable progress nationwide.
4. To improve the social and welfare conditions of lower-performing provinces in **Cluster 2**, targeted efforts are needed to enhance key indicators, particularly: Access to higher education (X_2), health infrastructure (X_3), child immunization coverage (X_4), sanitation (X_6), clean water access (X_7), Poverty reduction (X_9), and Human Development Index (X_{10}).

Author Contributions

Kristuisno Martsuyanto Kapiluka: Conceptualization, methodology, Writing-Original Draft, Software, Validation. Hari Wijayanto: Data Curation, Resources, Draft Preparation. Anwar Fitriantor: Formal Analysis, Validation. All authors discussed the results and contributed to the final manuscript.

Funding Statement

The authors gratefully acknowledge the financial support provided by the Indonesia Endowment Fund for Education (LPDP), which made this research possible.

Acknowledgment

The authors would like to thank all editors for their helpful guidance and support during the preparation and publication of this article.

Declarations

There are no conflicts of interest to report study.

REFERENCES

[1] Ö. Akay and G. Yüksel, "CLUSTERING THE MIXED PANEL DATASET USING GOWER'S DISTANCE AND K-PROTOTYPES ALGORITHMS," *Commun. Stat. - Simul. Comput.*, vol. 47, no. 10, pp. 3031–3041, Nov. 2018. doi: https://doi.org/10.1080/03610918.2017.1367806.

[2] E. U. Oti, M. O. Olusola, F. C. Eze, and S. U. Enogwe, "COMPREHENSIVE REVIEW OF K-MEANS CLUSTERING ALGORITHMS," *Int. J. Adv. Sci. Res. Eng.*, vol. 07, no. 08, pp. 64–69, 2021. doi: https://doi.org/10.31695/IJASRE.2021.34050.

[3] P. Jiang, J. Cao, W. Yu, and F. Nie, "A ROBUST ENTROPY REGULARIZED K-MEANS CLUSTERING ALGORITHM FOR PROCESSING NOISE IN DATASETS," *Neural Comput. Appl.*, vol. 37, pp. 6617–6632, Jan. 2025. doi: https://doi.org/10.1007/s00521-024-10899-4.

[4] X. Ao, Y. Zhang, Y. Zhou, and D. Xu, "RESEARCH ON WEIGHTED CLUSTER ANALYSIS METHOD OF PANEL DATA," *J. Phys. Conf. Ser.*, vol. 1848, no. 1, p. 12036, 2021. doi: https://doi.org/10.1088/1742-6596/1848/1/012036.

[5] J. Hu and S. Szymczak, "A REVIEW ON LONGITUDINAL DATA ANALYSIS WITH RANDOM FOREST," *Brief. Bioinform.*, vol. 24, no. 2, p. bbad002, Mar. 2023. doi: https://doi.org/10.1093/bib/bbad002.

[6] C. Genolini and B. Falissard, "KML: A PACKAGE TO CLUSTER LONGITUDINAL DATA," *Comput. Methods Programs Biomed.*, vol. 104, no. 3, pp. e112–e121, 2011, doi: https://doi.org/10.1016/j.cmpb.2011.05.008.

[7] N. G. P. Den Teuling, S. C. Pauws, and E. R. van den Heuvel, "A COMPARISON OF METHODS FOR CLUSTERING LONGITUDINAL DATA WITH SLOWLY CHANGING TRENDS," *Commun. Stat. Simul. Comput.*, vol. 52, no. 3, pp. 621–648, 2020. doi: https://doi.org/10.1080/03610918.2020.1861464.

[8] C. Genolini, X. Alacoque, M. Sentenac, and C. Arnaud, "KML AND KML3D: R PACKAGES TO CLUSTER LONGITUDINAL DATA," *J. Stat. Softw.*, vol. 65, no. 4 SE-Articles, pp. 1–34, Jun. 2015. doi: https://doi.org/10.18637/jss.v065.i04.

[9] S. Mullin *et al.*, "LONGITUDINAL K-MEANS APPROACHES TO CLUSTERING AND ANALYZING EHR OPIOID USE TRAJECTORIES FOR CLINICAL SUBTYPES," *J. Biomed. Inform.*, vol. 122, no. July, p. 103889, 2021. doi: https://doi.org/10.1016/j.jbi.2021.103889.

[10] X. Lu *et al.*, "*HEPATITIS B ANTIBODY TRAJECTORIES IN MEDICAL SCHOOL STUDENTS: AN EMPIRICAL COMPARISON OF LONGITUDINAL CLUSTERING METHODS*". *Research Square* Preprint 2025. doi: https://doi.org/10.21203/rs.3.rs-4899940/v1.

[11] S. Wahl *et al.*, "COMPARATIVE ANALYSIS OF PLASMA METABOLOMICS RESPONSE TO METABOLIC CHALLENGE TESTS IN HEALTHY SUBJECTS AND INFLUENCE OF THE FTO OBESITY RISK ALLELE," *Metabolomics*, vol. 10, Jun. 2014. doi: https://doi.org/10.1007/s11306-013-0586-x.

[12] A. Nas, S. Mulatsih, and M. Findi, "REGIONAL CLUSTERING BASED ON HDI COMPONENTS IN INDONESIA," *Int. J. Sci. Res. Sci. Eng. Technol.*, vol. 4099, pp. 21–25, 2021. doi: https://doi.org/10.32628/IJSRSET21813.

[13] A. Degirmenci and O. Karal, "EFFICIENT DENSITY AND CLUSTER BASED INCREMENTAL OUTLIER DETECTION IN DATA STREAMS," *Inf. Sci. (Ny).*, vol. 607, pp. 901–920, 2022. doi: https://doi.org/10.1016/j.ins.2022.06.013.

[14] E. Herman, K.-E. Zsido, and V. Fenyves, "CLUSTER ANALYSIS WITH K-MEAN VERSUS K-MEDOID IN FINANCIAL PERFORMANCE EVALUATION," 2022. doi: https://doi.org/10.3390/app12167985.

[15] N. R. Pradana Ratnasari, "COMPARATIVE STUDY OF K-MEAN, K-MEDOID AND HIERARCHICAL CLUSTERING USING DATA OF TUBERCULOSIS INDICATORS IN INDONESIA," *Indones. J. Life Sci.*, vol. 5, no. 2, pp. 9–20, 2023. doi: https://doi.org/10.54250/ijls.v5i02.181.

[16] N. F. Fahrudin and R. Rindiyani, "COMPARISON OF K-MEDOIDS AND K-MEANS ALGORITHMS IN SEGMENTING CUSTOMERS BASED ON RFM CRITERIA," *E3S Web Conf.*, vol. 484, 2024. doi: https://doi.org/10.1051/e3sconf/202448402008.

[17] T. Caliński and H. JA, "A DENDRITE METHOD FOR CLUSTER ANALYSIS," *Commun. Stat. - Theory Methods*, vol. 3, pp. 1–27, Jan. 1974. doi: https://doi.org/10.1080/03610927408827101.

[18] B. P. Statistik, "STATISTIK INDONESIA," www.bps.go.id. Accessed: Apr. 24, 2024. [Online]. Available: www.bps.go.id

[19] N. Aini, A. Lestari, M. N. Hayati, F. Deny, and T. Amijaya, "ANALISIS CLUSTER PADA DATA KATEGORIK DAN NUMERIK DENGAN PENDEKATAN CLUSTER ENSEMBLE (STUDI KASUS : PUSKESMAS DI PROVINSI KALIMANTAN TIMUR KONDISI DESEMBER 2017)," *J. EKSPONENSIAL Vol. 11*, vol. 11, pp. 117–126, 2020. doi: https://doi.org/10.30872/eksponensial.v11i2.652

[20] R. Juan, "FUSION CLUSTERING ANALYSIS OF MULTIVARIATE PANEL DATA," *J. Appl. Stat. Manag.*, 2013, [Online]. Available: https://api.semanticscholar.org/CorpusID:124374320

[21] D. Puspitasari, M. Wahyudi, M. Rizaldi, A. Nurhadi, K. Ramanda, and Sumanto, "K-MEANS ALGORITHM FOR CLUSTERING THE LOCATION OF ACCIDENT-PRONE ON THE HIGHWAY," *J. Phys. Conf. Ser.*, vol. 1641, no. 1, 2020. doi: https://doi.org/10.1088/1742-6596/1641/1/012086.

[22] G. Puentes, "COMPARISON BETWEEN NEURAL NETWORK CLUSTERING, HIERARCHICAL CLUSTERING, AND K-MEANS CLUSTERING: APPLICATIONS USING FLUIDIC LENSES," *Opt. Express*, vol. 33, no. 13, pp. 28405–28419, 2025. doi: https://doi.org/10.1364/OE.566535.

[23] T. M. Ghazal *et al.*, "PERFORMANCES OF K-MEANS CLUSTERING ALGORITHM WITH DIFFERENT DISTANCE METRICS," *Intell. Autom. Soft Comput.*, vol. 30, no. 2, pp. 735–742, 2021. doi: https://doi.org/10.32604/iasc.2021.019067.

[24] P. Arora, Deepali, and S. Varshney, "ANALYSIS OF K-MEANS AND K-MEDOIDS ALGORITHM FOR BIG DATA," *Procedia Comput. Sci.*, vol. 78, pp. 507–512, 2016. doi: https://doi.org/10.1016/j.procs.2016.02.095.

[25] E. Schubert and P. J. Rousseeuw, "FASTER K-MEDOIDS CLUSTERING: IMPROVING THE PAM, CLARA, AND CLARANS ALGORITHMS BT - SIMILARITY SEARCH AND APPLICATIONS," in Proceedings of the 12th International Conference on Similarity Search and Applications (SISAP 2019), G. Amato, C. Gennaro, V. Oria, and M. Radovanović, Eds., Lecture Notes in Computer Science, vol. 11807. Cham, Switzerland: Springer, 2019, pp. 171–187. doi: https://doi.org/10.1007/978-3-030-32047-8_16

[26] N. Sureja, B. Chawda, and A. Vasant, "AN IMPROVED K-MEDOIDS CLUSTERING APPROACH BASED ON THE CROW SEARCH ALGORITHM," *J. Comput. Math. Data Sci.*, vol. 3, no. July 2021, p. 100034, 2022. doi: https://doi.org/10.1016/j.jcmds.2022.100034.

[27] J. Baarsch and M. E. Celebi, "INVESTIGATION OF INTERNAL VALIDITY MEASURES FOR K-MEANS CLUSTERING," *Lect. Notes Eng. Comput. Sci.*, vol. 2195, pp. 471–476, 2012.

[28] A. Nowak-Brzezińska and I. Gaibei, "HOW THE OUTLIERS INFLUENCE THE QUALITY OF CLUSTERING?," *Entropy*, vol. 24, no. 7, 2022. doi: https://doi.org/10.3390/e24070917.