ICDSOS
The 3rd International Conference on Data Science and Official Statistics
November 27 - 28, 2025

"From Data to Impact: Leveraging Innovation in Data Science and Official Statistics for Data-Driven Policy Making and Global Competitiveness"

Application of K-Medoids for Regional Classification Based on Quality, Access, and Governance of Education in Indonesia

S Robiati[1,*], A Hakim[2], G F Dharmawan[3], and C Khotimah[4]

[1] Department of Statistics, Universitas Negeri Padang, Padang, Indonesia
[2,4] Pusdatin, Kementerian Pendidikan Dasar dan Menengah, Jakarta, Indonesia
[3] Kementerian Pendidikan Dasar dan Menengah, Jakarta, Indonesia

*Corresponding author's email: silfirobiati43@gmail.com

Abstract. Education is a fundamental foundation for individuals, yet substantial disparities persist across Indonesia, including both 3T (Disadvantaged, Frontier, and Outermost) and non-3T regions. Addressing the limited research on systematic regional mapping based on education indicators, this study analyzes 514 regencies/cities at the senior secondary level using 13 indicators covering three latent dimensions identified through Factor Analysis: education quality, quality of the learning process, and governance and educational participation. Data were processed through outlier detection, standardization, dimensionality reduction using Principal Component Analysis, factor score extraction, and K-Medoids clustering in RStudio. The optimal solution with three clusters was validated with a Davies–Bouldin Index of 1.44, confirming its effectiveness in capturing regional variation. Results reveal distinct spatial patterns in educational characteristics, where some 3T regions perform comparably to non-3T areas, while certain remote regions face challenges across all dimensions. These findings provide a basis for targeted, cluster-based policy interventions to improve education quality, expand access, and strengthen governance, supporting equitable educational development nationwide. The study demonstrates the utility of combining dimensionality reduction and clustering for evidence-based policy planning and highlights the importance of addressing regional disparities in education.

Keyword: Access, Cluster, Education, Governance, Quality

1. Introduction

Education is the foundation for determining one's future life. According to Article 31 paragraph 1 of the 1945 Constitution, every citizen has the right to education. Therefore, the Ministry of Primary and Secondary Education (KEMENDIKDASMEN) has various priority programmes to promote access to and quality of education in Indonesia, one of which is 13 years of compulsory education. There are three aspects that are prerequisites for achieving 13 years of compulsory education, namely access, quality, and governance.

Based on the 2024 National Socioeconomic Survey (Susenas)**,** the majority of out-of-school children (ATS) are within the 16–18-year-old age group, indicating that educational continuation at the upper secondary level remains a challenge. Furthermore, disparities in the School Participation Rate (APS) among regencies and cities persist across Indonesia. Data [1] shows that the 15 regencies/cities with the lowest APS are predominantly located in Central Papua Province. Interestingly, there are also several regencies/cities on Java Island with low APS, such as Bangkalan Regency (50.30) and Probolinggo

Regency (57.03) in East Java Province, as well as Wonosobo Regency in Central Java Province. The low APS in Wonosobo Regency is influenced by the relatively low quality of education and the high poverty rate [2]. Conversely, the highest APS was in East Java Province, namely Blitar City in East Java with an achievement of 94.58. This stark contrast indicates a significant imbalance in access to secondary education in Indonesia, not only in 3T areas but also outside 3T areas. To address this inequality in access, the government has launched the Indonesia Pintar (PIP) Programme. This programme aims to support school-age children from poor families, families at risk of poverty, or priority groups so that they continue to have access to education until they complete secondary school. In addition, this programme is also expected to prevent the risk of school dropouts and attract students who have dropped out of school to continue their education.

However, inequality in access to education in Indonesia is not only influenced by economic factors, but also by the geographical location of schools in remote areas. This situation is a major factor slowing down the distribution of educational facilities and infrastructure. In addition, limited access to transport exacerbates obstacles in the delivery of educational logistics, such as books, furniture and technological equipment to these areas [3].

It turns out that educational inequality issues like this also occur in other countries. In the Philippines, [4] found that disparities in school facilities were a major factor in educational inequality, especially in remote areas facing infrastructure limitations. This study maps the condition of educational facilities geographically and identifies patterns of disparity at the provincial level. Geographical aspects have proven to be an important factor in the establishment and supervision of nearly 60,000 schools spread across more than 7,000 islands in the Philippines.

Based on a review of the literature, research on regional grouping using indicators of quality, access, and governance in education is still limited. Therefore, this study proposes strategic mapping of regencies/cities based on these indicators at the senior secondary school level. The grouping approach enables the identification of groups of regions with similar characteristics, providing a valuable basis for planning secondary education development tailored to the needs of each cluster [5].

Clustering is a data analysis technique used to group objects according to their similarity [6]. K-Means and K-Medoids are clustering algorithms that belong to Non-Hierarchical or Partitional Clustering. K-Means is a cluster analysis method that uses the mean as the cluster centre. However, because the mean is not resistant to outliers, the K-Means algorithm becomes more sensitive to the presence of outliers. To overcome this problem, the K-Medoids method can be used to cluster data containing outliers. Unlike K-Means, K-Medoids uses medoids, which are objects located centrally within clusters, making them more resistant to outliers. Next [7] conducted research on grouping districts/cities in Indonesia using the Hierarchical, K-Means, and K-Medoids Clustering methods based on the Human Development Index (HDI). The results showed that the best method for grouping regencies/cities based on HDI was using K-Medoids with five clusters.

This study aims to identify clusters that will be formed based on indicators of quality, access, and governance of senior high school education in regencies/cities in Indonesia, as well as the characteristics of each cluster. It is hoped that this research will assist the government in identifying which regencies/cities require more attention in terms of the quality, access and governance of education, particularly at the senior secondary level.

2. Method Research

2.1. Data Sources

The data used in this study is secondary data obtained from the 2025 Education Report Card. This data consists of 514 regencies/cities in Indonesia and 13 variables used, which are presented in Table 1.

Table 1. Variable of Research.

Variable	Description	Classification	Unit of Data
X1	Literacy Skills	Quality	Persentase
X2	Numeracy Skills	Quality	Persentase
X3	Character	Quality	Persentase

X4	School Participation Rate (APS) (16-18)	Access	Persentase
X5	Net Participation Rate (APM) (16-18)	Access	Persentase
X6	Quality of Learning	Quality	Persentase
X7	Inclusive Climate	Quality	Persentase
X8	Learning Methods	Quality	Persentase
X9	Parent Participation	Governance	Persentase
X10	Student Participation	Access	Persentase
X11	Proportion of Local Government Budget Utilisation for Education	Governance	Persentase
X12	Education Unit Programmes and Policies	Governance	Persentase
X13	Percentage of Certified Teachers (Senior High School)	Quality	Persentase

The 2023 Education Report shows that the quality of education in Indonesia is measured through a number of key indicators grouped into several dimensions. In terms of quality, the indicators assessed include literacy and numeracy skills, character building of students, teaching quality and methods, inclusive school climate, and school participation rates [8]. In line with this [9] emphasises that improving the quality of education is not only determined by internal factors within schools, but is also greatly influenced by community involvement. The active role of parents in supporting their children's learning process at home is an important factor in strengthening academic success and character building. In addition, empowering schools through autonomy in resource management and policy-making allows educational units to be more flexible and responsive to local needs. Meanwhile, the success of education governance also depends heavily on the effective use of funds, whereby efficient budget allocation will ensure that improvements in the quality and accessibility of education can be achieved optimally.

2.2. Data Standardisation

Euclidean distance is among the most widely used distance measures; however, it is highly sensitive to variations in variable scales [10]. Consequently, data standardization is required when the variables differ substantially in their units of measurement. Thus, prior to conducting cluster analysis, the data must first be standardized.

$$z_{ij} = \frac{x_{ij} - \overline{x_J}}{s_j} \tag{1}$$

Description:

x_{ij}= observed value of the i-th individual on the j-th variable

s_j = standard deviation of variable j

$\bar{x}_j = mean\ of\ variable\ j$

z_{ij} = standardized value

2.3. Cluster

Cluster analysis is a multivariate technique aimed at grouping objects based on their shared characteristics. This method, often referred to as data segmentation, divides large datasets into smaller groups with similar attributes. Objects within the same cluster exhibit high similarity, while the similarity between clusters tends to be low. Thus, the approach seeks to minimize variation within clusters while maximizing differences across clusters [6]. Clustering methods are generally divided into two types. The first is hierarchical clustering, which groups objects in a structured and sequential manner based on their similarities, where the number of clusters is not predetermined. The second is non-hierarchical clustering**,** in which the number of clusters (*k*) must be specified in advance before the grouping process is carried out.

2.4. Kaiser-Meyer-Olkin (KMO)

The KMO test is used to assess the adequacy of the sample and whether it is suitable to represent the population. The KMO statistic ranges from 0 to 1. A KMO value less than 0.5 indicates that the sample is inadequate and not suitable for analysis, while values closer to 1 suggest higher sampling adequacy. The KMO formula is expressed as follows [11].

$$\text{KMO} = \frac{\sum_{j\neq k} r_{jk}^2}{\sum_{j\neq k} r_{jk}^2 + \sum_{j\neq k} q_{jk}^2} \quad (2)$$

Description:
r_{jk}^2 = Pearson correlation coefficient squared between variable j and k
q_{jk}^2 = Partial correlation coefficient squared between variable j and k
$Q = D.R^{-1}D$
D = $[(diag\ R^{-1})^{\frac{1}{2}}]^{-1}$

2.5. *Multicollinearity*

Multicollinearity refers to the presence of correlations among independent variables within a model. Ideally, there should be no correlation, or if present, the degree of multicollinearity should not be too high so as not to distort the interpretation of the analysis results. The detection of multicollinearity is commonly carried out using Pearson's correlation coefficient (r) between two independent variables, with the following notations:

$$r_{jk} = \frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2 \sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2}} \quad (3)$$

Description:
x_{ij}= observed values of individual i for variables j
x_{ik} = observation value of individual i for variable k
$\bar{x}_j, \bar{x}_k$ = average variables j and k
n = number of observations

2.6. *Principal Component Analysis (PCA)*

Principal components are applied to simplify high-dimensional data by reducing the number of dimensions [11]. The PCA algorithm begins by standardizing the data and calculating the covariance matrix, from which eigenvalues and eigenvectors are derived. The eigenvectors indicate the directions of maximum variance in the dataset, while the eigenvalues represent the amount of variance explained by each corresponding component. The proportion of variance explained by each principal component is calculated by dividing its eigenvalue by the sum of all eigenvalues, providing a measure of the relative importance of each component. By selecting eigenvectors associated with the largest eigenvalues, the principal components are chosen to transform the data into a lower-dimensional space, retaining most of the essential information. This approach facilitates subsequent analyses, such as clustering, by reducing redundancy and highlighting the most informative aspects of the data.

2.7. *Factor Analysis*

Factor Analysis (FA) is a statistical technique used to identify underlying latent variables that explain the correlations among observed variables, thereby reducing data complexity while retaining most of the original information. In this study, FA was applied to extract key dimensions of educational performance across regencies/cities. The process began with data standardization and checking assumptions such as linearity and multicollinearity, followed by factor extraction using Principal Component Analysis. The number of factors retained was determined based on eigenvalues greater than 1 and cumulative variance explained. Rotation (Varimax) was applied to simplify the factor structure, and factor loadings were examined to interpret each latent dimension. Factor scores generated from this process were then used as input for subsequent clustering analysis, facilitating the classification of regions with similar educational characteristics.

2.8. *Elbow Method*

The Elbow method is applied to determine the optimal number of clusters (K) by calculating the Within

Cluster Sum of Squares (WCSS) for each value of c from 1 to k. WCSS represents the total squared distance between data points and their respective cluster centroids. As the number of clusters increases, the WCSS value decreases. When c = 1, WCSS reaches its highest value. The plot typically shows a sharp decline at the beginning, forming an “elbow” shape, after which the curve flattens and runs almost parallel to the X-axis [12]. Algorithm Elbow Method the following formula:
Step 1: Apply a clustering algorithm for different values of c, ranging from 1 to k.
Step 2: For each c, compute the total *Within-Cluster Sum of Squares* (WCSS) using the following formula:

$$\sum_{c=1}^{k}\sum_{i\in S_c}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{cj})^2 \qquad (4)$$

Where S_c represents the set of observations in the C^{th} cluster and $\bar{x}_{cj}$ is the f^{th} variable of the cluster center for the C^{th} cluster.
Step 3: Plot the WCSS values against the number of clusters C
Step 4: The point at which the curve shows a noticeable bend (the “elbow”) is generally considered the optimal number of clusters.

2.9. *K-Medoids*

K-Medoids is one of the methods known as Partitioning Around Medoids (PAM), an extension of K-Means that is sensitive to outliers. This method uses individual objects (medoids) as cluster centres [13]. The stages of clustering using the K-Medoids algorithm can be described as follows [14]:
Step 1: Determine the number of clusters to be formed using the Elbow method.
Step 2: Randomly select initial medoids as many as the predefined number of clusters (k).
Step 3: Calculate the distance of each non-medoid object to the initial medoids of every cluster, and assign each object to the nearest medoid using the following distance measure:

$$d_{i,j} = \sqrt{\sum_{k=1}^{p}(x_{ik}-x_{jk})^2} \qquad (5)$$

Description:
d_{ij} = Euclidean distance between object i and object j
x_{ik}, = Observed value between object i of variable k
x_{jk} = Observed value between object j of variable k
p = Number of Variables
Step 4: Compute the total cost (sum of all distances):

$$Total\ Cost = \sum_{C=1}^{m}\sum_{i=1}^{n} d_{i,C} \qquad (6)$$

Step 5: Calculate the difference in total cost by comparing the new distance with the previous one. If the difference is less than zero, replace the object with the current medoid to form a new set of medoids.
Step 6: Repeat steps 3–5 until no changes occur in the medoid members.

2.10. *Bartlett Test*

The Bartlett test is designed to examine the null hypothesis that the correlation matrix is an identity matrix, implying no correlations among the variables. If the test yields a significant p-value (e.g., p < 0.05), the null hypothesis is rejected, indicating that the variables in the dataset are sufficiently correlated to justify the use of dimensionality reduction techniques such as factor analysis or principal component analysis PCA[15].

$$X^2 = -\left[(N-1) - \frac{(2_p+5)}{6}\right] In|R| \qquad (7)$$

Description:
X^2 = chi-square test statistic
N = sample size
p = number of variables used
|R| = determinant of the correlation matrix

2.11. *Kruskal-Wallis*

The Kruskal-Wallis test, a non-parametric alternative to one-way ANOVA, is employed to determine whether there are significant differences among the medians of three or more groups. This test does not require the data to follow a normal distribution, making it suitable for ordinal data or continuous data that are not normally distributed[16].

2.12. *Davies Bouldin Index (DBI)*

The Davies-Bouldin Index (DBI) is a validity measure commonly employed to evaluate both the number of clusters formed and the overall quality of the clustering results. This index provides an assessment by examining the ratio between within-cluster similarity and between-cluster separation. A lower DBI value indicates that the clusters are more compact and well-separated from one another, which reflects a better clustering structure. Therefore, the smaller the DBI, the more optimal the clustering solution is considered to be. The mathematical formulation of DBI as proposed by [17] is as follows.

$$\text{DB} = \frac{1}{k}\sum_{c=1}^{k} R_c \tag{8}$$

Description:
DB = davies bouldin index
R_p = cluster similarity measure (maximum)

2.13. Analysis techniques

This methodology describes the essential steps in applying the proposed approach to determine the most appropriate target data clusters. The overall process supports the discovery of meaningful patterns in the data and guides more effective decisions regarding the allocation of aid (see Figure 2).

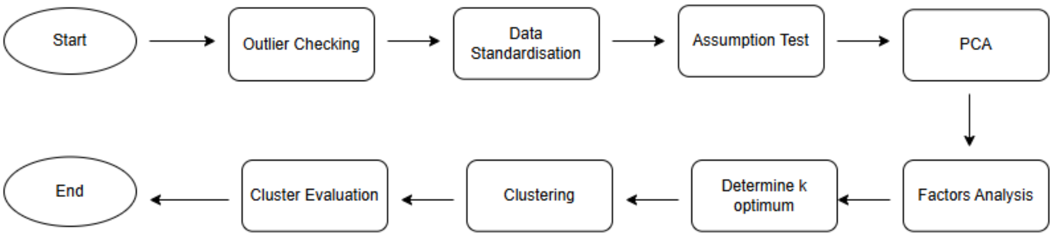

Figure 2. Research flow chart.

This research was carried out through several systematically structured analytical stages, as illustrated in the research flow diagram. The initial steps, such as outlier checking and data standardization, are conducted to ensure data quality and minimize potential bias in subsequent statistical procedures. Once the data meet the required assumptions, *Principal Component Analysis (PCA)* is applied to reduce data dimensionality and identify the most influential variables. The extracted principal components then serve as the foundation for *Factor Analysis*, which aims to uncover latent dimensions underlying the observed variables. The resulting factor scores are used as input for the clustering process using the *K-Medoids* algorithm, as standardized and reduced data tend to produce more stable and interpretable cluster structures. The determination of the optimal number of clusters and subsequent cluster evaluation are performed to ensure that the resulting groups accurately represent distinct regional characteristics. Thus, each analytical stage supports one another, forming a coherent workflow that effectively illustrates the spatial patterns and disparities in educational performance across regions in Indonesia.

3. Result and Discussion

This section presents the results of data analysis conducted based on predetermined research variables. The analysis begins with data exploration to determine the distribution characteristics of each variable. This stage is important to ensure data validity before proceeding to further analysis.

3.1. Outlier Checking

Prior to the main analysis, an outlier detection procedure was conducted to identify any extreme values within the dataset. This step is crucial to ensure the validity and reliability of the subsequent factor and clustering analyses, as outliers can distort statistical relationships and influence clustering results. Outlier detection was performed using a boxplot visualization to identify data points lying beyond acceptable boundaries, as illustrated in the figure below.

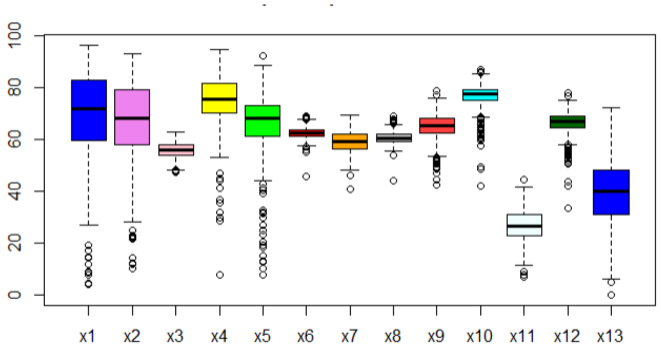

Figure 3. Boxplot for each variable.

Based on the image above, it shows that each variable used indicates the presence of outliers. This indicates that there are disparities in quality, access, and governance in Indonesia. Therefore, the use of K-Medoids is appropriate because this method is used on data containing outliers.

3.2. Data Standardisation

Prior to performing cluster analysis, the data must be standardised using the z-score method to ensure that all variables are measured on the same scale. Standardisation produces data with a uniform scale and distribution, thereby facilitating analysis and comparison across variables. The standardisation output highlights variations in z-score values among regencies/cities based on the dimensions of quality, access, and educational governance.

3.3. Assumption Test

3.3.1 KMO and Barlett's Test

After confirming that the dataset was free from outliers, an assumption test was carried out to evaluate the suitability of the data for factor analysis. The Kaiser-Meyer-Olkin (KMO) measure and Bartlett's Test of Sphericity were applied to assess sampling adequacy and inter-variable correlations. A high KMO value indicates sufficient correlation among variables for factor extraction, while a significant Bartlett's Test result confirms that the correlation matrix differs significantly from the identity matrix, thereby validating the data's appropriateness for factor analysis.

Table 3. Result of the KMO Test.

X1	**X2**	**X3**	**X4**	**X5**	**X6**	**X7**	**X8**	**X9**	**X10**	**X11**	**X12**	**X13**
0.87	0.88	0.92	0.91	0.93	0.80	0.93	0.74	0.91	0.89	0.89	0.90	0.96

The result of the KMO test indicate have KMO values above 0.5. This suggests that the sample used is adequate and representative of the population. Therefore, the assumption of sample representativeness is fulfilled. The next step is to conduct a correlation test to examine whether there are relationships among the variables that may indicate multicollinearity.

Table 4. Result Bartlett's Test.

chisq	**p-value**
126.6408	0.000414

The results of Bartlett's test revealed a p-value of less than 0.005, indicating significant differences in variances across groups. This confirms that applying PCA is appropriate for dimensionality reduction, providing a suitable basis for subsequent clustering analysis.

3.3.2 Correlation

Subsequently, a correlation matrix analysis was conducted to examine the strength of relationships among the variables. High correlations between several variables suggest the presence of underlying latent dimensions within the dataset. These findings provide the foundation for exploratory factor analysis, enabling the identification of interrelated patterns among educational indicators.

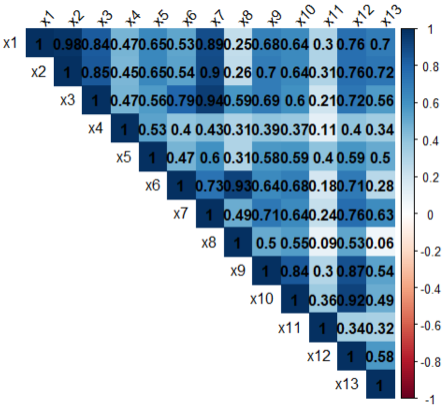

Figure 4. Correlation Test.

The correlation analysis reveals several variable pairs with very high correlations (r > 0.80), including x1–x2 (0.98), x1–x7 (0.89), x2–x7 (0.90), x3–x7 (0.94), and x9–x12 (0.87). Correlation values approaching 1 indicate a very strong relationship between these variables. This suggests the presence of potential multicollinearity, where independent variables share overlapping information. In the context of multivariate analysis such as PCA or clustering, multicollinearity should be carefully considered, as it may influence the interpretability of results. Therefore, the application of Principal Component Analysis (PCA) becomes relevant to reduce data dimensionality while addressing redundancy among variables.

3.4. PCA

Principal Component Analysis (PCA) was conducted to reduce the dimensionality of the 13 educational indicators and to identify the most influential components explaining the variance in the data. The PCA results revealed that three principal components have eigenvalues greater than one, which together explain a significant proportion of the total variance. These components were then subjected to Factor Analysis (FA) to extract the underlying latent dimensions that represent broader educational constructs. The factor loading matrix, as presented in Table 5, indicates strong correlations between several variables and their respective factors.

Table 4. PCA Summary.

Component	**Eigen Value**	**Proportion of Variance**	**Cumulative Variance (%)**
PC1	7.9015	0.6078	60.10
PC2	1.5455	0.1189	72.67
PC3	1.0055	0.0773	80.40
PC4	0.8253	0.0634	86.75
PC5	0.6272	0.0482	91.57
PC6	0.3812	0.0293	94.51
PC7	0.3259	0.0251	97.10

PC8	0.1802	0.0138	98.40
PC9	0.0696	0.0053	98.94
PC10	0.0554	0.0042	99.36
PC11	0.0434	0.0033	99.70
PC12	0.0201	0.0015	99.85
PC13	0.0185	0.0014	100.00

Based on the PCA summary presented in Table above, the eigenvalues of each component were calculated to determine the appropriate number of retained components. The selection criterion applied was eigenvalue ≥ 1. The results indicate that three principal components (PC1, PC2, and PC3) meet this criterion. Collectively, these three components explain 80.4% of the total variance. Therefore, the use of three principal components is considered sufficient to represent the information contained in all analyzed variables.

3.5. Factors Analysis

Based on the PCA results, which identified three principal components explaining 80.4% of the total variance, a subsequent Factor Analysis (FA) was conducted to interpret the underlying structure of these components in more detail. The objective of FA is to identify the latent constructs that account for the observed correlations among the educational indicators. By examining the factor loading matrix, as presented in the table below, it becomes possible to determine which variables have the strongest association with each extracted factor. This approach allows for a more meaningful interpretation of the educational dimensions that characterize regional differences.

Table 5. Loading Factors.

Variable	PA1	PA2	PA3
X1	0.876	0.142	0.398
X2	0.888	0.143	0.403
X3	0.787	0.534	0.197
X4	0.390	0.232	0.226
X5	0.501	0.192	0.470
X6	0.346	0.895	0.245
X7	0.818	0.422	0.281
X8	0.000	0.959	0.156
X9	0.423	0.393	0.664
X10	0.269	0.431	0.828
X11	0.182	0.000	0.391
X12	0.449	0.415	0.729
X13	0.629	0.000	0.433

Based on the factor loading results presented in Table 5, three main factors were identified, each representing distinct latent dimensions of educational performance across regions in Indonesia. Factor 1 (PA1) shows high loadings on key variables such as literacy rate (0.876), numeracy rate (0.888), school participation rate (0.787), and average years of schooling (0.818). These indicators are strongly associated with students' academic achievement and learning outcomes. Therefore, this factor represents Educational Quality, reflecting regions with strong educational performance characterized by high literacy and numeracy levels, as well as broad school participation.

Factor 2 (PA2) is dominated by teacher-to-student ratio (0.895) and teacher qualification index (0.959). These variables emphasize aspects of instructional efficiency and teaching quality. Accordingly, this factor is interpreted as Quality of the Learning Process, indicating the effectiveness of classroom practices, teacher competence, and the overall quality of instructional interaction. Regions with high scores in this factor tend to exhibit well-managed and efficient learning systems supported by qualified educators.

Factor 3 (PA3) presents high loadings for school management index (0.664), educational participation rate (0.828), and community involvement in education (0.729). This factor captures Educational Governance and Participation, focusing on managerial and social aspects of education. It highlights the implementation of educational policies, institutional management practices, and the degree of community engagement in supporting education sustainability.

Overall, the three extracted factors—Educational Quality, Quality of the Learning Process, and Educational Governance and Participation—explain 80.4% of the total data variance. This indicates that these three latent dimensions comprehensively describe the structural characteristics of the educational system across regions in Indonesia. The results provide deeper insights into how disparities in educational quality can be understood through a combination of learning outcomes, teaching effectiveness, and governance practices supported by community participation.

3.6. Determine k cluster

To ensure that the clustering process produces meaningful and well-separated groups, it is essential to determine the optimal number of clusters (*k*) before performing the K-Medoids analysis. In this study, the Elbow Method was employed using the factoextra package and the fviz_nbclust function in RStudio. The method evaluates the total within-cluster sum of squares (WSS) for different cluster numbers and identifies the point where the rate of decrease in WSS begins to level off, indicating the most suitable number of clusters.

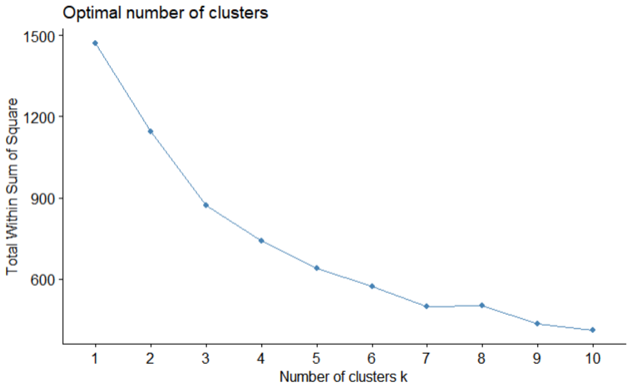

Figure 5. Elbow Plot of K-Medoids.

As shown in the resulting plot, the decline in WSS became noticeably less significant after the third cluster. This pattern suggests that increasing the number of clusters beyond three does not substantially improve clustering performance. Therefore, the optimal number of clusters (k = 3) was selected for the subsequent K-Medoids clustering analysis, ensuring an effective balance between model simplicity and explanatory power.

3.7. K-Medoids

Based on the determination of the optimal k value, four clusters were identified as the optimal solution. Subsequently, cluster analysis was conducted using the K-Medoids method with the pam function from the cluster package. The clustering results are summarized in the Table 6.

Table 6. Average of each cluster.

Cluster	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13
1	64.50	62.30	54.25	72.93	64.11	61.16	57.03	59.41	64.12	76.34	27.58	65.61	38.23
2	54.44	53.39	55.37	73.39	59.77	63.78	57.50	62.96	63.46	75.34	24.27	64.65	28.20
3	86.52	82.76	58.48	79.33	72.77	63.51	63.40	61.09	67.23	77.86	27.28	68.44	49.04

Based on the factor loading results presented in Table 5, three main factors were identified that represent the latent dimensions within the educational data, namely: (1) education quality, (2) quality of the learning process, and (3) governance and educational participation. These three factors served as the basis for constructing factor scores, which were subsequently used in the clustering process employing the K-Medoids method.

Table 6 presents the average values of each variable across clusters. Cluster 3 stands out with the highest mean values for almost all quality-related indicators, such as literacy (86.52), numeracy (82.76), and student participation (77.86). This indicates that regions belonging to Cluster 3 have high educational quality and active participation from both students and schools. This cluster reflects areas with a well-established learning system and strong support for educational excellence. Conversely, Cluster 1 shows moderate values for most variables, with relatively strong performance in school participation indicators (APS = 72.93) and net enrollment rate (APM = 64.11). This suggests that regions in Cluster 1 have relatively good access to education, although the quality of learning still requires improvement. Meanwhile, Cluster 2 records the lowest averages for almost all indicators, including literacy (54.44) and numeracy (53.39), depicting areas with low educational quality, possibly due to limited resources, teacher competency, and local education policy support.

After performing the K-Medoids clustering with the optimal number of clusters (k = 3), the mean values of each latent dimension (PA1, PA2, and PA3) were calculated for every cluster to examine the distinct characteristics of each group. To statistically verify whether these differences among clusters were significant, the Kruskal–Wallis test was subsequently conducted. This non-parametric test is suitable for assessing variations in median values across multiple independent groups when the data do not meet the assumption of normality.

Table 7. Result Kruskal-wallis test.

Dimension	Chi-squared	p-value
PA1	339.92	$< 2.2e^{-16}$
PA2	226.79	$< 2.2e^{-16}$
PA3	85.45	$< 2.2e^{-16}$

Based on the Kruskal–Wallis test results, all three dimensions (PA1, PA2, and PA3) show high chi-squared values with p-values < 0.005, indicating that these dimensions differ significantly across clusters. This finding supports the validity of the K-Medoids clustering results, suggesting that each cluster represents distinct regional profiles in terms of the three underlying latent dimensions—PA1 (Educational Quality), PA2 (Learning Process Quality), and PA3 (Governance and Educational Participation).

3.8. Cluster Evaluation

Following the cluster analysis, the next step was to evaluate the clustering results using the Davies-Bouldin Index (DBI) with the help of RStudio, specifically employing the Index.DB function from the ClusterSim package. The calculation yielded a DBI value of 1.44 for the K-Medoids analysis with three clusters. According to [11], a lower DBI value indicates better clustering quality, with values closer to 0 reflecting optimal cluster separation. Therefore, a DBI score of 1.44 can be considered good, as it adequately represents the heterogeneity among regions within each cluster, although there remains room for improvement in cluster separation.

3.9. Discussion

Overall, the analysis indicates significant disparities across regions, highlighting the need for policy interventions tailored to the characteristics of each cluster. The clustering results presented in Table 5 are further visualised in Figure 6 to facilitate interpretation of the grouping of regencies/cities in Indonesia based on the dimensions of quality, access, and governance in education for the year 2025 with the help of Rstudio, specifically employing the shapefile JSON and leaflet package.

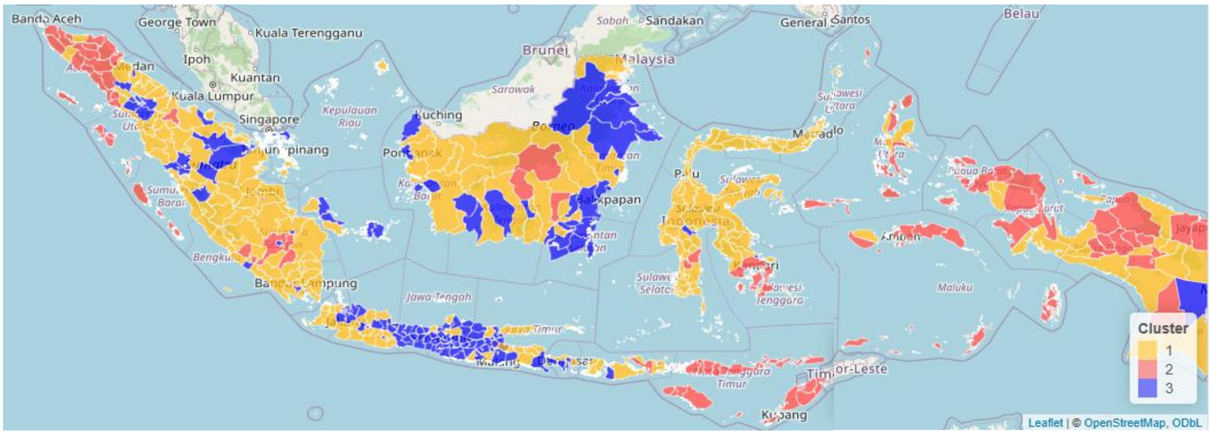

Figure 6. Map of the Distribution of Regencies/Cities in Indonesia Based on Cluster Results.

The clustering analysis of regencies/cities in Indonesia using the K-Medoids method identified three clusters based on indicators of education quality, learning processes, and governance. Each color on the map represents a different cluster, reflecting variations in educational characteristics across regions. Cluster 3 (blue) includes regions with strong performance in literacy, numeracy, and student participation, supported by well-established learning systems, adequate infrastructure, and active community engagement. These regions are mainly concentrated in parts of Java, Kalimantan, and Sulawesi, highlighting areas that have successfully balanced education quality and governance. Interestingly, some 3T areas (frontier, outermost, and disadvantaged regions) are included in this cluster, indicating that appropriate policies can foster positive educational outcomes even in challenging contexts. Cluster 1 (orange) comprises regions with mixed performance, showing relatively good access to education, such as school participation, yet still facing challenges in achieving consistent learning outcomes. These areas are widely distributed across Sumatra, Kalimantan, Sulawesi, as well as parts of Java and Papua, reflecting ongoing improvements in educational opportunities alongside the need for further support to enhance teaching quality and governance. Cluster 2 (red) consists of regions with more limited educational outcomes, with constraints in literacy, numeracy, and overall learning effectiveness. These regions are mostly located in remote and less developed areas, including parts of Sumatra, Kalimantan, Papua, and other 3T regions. This highlights persistent disparities and the urgent need for targeted interventions to improve educational services and equity.

Overall, the findings highlight significant regional disparities in education across Indonesia. Geographic remoteness, socio-economic conditions, and governance effectiveness play crucial roles in shaping cluster membership, with urban areas generally performing better than remote regions. Based on these insights, several policy recommendations can be formulated to guide decision-makers in developing targeted operational strategies. A concise summary is presented in the table below.

Table 8. Policy Recommendations.

Cluster	Number of Regions	Medoid Regions	Main Targets	Programs	Outcome Indicators
3	200	Banjarmasin City, South Kalimantan Province	Maintain high quality in literacy, numeracy, character development, school participation rate (APS, ages 16–18), and net participation rate (APM, ages 16–18); Improve	Best practice sharing programs, school digitalization, incentives for honorary teachers, targeted scholarship programs (PIP)	Stable or increasing literacy & numeracy scores, APS & APM (16–18) >95%, high proportion of certified teachers maintained, improved utilization of local government education budget (APBD)

			quality of learning, inclusive school climate, proportion of local government budget utilization for education, and percentage of certified senior high school teachers.		
1	399	Maros Regency, South Sulawesi Province	Improve quality (literacy, numeracy, character, quality of learning, inclusive climate), access (school participation rate [APS, ages 16–18], net participation rate [APM, ages 16–18]), and governance (proportion of local government budget utilization for education)	Targeted scholarship programs (PIP), conditional cash transfers for vulnerable families, school management mentoring, certified teacher training and incentives for honorary teachers, additional DAK Fisik for facilities and infrastructure (educational transportation, classroom renovation, sanitation)	APS and APM increased by at least 5%, significant increase in certified teachers, improved literacy and numeracy scores
2	505	North Nias Regency, North Sumatra Province	Improve quality (literacy, numeracy, character, quality of learning, inclusive climate), access (school participation rate [APS, ages 16–18], net participation rate [APM, ages 16–18]), and governance (proportion of local government	Additional DAK Fisik for school infrastructure, strengthening certified teacher training and incentives for honorary teachers, targeted scholarship programs (PIP), school management mentoring.	Increase literacy & numeracy by 5%, APS and APM up 3–5% within 12 months, certified teachers increased by 10%

		budget utilization for education)

4. Conclusion

This study provides a comprehensive understanding of educational disparities across Indonesia by systematically analyzing regional data through PCA, Factor Analysis, and K-Medoids clustering. The identification of three latent dimensions—education quality, quality of the learning process, and governance and educational participation—offers a robust framework for interpreting regional variations in educational performance. The optimal clustering solution, validated with a Davies–Bouldin Index (DBI) of 1.44, confirms that the clusters effectively capture differences among regions. The results demonstrate that spatial patterns in education are closely linked to local governance, infrastructure, and access, highlighting areas where targeted interventions are essential. This research underscores the necessity of cluster-based, evidence-driven policy strategies to enhance education quality, ensure equitable access, and strengthen governance mechanisms nationwide. By providing a methodological framework that integrates dimensionality reduction and clustering, this study contributes both theoretically and practically to the design of educational policies aimed at reducing disparities and promoting inclusive development across diverse regions.

Acknowledgement

The author would like to express his deepest gratitude and appreciation to the Data and Information Technology Centre (Pusdatin), Ministry of Primary and Secondary Education, for the data support and data science training provided. This support has been invaluable in conducting the research and compiling this paper.

References

[1] Kementerian Pendidikan Dasar dan Menengah, “Rapor Pendidikan Kabupaten/Kota di Indonesia Tahun 2025,” 2025.

[2] M. S. Anwar, “Ketimpangan aksesibilitas pendidikan dalam perpsektif pendidikan multikultural,” *Foundasia*, vol. 13, no. 1, pp. 1–15, 2022, doi: 10.21831/foundasia.v13i1.47444.

[3] R. Fadillah, R. Desmaryani, and A. Lestari, “Analisis Ketimpangan Sarana Dan Prasarana Pendidikan Di Daerah Pedesaan,” *J. Adijaya Multidisplin*, vol. 03, no. 02, pp. 217–225, 2025, [Online]. Available: https://e-journal.naureendigition.com/index.php/mj

[4] L. L. Figueroa, S. Lim, and J. Lee, “Spatial analysis to identify disparities in Philippine public school facilities,” *Reg. Stud. Reg. Sci.*, vol. 3, no. 1, pp. 1–27, 2016, doi: 10.1080/21681376.2015.1099465.

[5] C. E. Widiantoro, S. P. Putri, T. Purwandari, and U. Padjadjaran, “Pengelompokkan Provinsi di Indonesia Berdasarkan Indikator Pendidikan Jenjang SMA / Sederajat dengan Analisis Klaster Non Hierarki,” vol. 3, pp. 1–9, 2024.

[6] A. A. Mattjik and I. M. Sumertajaya, *Sidik Peubah Ganda dengan Menggunakan SAS*. 2011.

[7] E. Luthfi and A. W. Wijayanto, “Analisis Perbandingan Metode Hirearchical, K-Means, dan K-Medoids Clustering Dalam Pengelompokkan Indeks Pembangunan Manusia Indonesia Comparative Analysis of Hirearchical, K-Means, and K-Medoids Clustering and Methods in Grouping Indonesia’s Human,” *Inovasi*, vol. 17, no. 4, pp. 761–773, 2021.

[8] Kemdikbud, “Rapor Pendidikan Indonesia Tahun 2023,” *Merdeka Belajar*, p. 2023, 2023, [Online]. Available: https://raporpendidikan.kemdikbud.go.id/login

[9] Safiq Maulido, Popi Karmijah, and Vinanda Rahmi, “Upaya Meningkatkan Pendidikan Masyarakat Di Daerah Terpencil,” *J. Sade. Publ. Ilmu Pendidikan, pembelajaran dan Ilmu Sos.*, vol. 2, no. 1, pp. 198–208, 2023, doi: 10.61132/sadewa.v2i1.488.

[10] A. Gere, “Recommendations for validating hierarchical clustering in consumer sensory projects,” *Curr. Res. Food Sci.*, vol. 6, no. February, p. 100522, 2023, doi: 10.1016/j.crfs.2023.100522.

[11] A. C. Rencher and W. F. Christensen, “Méthods of multivariate analysis. a john wiley & sons,” *Inc. Publ.*, vol. 727, pp. 230–2218, 2002.

[12] S. A. P. Raj and Vidyaathulasiraman, “Determining Optimal Number of K for e-Learning Groups Clustered using K-Medoid,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, pp. 400–407, 2021, doi: 10.14569/IJACSA.2021.0120644.

[13] A. Hoerunnisa, G. Dwilestari, F. Dikananda, H. Sunana, and D. Pratama, “Komparasi Algoritma K-Means Dan K-Medoids Dalam Analisis Pengelompokan Daerah Rawan Kriminalitas Di Indonesia,” *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 1, pp. 103–110, 2024, doi: 10.36040/jati.v8i1.8249.

[14] Riska, *Analisis klaster k-means dan k-medoids dalam pengelompokan provinsi di indonesia berdasarkan rumah tangga usaha pertanian subsektor st2023 skripsi*. 2024.

[15] F. Asabuwa Ngwabebhoh *et al.*, “Preparation and Characterization of Nonwoven Fibrous Biocomposites for Footwear Components.,” *Polymers (Basel).*, vol. 12, no. 12, Dec. 2020, doi: 10.3390/polym12123016.

[16] A. C. MACUNLUOGLU and G. OCAKOĞLU, “Comparison of the performances of non-parametric k-sample test procedures as an alternative to one-way analysis of variance,” *Eur. Res. J.*, vol. 9, no. 4, pp. 687–696, 2023, doi: 10.18621/eurj.1037546.

[17] G. R. Suraya and A. W. Wijayanto, “Comparison of Hierarchical Clustering, K-Means, K-Medoids, and Fuzzy C-Means Methods in Grouping Provinces in Indonesia according to the Special Index for Handling Stunting,” *Indones. J. Stat. Its Appl.*, vol. 6, no. 2, pp. 180–201, 2022, doi: 10.29244/ijsa.v6i2p180-201.