**ComTech: Computer, Mathematics and Engineering Applications**, 16(1), June 2025, 23−33 P-ISSN: 2087-1244
**DOI:** 10.21512/comtech.v16i1.12274 E-ISSN: 2476-907X

# The Implementation of the Fuzzy C-Means Method in Handling Outlier Data in the 2021 Village Potential Data of Bengkulu Province

**Intan Juliana Panjaitan[1], Indahwati[2]*, and Farit Mochamad Afendi[3]**

[1-3]Departemen Statistika, FMIPA, IPB University
Jawa Barat, Indonesia 16680
[1]julianaintan@apps.ipb.ac.id; [2]indahwati@apps.ipb.ac.id; [3]fmafendi@apps.ipb.ac.id

**Received:** 30th September 2024/ **Revised:** 11th December 2024/ **Accepted:** 11th December 2024

**How to Cite:** Panjaitan, I., Indahwati, & Afendi, F. M. (2025). The implementation of the Fuzzy C-Means method in handling outlier data in the 2021 Village Potential Data of Bengkulu Province. *ComTech: Computer, Mathematics and Engineering Applications, 16*(1), 23−33. https://doi.org/10.21512/comtech.v16i1.12274

***Abstract*** **-** Clustering groups aims to ensure similarity within clusters and disparity between them. The research evaluated the Fuzzy C-Means method's effectiveness in clustering large datasets containing outliers, focusing on the 2021 Village Potential data from Bengkulu Province. The dataset, comprising 1,514 observations from villages and urban villages, provided a comprehensive resource for understanding regional development. Outliers, a common challenge in cluster analysis, were detected using univariate and multivariate methods, revealing substantial variability. PCA was applied, improving clustering quality to address multicollinearity among variables. In the results, the fuzzifier (w) parameter in the FCM method plays a crucial role in controlling the degree of membership for data points in clusters, which can potentially reduce the impact of outliers, enhancing clustering robustness and accuracy. The FCM method effectively produces clusters with high intra-cluster homogeneity and inter-cluster heterogeneity. Using the Elbow method, three optimal clusters are identified. Cluster 1, dominated by villages in Bengkulu City, is the most advanced, with superior infrastructure and services, but the fewest villages business units, necessitating economic empowerment. Cluster 2, comprising villages in North Bengkulu Regency, demonstrates moderate development but suffers from poor transportation access, requiring improvements to support socio-economic activities. Cluster 3, dominated by villages in Kaur Regency, is the least developed, with limited basic services and infrastructure, highlighting the need for substantial investments in governance and essential services. These findings provide actionable insights for village development in Bengkulu Province, supporting targeted policies tailored to each cluster's unique characteristics.

***Keywords*****:** Fuzzy C-Means (FCM) method, outlier data, village potential data

## I. INTRODUCTION

Cluster analysis is a statistical method that identifies groups of objects based on shared characteristics. It categorizes similar elements as research subjects, where these elements exhibit high homogeneity within clusters and significant heterogeneity between clusters (Hennig, 2019). Cluster analysis can be approached in two ways: hierarchical and non-hierarchical (partitioning). The hierarchical approach organizes observation objects in a structured manner based on their similarities, while the non-hierarchical approach assigns objects to predefined clusters (Oti et al., 2021).

A common challenge in cluster analysis is the presence of outlier data points that significantly deviate from the general pattern of other observations (Bieber et al., 2023). Outliers can distort analysis results, leading to clusters that do not accurately represent the data (Nowak-Brzezińska & Łazarz, 2021). Although the simplest solution is to remove outliers, this is not always ideal. Alternative methods are needed to handle outliers without discarding them. According to Wu (2012) in Zhou and Yang (2019), the Fuzzy C-Means method is also robust against outliers, particularly when using an optimal weighting value ($w$). It is also further demonstrated that the Fuzzy C-Means method outperforms the K-Means method in handling outlier data (Kenger et al., 2023).

Besides outliers, another challenge in cluster analysis is dealing with large datasets. Inefficient methods can result in lengthy computation times, making it essential to choose a clustering method that

*Corresponding Author

is both effective and efficient. Previous research shows the effectiveness of the Fuzzy C-Means technique in clustering of the banking data. The technique proves to be a valuable tool for analyzing large and complex datasets, offering advantages over traditional clustering algorithms (Choudhary & Saxena, 2023). Fuzzy C-Means determines the inclusion of each observation in a cluster based on the degree of membership. Fuzzy C-Means is popular for clustering due to its effectiveness with large datasets, robustness against outliers, and superior performance compared to K-Means (Hassan et al., 2020). Due to its flexibility and robustness for ambiguity, Fuzzy C-Means has been widely used (S. Zhou et al., 2021). It is found that the Fuzzy C-Means algorithm shows better performance than the K-Means algorithm (Ahmadov, 2023).

The research evaluates the performance of the Fuzzy C-Means method in clustering large datasets containing outliers. Despite being well-established, Fuzzy C-Means remains relevant due to its strengths in addressing challenges in cluster analysis. Its stability, ease of implementation and interpretation, and overall effectiveness are key factors behind its continued use. The research is expected to provide insights into the reliability of Fuzzy C-Means in handling complex data conditions and to serve as a reference for selecting appropriate clustering methods for large-scale data analysis, particularly those involving outliers.

## II. METHODS

In the research, the data utilized are sourced from the 2021 Village Potential data provided by Badan Pusat Statistik (BPS), which plays a crucial role in understanding the socio-economic dynamics of rural areas. The observation unit encompasses all villages and urban villages in Bengkulu Province, amounting to 1,514 distinct villages and urban areas. The selection of research variables is meticulously based on prior studies that conducted comprehensive cluster analyses (Azrahwati et al., 2022; Chrisinta et al., 2020) and is further refined according to the parameters set forth in the 2018 Village Development Indeks (Badan Pusat Statistik, 2019, 2022), which encompasses five key dimensions that are essential for evaluating village development.

The data for the research is sourced from Statistics Indonesia (Badan Pusat Statistik (BPS)), specifically the 2021 Village Potential dataset for Bengkulu Province. Bengkulu's diverse geographical characteristics, including urban areas, remote rural areas, and border regions, often result in significant disparities in access to basic infrastructure, public services, and economic facilities among villages. Village Potential refers to the resources and capabilities of a village that can be leveraged to promote development and improve community welfare (Supandi et al., 2020). Clustering Village Potential data can provide valuable insights for deeper analysis and inform Village Potential data includes 1,514 village/urban village observations, representing the large datasets typically encountered in statistical analysis. Given the indication of outliers in this data, it is used for empirical data analysis in the research.

Consequently, the variables employed are predominantly numerical and represent critical facets of village development that are integral to policy formulation and resource allocation. These variables encompass basic services, which signify the aspect of fulfilling fundamental needs, including the number of educational facilities (X1), health facilities (X2), and health workers (X3), which are available within each village. Additionally, infrastructure conditions represent both basic and supplementary needs related to the availability of economic infrastructure, energy resources, communication systems, and information technology, which include the number of economic facilities and infrastructure (X4), families utilizing electricity (X5), and Base Transceiver Stations (BTS) (X6). Accessibility and transportation involve access to transportation facilities and infrastructure that support the village's socio-economic activities, specifically, travel time from the village head office to the sub-district office (X7) and travel time from the village head office to the regent or mayor office (X8). Public services reflect efforts to meet service needs related to community activities, particularly in terms of sports facilities. They are represented by the number of poverty certificates issued by the village in 2020 (X9). Lastly, governance reflects the performance of village governance and its support mechanisms, represented by the number of village business units (X10) and village government officials (X11).

The Fuzzy C-Means method, utilized to group the villages based on these diverse variables effectively, is an innovative approach in cluster analysis. Fuzzy C-Means is a non-hierarchical method where the degree of membership is crucial in determining the inclusion of each data point in a specific cluster (Singh et al., 2023). This method is frequently employed in clustering because it produces smooth and effective results, has the capacity to handle larger datasets, and demonstrates robustness against outliers. The Manhattan Distance is applied to detect potential outliers in the data. It measures the absolute distance between each data point and others, helping to identify points that significantly deviate from the general pattern.

The Fuzzy C-Means method is utilized to effectively group the villages based on these diverse variables, which is an innovative approach in cluster analysis. Fuzzy C-Means is a non-hierarchical method where the degree of membership is crucial in determining the inclusion of each data point in a specific cluster (Singh et al., 2023). This method is frequently employed in clustering because it produces smooth and effective results, has the capacity to handle larger datasets, and demonstrates robustness against outliers. The effectiveness of the Fuzzy C-Means method is significantly influenced by the choice of the weighting exponent ($w$), a topic that has

been thoroughly investigated by researchers such as Pal and Bezdek (1995) in Abdellahoum et al. (2021) and Wang et al. (2021). Specifically, previous research suggests using a weighting value of $w \in [1.5, 2.5]$. In contrast, another previous research proposes an upper theoretical limit for $w$, which is essential in preventing the sample mean from becoming the unique optimization of the Fuzzy C-Means objective function. The Fuzzy C-Means algorithm adheres to a series of systematic steps to cluster data efficiently. Initially, the number of clusters ($k$) to be formed is determined, with the prerequisite that $k \geq 2$. Furthermore, a suitable weighting value ($w$) is selected such that $w > 1$. Subsequently, random membership values $u_{ip}$ are generated for each data point across the clusters, following the condition outlined in Eq. (1). It shows that $u_{ip}$ is the $i$ random number in the $p$ cluster. The cluster centers for each cluster are computed using Eq. (2), after establishing membership values. Here, $c_{pj}$ represents the center of the $p$ cluster on the $j$ variable, $u_{ip}$ is the $i$ random number in the $p$ cluster, $x_{ij}$ is the $i$ object standardized on the $j$ variable, and $w$ is the weighting exponent. Following this, the membership degree for each object in each cluster is computed utilizing Eq. (3). The distance between the $i$ object and the $p$ cluster center, $d(x_i, c_p)$ is calculated according to Eq. (4). Subsequently, objects are assigned to clusters based on the nearest distance, employing Eq. (5).

$$\sum_{p=1}^{k} u_{ip} = 1 \tag{1}$$

$$c_{pj} = \frac{\sum_{i=1}^{n} (u_{ip}{}^{w} x_{ij})}{\sum_{i=1}^{n} (u_{ip})^{w}} \tag{2}$$

$$u_{ip} = \left[ \frac{d(x_i, c_p)^{\frac{-2}{w-1}}}{\sum_{p=1}^{k} d(x_i, c_p)^{\frac{-2}{w-1}}} \right] \tag{3}$$

$$d(x_i, c_p) = \left[ \sum_{j=1}^{l} d(x_{ij} - c_{pj})^2 \right]^{\frac{1}{2}} \tag{4}$$

$$w_{ip} = \begin{cases} 1, & if\ u_{ip} = \max(u_{i1}, u_{i2}, \dots, u_{ip}) \\ 0, & otherwise \end{cases} \tag{5}$$

These steps are iteratively repeated until the cluster centers stabilize and no members switch clusters (Mahmudi et al., 2021). The iteration process ceases when the change in membership values between iterations is less than a predefined threshold $\varepsilon$, $max_{ip}\{|u_{ip}^{k+1} - u_{ip}^{k}|\} < \varepsilon$. The $\varepsilon$ is usually set to $\varepsilon = 1 \times 10^{-3}$.

Once the Fuzzy C-Means clustering method has been effectively applied, the subsequent step is to conduct a comprehensive and detailed analysis of the resulting data to generate the desired clusters. This analysis is carried out using R version 4.3.1, along with several supporting packages designed to enhance data handling and visualization capabilities. Conducting thorough data exploration is the initial step of the data analysis procedure, ensuring that the research methodology and findings can be thoroughly understood and replicated based on the provided details. Principal Component Analysis (PCA) is applied to address multicollinearity issues among variables, thereby improving the stability and interpretation of the clustering results. After that, the optimal number of clusters is determined using the Elbow method, followed by applying the Fuzzy C-Means method to cluster the data. The Kruskal-Wallis test is conducted to assess whether the selected variables significantly distinguish the clusters to further validate the differences between clusters. Subsequently, the weighting exponent ($w$) is optimized within the range recommended by Pal and Bezdek, $w \in [1.5, 2.5]$, to ensure effective and stable clustering. The clustering results are then interpreted and analyzed in depth, culminating in the formulation of conclusions drawn from the comprehensive analysis. The research stages are explained in Figure 1 (see Appendices).

## III. RESULTS AND DISCUSSIONS

The data analyzed are the 2021 Village Potential data, with the unit of analysis being villages and urban villages in Bengkulu Province, totaling 1,514 units comprising 1,342 villages and 172 urban villages. The analysis begins with data exploration to determine if the data contain outliers. Outlier detection is conducted both univariately and multivariately using Mahalanobis distance. The univariate detection reveals an average outlier percentage of 7.94% across the 11 variables, while Mahalanobis distance detection shows an outlier percentage of 8.52%. Figure 2 (see Appendices) demonstrates that all variables contain outliers, as shown by the box plot. It indicates significant diversity in Bengkulu Province's geography, economy, and development, contributing to the high variability observed.

Following the comprehensive detection of outliers in the dataset, the subsequent step involves a meticulous examination of the correlation between the various variables under consideration. Figure 3 illustrates that most variable combinations show low correlation values, indicating a weak relationship according to the Pearson correlation method. Furthermore, it is noteworthy that some variables are entirely uncorrelated. For instance, there is no significant correlation between X9, which represents the number of poverty letters issued by the village in 2020, and X7, which denotes the travel time from the village head's office to the sub-district office. In addition, Figure 3 (see Appendices) also reveals that certain variables demonstrate relatively high correlation values, specifically those that reach or exceed the threshold of 0.7 ($\rho \geq 0.7$). Among these, the most substantial correlation is identified between X5, representing the number of households utilizing electricity, and X1, which indicates the total number

of educational facilities within the village, with a remarkably high correlation coefficient of 0.79. This significant correlation suggests that as the number of households using electricity increases, there tends to be a corresponding increase in the total number of educational facilities available in the village, reflecting an underlying relationship between energy access and educational infrastructure. Travel time from the village head office to the sub-district office (X7) and the number of poverty certificates issued by the village in 2020 do not have a correlation. It means that factors influencing travel time (like infrastructure or geography) do not directly affect the number of poverty certificates, which are more likely influenced by other socio-economic factors.

High correlation between variables can pose a problem in cluster analysis by causing multicollinearity. Multicollinearity can reduce the stability of the analysis results, as highly correlated variables provide redundant information without adding new insights. PCA is used to eliminate correlations between variables. PCA transforms the highly correlated variables into a set of uncorrelated Principal Components (PCs), thereby stabilizing the cluster analysis results and making them easier to interpret.

Table 1 presents the eigenvalue decomposition obtained from the PCA. It shows that PC1, PC2, PC3, PC4, and PC5 explain the variance of the original variables as follows: 44.08%, 10.25%, 9.04%, 8.29%, and 7.91%, respectively. The cumulative proportion of variance explained by these five components is 79.57%. These five components adequately represent the 11 variables, as they account for 79.57% of the data variance. The results of the PCA show the loadings for the five principal components, detailed in Table 2. Each row represents the original variables (X1 to X11), and each column represents the loadings of these variables on the corresponding principal components.

The information obtained from Table 2 provides insights into the relationships between the original variables and the PCs derived from the analysis. PC1 indicates that higher values are associated with a lower number of households using electricity (X5), fewer educational facilities (X1), fewer economic infrastructure facilities (X4), fewer health facilities (X2), and fewer village government officials (X11). In contrast, PC2 shows that higher values correlate with shorter travel times from the village head's office to both the sub-district office (X7) and the regent/mayor's office (X8). Similarly, PC3 suggests that higher values are linked to fewer village business units (X10). On the other hand, PC4 reveals that higher values are associated with longer travel times from the village head's office to the regent/mayor's office (X8) and a lower number of village business units (X10). Finally, PC5 indicates that higher values are related to shorter travel times from the village head's office to the sub-district office (X7) and fewer health workers (X3). This analysis of PCs highlights the nuanced interrelationships between the variables in the dataset. Base Transceiver Stations (X6) and the number of poverty certificates issued by the village in 2020 (X9) show relatively low loadings across the five PCs. X6 contributes most to PC4 (0.258), and X9 to PC4 (-0.309), but neither is strong enough to be considered a key differentiating factor. The result suggests that their influence on village characteristics is limited in the PCA structure.

The elbow method is meticulously applied to determine the optimal number of clusters for the analysis. As illustrated in Figure 4 (see Appendices), the most significant change in the graph, which resembles the shape of an elbow, occurs specifically when $k$=3. This observation indicates a pivotal moment in the clustering process, as the rate of decrease in the Sum of Squared Errors (SSE) becomes less pronounced beyond this point. After reaching $k$=3, the subsequent addition of more clusters leads to a more gradual decline in the SSE, suggesting that the clusters become increasingly similar to one another, resulting in a diminishing return on the quality of clustering. Therefore, the

Table 1 Eigenvalue Decomposition from the Principal Component Analysis (PCA)

| Principal Component | Eigenvalue | Variance Proportion (%) | Cumulative Variance Proportion (%) |
|---|---|---|---|
| 1 | 4.85 | 44.08 | 44.08 |
| 2 | 1.13 | 10.25 | 54.33 |
| 3 | 0.99 | 9.04 | 63.37 |
| 4 | 0.91 | 8.29 | 71.66 |
| 5 | 0.87 | 7.91 | 79.57 |
| 6 | 0.66 | 6.00 | 85.57 |
| 7 | 0.44 | 4.02 | 89.59 |
| 8 | 0.40 | 3.68 | 93.27 |
| 9 | 0.31 | 2.78 | 96.05 |
| 10 | 0.26 | 2.35 | 98.40 |
| 11 | 0.18 | 1.60 | 100.00 |

optimal number of clusters chosen for the analysis is three, as this particular value strikes a commendable balance between effective cluster formation and the preservation of essential information. Choosing three clusters allows for meaningful differentiation among the data points while avoiding overfitting, ensuring that the analysis remains insightful and manageable.

The clustering analysis results are presented in Figure 5 (see Appendices), which displays the clustering plot generated using the Fuzzy C-Means method. In this analysis, the optimal mm value obtained is 1.5, with a within-cluster to between-cluster ratio of 0.00043. This value indicates that the Fuzzy C-Means method effectively separates the clusters. It is evident from the plot, which demonstrates high homogeneity within each cluster and significant heterogeneity between clusters. Thus, this clustering successfully identifies distinct groups of villages based on the analyzed variables.

To confirm the statistical validity of the clustering, a Kruskal-Wallis test is conducted to test whether the variables distinguish between the clusters. The null hypothesis (H0) asserts that the variables do not distinguish between clusters, while the alternative hypothesis (H1) asserts that the variables distinguish between clusters. If the significance value is less than 0.05, H0 is rejected. Table 3 shows that all variables reveal significant differences between clusters, with p-values less than 0.05. The results indicate substantial differences in these variables across clusters. These results are consistent with the descriptive analysis presented in Figure 6 (see Appendices). Figure 6 shows the distribution of variable values for each cluster using box plots, demonstrating significant differences in these variable values across the clusters.

Table 4 provides additional information on the average values for each variable in the clusters produced by the Fuzzy C-Means method. The clustering results

Table 2 Loadings for the Five Principal Components (PCs)

| | PC1 | PC2 | PC3 | PC4 | PC5 |
|---|---|---|---|---|---|
| X1 | -0.398 | -0.037 | -0.026 | 0.012 | 0.021 |
| X2 | -0.381 | -0.018 | -0.126 | 0.064 | -0.046 |
| X3 | -0.244 | 0.180 | -0.337 | 0.336 | -0.484 |
| X4 | -0.389 | 0.025 | -0.044 | 0.058 | 0.006 |
| X5 | -0.412 | -0.059 | 0.040 | -0.098 | 0.104 |
| X6 | -0.277 | 0.059 | -0.171 | 0.258 | -0.195 |
| X7 | 0.037 | -0.692 | 0.120 | -0.266 | -0.652 |
| X8 | 0.051 | -0.641 | -0.094 | 0.640 | 0.393 |
| X9 | -0.324 | -0.136 | 0.255 | -0.309 | 0.230 |
| X10 | 0.083 | -0.166 | -0.860 | -0.424 | 0.196 |
| X11 | -0.351 | -0.147 | 0.097 | -0.220 | 0.209 |

Table 3 Kruskal-Wallis Test Results for Significant Differences between Variables in the Fuzzy C-Means Method

| Variable | Chi Squared | Degrees of Freedom | P-Value |
|---|---|---|---|
| X1 | 600.16 | 2 | 0.00 |
| X2 | 600.72 | 2 | 0.00 |
| X3 | 494.11 | 2 | 0.00 |
| X4 | 705.81 | 2 | 0.00 |
| X5 | 805.22 | 2 | 0.00 |
| X6 | 559.19 | 2 | 0.00 |
| X7 | 39.54 | 2 | 0.00 |
| X8 | 62.01 | 2 | 0.00 |
| X9 | 227.91 | 2 | 0.00 |
| X10 | 73.03 | 2 | 0.00 |
| X11 | 392.92 | 2 | 0.00 |

Table 4 Average Values of Each Variable for Each Cluster

| | Cluster 1 | Cluster 2 | Cluster 3 |
|---|---|---|---|
| X1 | 12.32 | 4.27 | 2.12 |
| X2 | 9.04 | 2.68 | 1.04 |
| X3 | 17.06 | 6.76 | 2.35 |
| X4 | 93.71 | 30.68 | 11.93 |
| X5 | 1648.28 | 491.17 | 217.87 |
| X6 | 1.80 | 0.82 | 0.10 |
| X7 | 12.13 | 17.73 | 17.52 |
| X8 | 37.35 | 74.64 | 65.29 |
| X9 | 358.80 | 59.97 | 31.14 |
| X10 | 0.38 | 1.11 | 0.97 |
| X11 | 43.99 | 20.01 | 13.75 |

are as follows. Cluster 1 consists of 94 villages/urban villages in Bengkulu Province, predominantly in Bengkulu City. According to Table 4, this cluster has the highest counts of educational facilities, health facilities, health workers, economic infrastructure, electricity users, and village government officials. It also features the shortest travel times from the village head office to the sub-district and regent/mayor offices, indicating proximity. Additionally, it has the fewest village business units.

Cluster 2 comprises 472 villages/urban villages in Bengkulu Province, mainly in North Bengkulu Regency. This cluster has a substantial number of educational facilities, health facilities, health workers, economic infrastructure, electricity users, and village government officials. It also has the longest travel times from the village head office to the sub-district and regent/mayor offices, indicating a considerable distance. Furthermore, it has the highest number of village business units.

Cluster 3 includes 948 villages/urban villages in Bengkulu Province, primarily in Kaur Regency. This cluster has the fewest educational facilities, health facilities, health workers, economic infrastructure, electricity users, and village government officials. Travel times from the village head office to the sub-district and regent/mayor offices are intermediate, suggesting moderate distances. The number of village business units is also moderate.

The distribution of villages in the optimal clusters obtained is presented in the form of a thematic map in Figure 7 (see Appendices). The area on the map is dominated by red, representing cluster 3, which is mostly located in Kaur Regency. Cluster 2, marked in yellow, is predominantly found in North Bengkulu Regency. Meanwhile, cluster 1, marked in green, is mostly located in Bengkulu City.

## IV. CONCLUSIONS

Based on the analysis of the 2021 Village Potential data, the optimal number of clusters identified is three, each with distinct characteristics: cluster 1 is the most favorable and dominated by villages/urban villages in Bengkulu City. It features the highest numbers of educational facilities, health services, economic infrastructure, electricity users, and village officials. Additionally, it has the shortest travel times to the sub-district and regent/mayor offices and the fewest village business units. Cluster 2 primarily consists of villages/urban villages in North Bengkulu Regency. This cluster has a substantial number of educational facilities, health services, economic infrastructure, electricity users, and village officials. It also experiences the longest travel times from village offices to the sub-district and regent/mayor offices and has the highest number of village business units. Cluster 3 includes villages/urban villages in Kaur Regency. This cluster is less favorable, with the fewest educational facilities, health services, economic infrastructure, electricity users, and village officials. The travel times from village offices to the sub-district and regent/mayor offices are moderate, and the number of village business units is relatively low.

The results of the clustering analysis successfully address the research problem by providing a clear and organized classification of the villages in Bengkulu Province based on various socio-economic factors. The use of the Fuzzy C-Means method has shown its ability to produce distinct and meaningful clusters. It confirms that the experiment's goal of grouping the villages effectively has been achieved.

Despite the strengths of the Fuzzy C-Means method in this analysis, several limitations should be noted. Firstly, the method's effectiveness may vary with different datasets, particularly those with varying levels of noise and outliers. The reliance on a specific range for the weighting exponent w can also influence clustering results, necessitating careful selection based on dataset characteristics. Secondly, the research findings are contextual to Bengkulu Province, limiting generalizability to other regions with different socio-economic dynamics. Hence, future research is encouraged to explore alternative clustering methods

that may offer better performance in handling datasets with outliers. It can provide more nuanced insights into village development, potentially leading to more effective policy recommendations.

## AUTHOR CONTRIBUTIONS

Conceived and designed the analysis, I. J. P., I., and F. M. A; Collected the data, I. J. P., I., and F. M. A; Contributed data or analysis tools, I. J. P., I., and F. M. A; Performed the analysis, I. J. P., I., and F. M. A; and Wrote the paper, I. J. P., I., and F. M. A.

## DATA AVAILABILITY

The data that support the findings of this research are available from Statistics Indonesia (Badan Pusat Statistik – BPS RI). Restrictions apply to the availability of these data, which were used under license for the purpose of this study. Data are available from the author (Intan Juliana Panjaitan) with written permission from BPS RI.

## REFERENCES


Abdellahoum, H., Mokhtari, N., Brahimi, A., & Boukra, A. (2021). CSFCM: An improved Fuzzy C-Means image segmentation algorithm using a cooperative approach. *Expert Systems with Applications*, *166*. https://doi.org/10.1016/j.eswa.2020.114063

Ahmadov, E. Y. (2023). Comparative analysis of K-Means and Fuzzy C-Means algorithms on demographic data using the PCA method. *Problems of Information Technology*, *14*(1), 15–22. https://doi.org/10.25045/jpit.v14.i1.03

Azrahwati, Nusrang, M., Aidid, M. K., & Rais, Z. (2022). K-Means cluster analysis for grouping districts in South Sulawesi province based on village potential. *ARRUS Journal of Mathematics and Applied Science*, *2*(2), 73–82. https://doi.org/10.35877/mathscience739

Badan Pusat Statistik. (2019, May 9). *Indeks pembangunan desa 2018*. https://www.bps.go.id/id/publication/2019/05/09/4edae4bd6c18d24b1b4273fe/ indeks-pembangunan-desa-2018.html

Badan Pusat Statistik. (2022, March 24). *Statistik potensi desa Indonesia 2021*. https://www.bps.go.id/id/publication/2022/03/24/ceab4ec9f942b1a4fdf4cd08/statistik-potensi-desa-indonesia-2021.html

Bieber, M., Verhagen, W. J. C., Cosson, F., & Santos, B. F. (2023). Generic diagnostic framework for anomaly detection—Application in satellite and spacecraft systems. *Aerospace*, *10*(8), 1–24. https://doi.org/10.3390/aerospace10080673

Choudhary, B., & Saxena, V. (2023). Fuzzy C-Mean technique for accessing large database of banking sector. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(4), 263–271.

Chrisinta, D., Sumertajaya, I. M., & Indahwati. (2020). Evaluasi kinerja metode cluster ensemble dan latent class clustering pada peubah campuran. *Indonesian Journal of Statistics and Its Applications*, *4*(3), 448–461.

Hassan, A. A. H., Shah, W. M., Othman, M. F. I., & Hassan, H. A. H. (2020). Evaluate the performance of K-Means and the Fuzzy C-Means algorithms to formation balanced clusters in wireless sensor networks. *International Journal of Electrical and Computer Engineering*, *10*(2), 1515–1523. https://doi.org/10.11591/ijece.v10i2.pp1515-1523

Hennig, C. (2019). Cluster validation by measurement of clustering characteristics relevant to the user. In C. H. Skiadas & J. R. Bozeman (Eds.), *Data analysis and applications 1: Clustering and regression, modeling-estimating, forecasting and data mining*. Wiley. https://doi.org/10.1002/9781119597568.ch1

Kenger, O. N., Kenger, Z. D., Ozceylan, E., & Mrugalska, B. (2023). Clustering of cities based on their smart performances: A comparative approach of Fuzzy C-Means, K-Means, and K-Medoids. *IEEE Access*, *11*, 134446–134459. https://doi.org/10.1109/ACCESS.2023.3333753

Mahmudi, Goejantoro, R., & Amijaya, F. D. T. (2021). Perbandingan metode C-Means dan Fuzzy C-Means pada pengelompokan kabupaten/kota di Kalimantan berdasarkan indikator IPM tahun 2019. *Jurnal EKSPONENSIAL*, *12*(2), 193–200. https://doi.org/10.30872/eksponensial.v12i2.814

Nowak-Brzezińska, A., & Łazarz, W. (2021). Qualitative data clustering to detect outliers. *Entropy*, *23*(7), 1–27. https://doi.org/10.3390/e23070869

Oti, E. U., Olusola, M. O., Eze, F. C., & Enogwe, S. U. (2021). Comprehensive review of K-Means clustering algorithms. *International Journal of Advances in Scientific Research and Engineering*, *7*(8), 64–68. https://doi.org/10.31695/ijasre.2021.34050

Singh, P., Rathee, N., Sharda, S., & Kumar, S. (2023). Comparative study of rough set-based FCM and K-Means clustering for tumor segmentation from brain MRI images. *Revue d'Intelligence Artificielle*, *37*(4), 921–927. https://doi.org/10.18280/ria.370412

Supandi, A., Saefuddin, A., & Sulvianti, I. D. (2020). Two step cluster application to classify villages in Kabupaten Madiun based on village potential data. *Xplore: Journal of Statistics*, *10*(1), 12–26.

Wang, H. Y., Wang, J. S., & Wang, G. (2021). Combination evaluation method of Fuzzy C-Mean clustering validity based on hybrid weighted strategy. *IEEE Access*, *9*, 27239–27261. https://doi.org/10.1109/ACCESS.2021.3058264

Zhou, K., & Yang, S. (2019). Fuzzifier selection in Fuzzy C-Means from cluster size distribution perspective. *Informatica*, *30*(3), 613–628. https://doi.org/10.15388/informatica.2019.221

Zhou, S., Li, D., Zhang, Z., & Ping, R. (2021). A new membership scaling Fuzzy C-Means clustering algorithm. *IEEE Transactions on Fuzzy Systems*, *29*(9), 2810–2818. https://doi.org/10.1109/TFUZZ.2020.3003441

APPENDICES

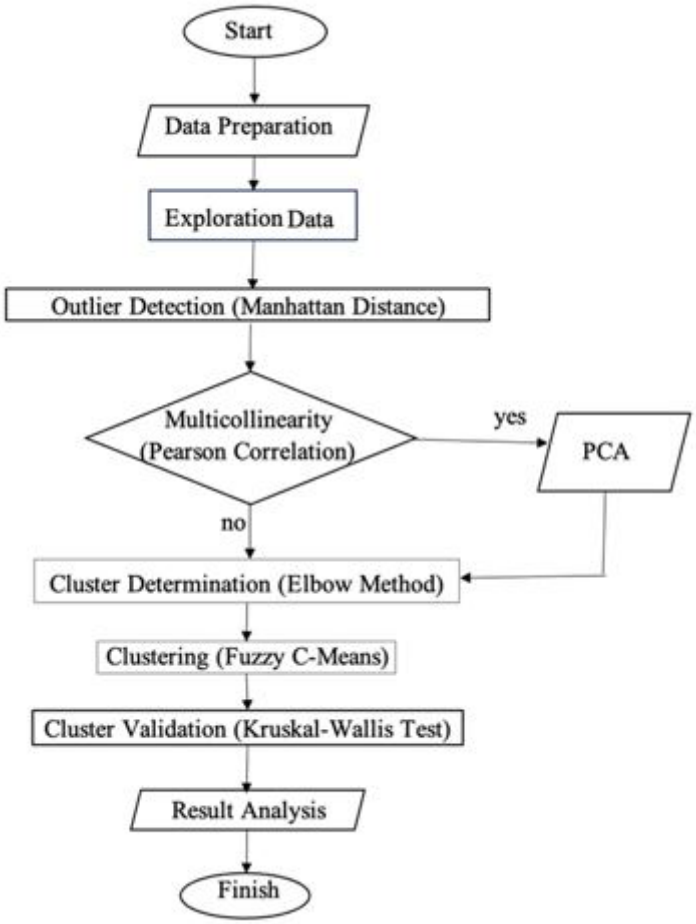


Figure 1 Research Framework. It has Principal Component Analysis (PCA).

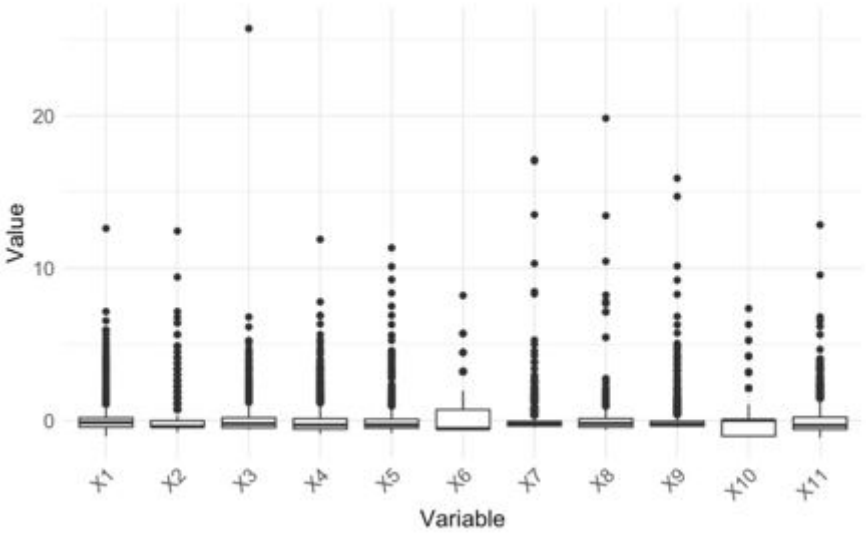


Figure 2 Box Plot of Each Variable

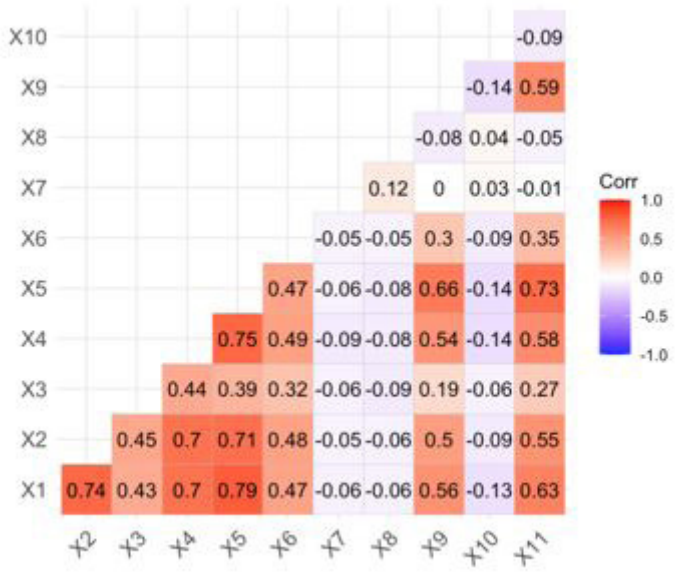


Figure 3 Pearson Correlation between Variables in the 2021 Bengkulu Province Village Potential data

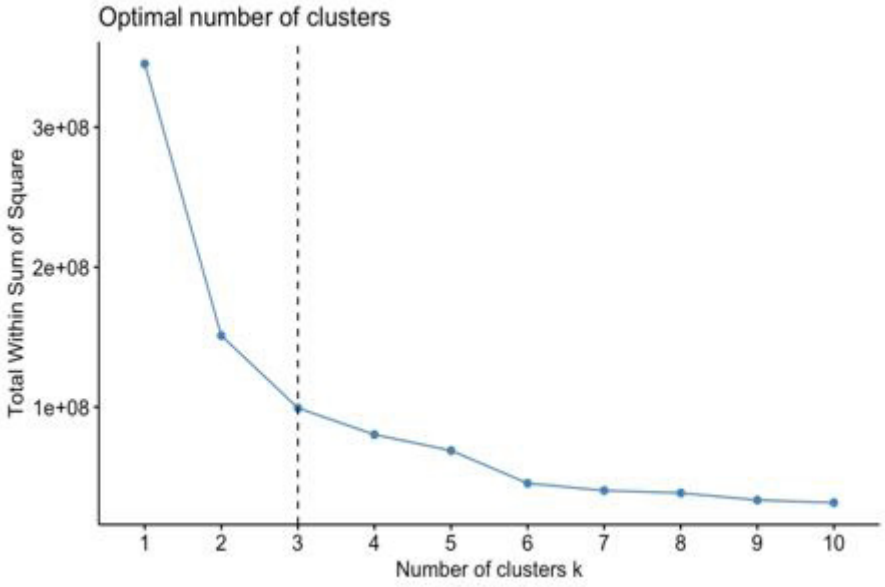


Figure 4 Optimal Number of Clusters Determined Using the Elbow Method

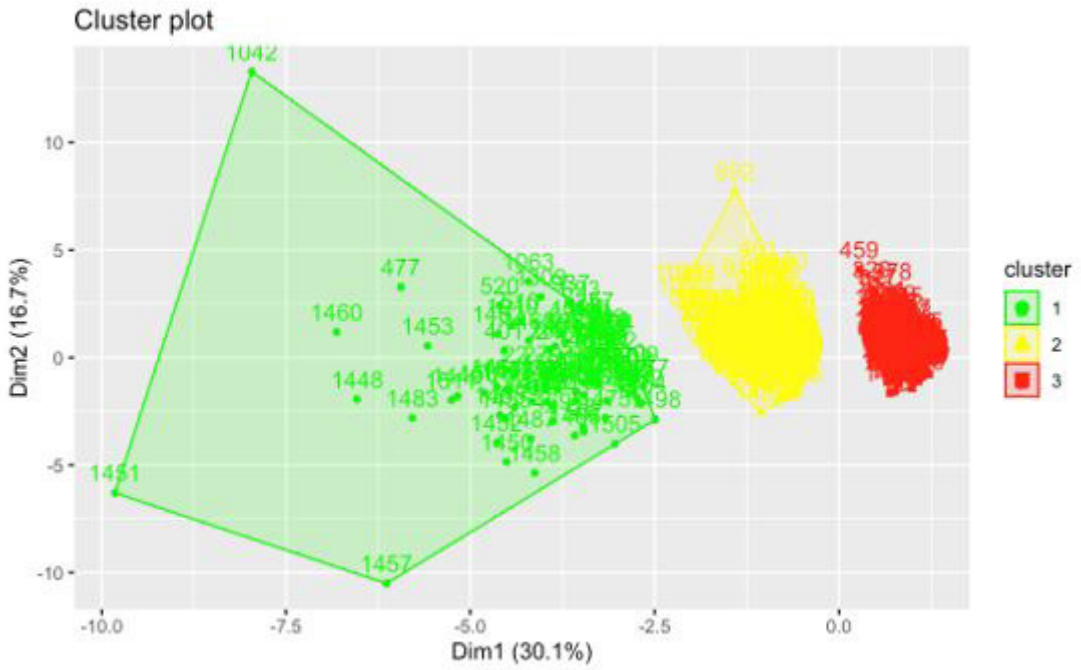


Figure 5 The Cluster Plots Obtained from the Fuzzy C-Means Method

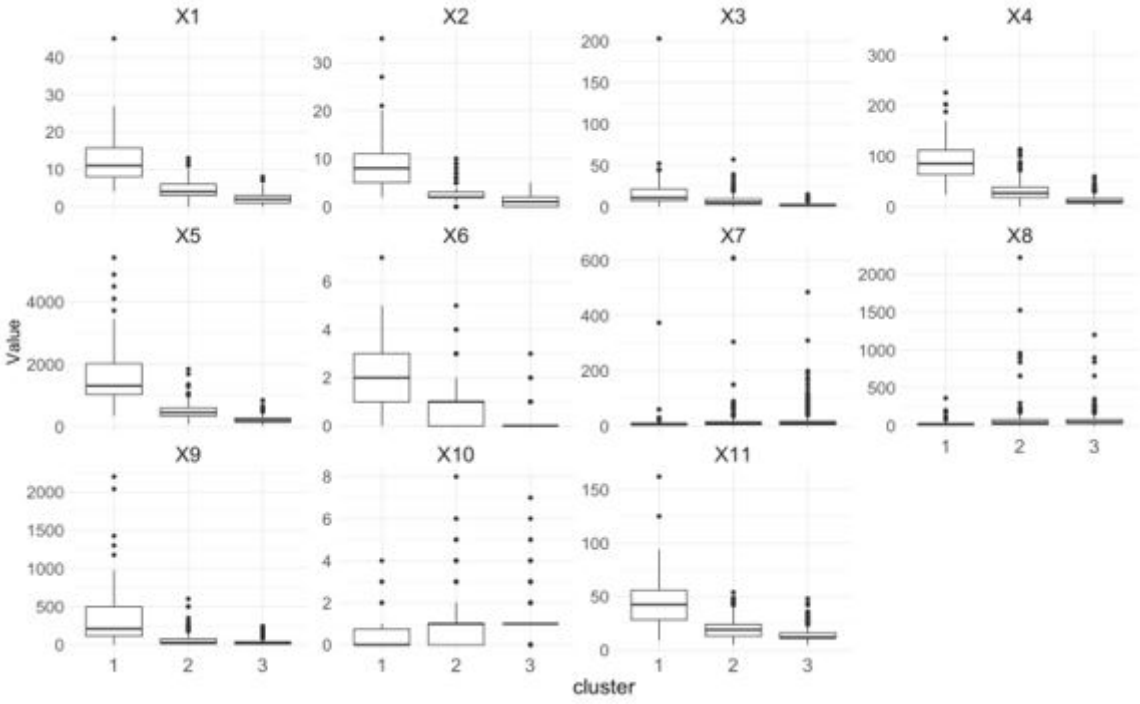


Figure 6 Box Plot of Each Variable in Each Group from the Fuzzy C-Means Analysis Results

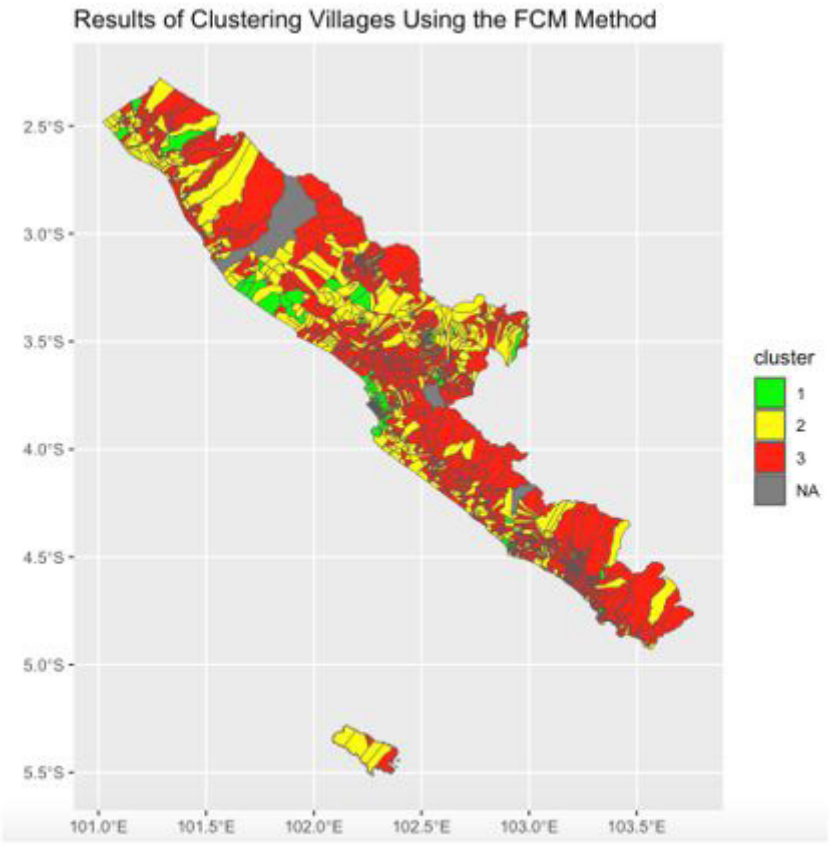


Figure 7 Thematic map of village levels in Bengkulu Province based on optimal clustering using the Fuzzy C-Means method