Contents lists available at GrowingScience

International Journal of Data and Network Science

homepage: www.GrowingScience.com/ijds

Unraveling educational networks: Data-driven exploration through multivariate regression, geographical clustering, and multidimensional scaling

Restu Arisanti[a*], Yuyun Hidayat[a], Irlandia Ginanjar[a], Titi Purwandari[a], Arum Putri Juniarsih[a] and Janatin[a]

[a] *Department of Statistics, Padjadjaran University, Indonesia*

CHRONICLE

Article history:
Received: November 6, 2023
Received in revised format: November 24, 2023
Accepted: December 22, 2023
Available online: December 22, 2023

Keywords:
School Participation
Education
Multivariate Regression
Geographical Classification
Multidimensional Scaling

ABSTRACT

Enhancing rates of school participation holds significant importance for a nation's educational achievements. This research employs a comprehensive approach that combines various methodologies, including multivariate regression analysis, geographic categorization, and multidimentional visualization, to examine the factors influencing school enrollment in Indonesia. Through the integration of diverse data sources, we investigate the connections among variables such as economic status, school accessibility, educational quality, and societal considerations concerning enrollment rates. This discrete impact of each factor on enrollment variations is analyzed through multivariate regression. Geospatial clustering analysis reveals enrollment trends in different regions, while multidimensional visualization untangles the intricate interplay of influencing factors. This holistic approach facilitates a nuanced comprehension of these dynamics within Indonesia's varied geographical and society offering guidance in the formulation of more efficient strategies to improve school attendance, tackle enrollment disparities, and advocate for inclusive education based on fundamental determinants.

1. Introduction

Education is a fundamental human right and a driving force in societal and economic progress. It is required to meet the Sustainable Development Goals 4 (*SD*Gs 4), a set of worldwide goals aimed at providing universal access to high-quality education. Indonesia has made considerable progress in improving educational opportunities as a country committed to accomplishing these goals. Attendance rates vary at several levels, including *SD*, *SMP*, *SMA* and *SMK*. Understanding the factors driving these rates is critical for developing policies and activities that will successfully improve educational results.

A crucial indicator of a nation's educational accessibility and quality is the school participation rate. Despite the National Education System Law No. 20 of 2003's emphasis on the value of equal and outstanding education, a number of barriers still prevent many Indonesian students from enrolling in school. A major issue is the unequal distribution of educational access among regions. Even though the law places a strong emphasis on the idea of educational equality, access to schools can occasionally be hindered, especially in distant areas, by insufficient infrastructure and economic inequality across locations.

Conversely, the caliber of education bears notable influence over school enrollment metrics. Legislative provisions in Law No. 14 of 2005 concerning Teachers and Lecturers outline benchmarks for educators to uphold educational excellence. Nevertheless, certain regions grapple with an insufficiency of adept instructors and requisite provisions, potentially dampening students' fervor and motivation for academic engagement. This, in turn, directly reverberates on the rates of school participation, with diminished educational attainment potentially prompting locals to opt for less frequent enrollment of their progeny.

[*] Corresponding author.
E-mail address: r.arisanti@unpad.ac.id (R. Arisanti)

ISSN 2561-8156 (Online) - ISSN 2561-8148 (Print)

doi: 10.5267/j.ijdns.2023.12.020

On the contrary, the impact of education quality on school enrollment rates is significant. Professional standards for educators are delineated in Law No. 14 of 2005 on Teachers and Lecturers to ensure the provision of quality education. However, the inadequacy of proficient educators and necessary resources in certain areas can undermine students' eagerness and drive to attend school. This directly impacts school enrollment rates, as lower education levels can lead residents to send their children to school with reduced frequency.

The purpose of this research is to look into the factors that influence Indonesian students' participations rates at various educational levels. We hope to provide a full knowledge of the numerous mechanisms that lead to variable participation rates by using an integrated technique that incorporates multivariate regression, spatial clustering, and multidimensional scaling. The purpose of this research is to discover the elements that have a substantial impact on school attendance rates in order to shed light on their current situation.

Using multivariate regression to analyze the relationship between various factors and school attendance rates at different educational levels. It explores the individual and combined effects of these factors, providing insights into participation rates. Geographic clustering analysis helps identify spatial patterns and regional variations in school attendance rates, aiding in targeted interventions and resource allocation. Multidimensional scaling analysis is utilized to understand relationships between variables and participation rates, highlighting key factors influencing attendance. Understanding these factors is crucial for evidence-based policies to enhance participation rates, ultimately promoting sustainable development, social mobility, poverty reduction, and achieving SDG 4, breaking the cycle of poverty through education.

The findings of this study will aid in allocating resources to the places most in need and will encourage evidence-based decisions. We can build targeted and effective measures to increase school enrollment rates across the country by taking a holistic approach that blends quantitative research with geographical awareness.

2. Materials and Methods

2.1 Multidimensional Scaling

Multidimensional Scaling (MDS) analysis is a multivariable technique that can be used to determine the similarity between any pair of N observed elements and to plot elements in multiple dimensions based on the proximity between elements and their similarity elements (Johnson & Winchern, 2007). When the distance value is *SMA*ller, similarity means the object is more similar, while dissimilarity itself means that the object becomes progressively more dissimilar as the distance value is larger (Rabinowitz, 1975).

This analysis is used to determine the relationship of interdependence or interdependence between variables or data. This visual perception map is executed in a multidimensional map (Adlakha & Sharma, 2019). Based on the scale of the data used, multidimensional scaling analysis is divided into multidimensional metric and multidimensional nonmetric analysis (Johnson & Winchern, 2007).

Multidimensional scaling not metric

The distance data used in this scaling is ordinal scaled data. Rabinowitz (1975) describes several analysis phases when performing a multidimensional scaling analysis, including (Rabinowitz, 1975):

1. Calculation of the distance matrix using the Euclidean distance value. Euclidean distance is used to calculate the inter-object proximity between the first object and the j-th object perception map with the following formula.

$$d_{\text{ij}} = \sqrt{\sum_{h=i}^{n} \left(x_{ih} - x_{jh}\right)^2}, \tag{1}$$

2. Find the eigenvalue and eigenvector using the following formula

$$det\ (B - \lambda I) \text{ And } det\ (B - \lambda I)X, \tag{2}$$

3. Forming object coordinates based on eigenvectors $\boldsymbol{X} = [X_1, X_2]$, then the next calculation $\widehat{D}$. That's Euclidean distance is formed by coordinates.
4. Calculate the voltage value using the following formula.

$$S = \left(\frac{\sum_{i=j}^{n} \left(d_{ij} - \hat{d}_{ij} \right)^2}{\sum_{i=j}^{n} d_{ij}^2} \right), \tag{3}$$

From the stress value, it can be seen that the lower the stress value, the better the resulting model. The following are guidelines for criteria that can be used to assess the feasibility of models using stress values (Johnson and Winchern, 2007).

Table 1
Criteria for the value of the emphasis on the feasibility of the model.

Stress values	Goodness of fit
20 %	Poor
10 %	Fair
5 %	Good
2,5 %	Excellent
0%	Perfect

Source: Johnson and Wichern (2007)

2.2 Panel-Regression

The regression analysis of the panel data is the result of observing several people, each observed in several consecutive periods (time units) (Bai and Kunpeng, 2014).

Estimation Model of Panel Data Regression Analysis

Common Effects Model

Model that there is no difference in the intercept and slope values in the regression results, either due to differences between individuals or between times. In general, the equation of the common effects model is as follows (Baltagi, 2008).

$$Y_{it} = \beta_0 + \sum_{k=1}^{k} \beta_k X_{kit} + u_{it}, \tag{4}$$

for $i = 1,2, \dots, N; t = 1,2, \dots, T; k = 1,2, \dots, K$

Fixed Effect Model

Estimation method regression Panel data on the fixed effect model use the technique of adding a dummy variable or Least Square Dummy Variable (LSDV). There are two assumptions in the Fixed Effect Model namely as follows (Hsiao, 2003).

Step 0. The slope value is constant, but the intercept varies between units;

$$Y_{it} = \beta_{0i} + \sum_{k=1}^{k} \beta_k X_{kit} + u_{it}, \tag{5}$$

Step 1. The slope value is constant, but the intercept varies between individuals and between periods.

$$Y_{it} = \beta_{0it} + \sum_{k=1}^{k} \beta_k X_{kit} + u_{it}, \tag{6}$$

Random Effect Model

The panel data regression estimation in the random effects model employs the Generalized Least Squares (GLS) method. There are two key assumptions regarding the random effect (Hsiao, 2003).

Step 2. The intercept and slope vary from person to person;

$$Y_{it} = \beta_{0i} + \sum_{k=1}^{k} \beta_{ki} X_{kit} + u_{it}, \tag{7}$$

Step 3. The intercept and slope differ between individuals and over time.

$$Y_{it} = \beta_{0it} + \sum_{k=1}^{k} \beta_{kit} X_{kit} + u_{it}, \tag{8}$$

Selection of Panel Data Regression Model

Chow Test

The Chow test is a test performed to select one of the models in panel data regression, namely between the Fixed Effect Model and with Common Effect Model (Ioan et al., (2020). This test was conducted with the following hypothesis (Binkley et al., 2018).
$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_N = \alpha$ (Common Effect Model)
H_1 : at least there is oneα_I otherwise (Fixed Effect Model)

The basis for rejection is determined based on the F-statistic tests as follows (Baltagi, 2008).

$$Chow = \frac{RSS_1 - RSS_2/(N-1)}{RSS_2/(NT-N-K)}, \tag{9}$$

The test statistics for the Chow test follow the distribution of the F-statistics, which is $F_{(N-1,NT-N-K);\alpha}$ With the test criteria whether the statistical Chow value is greater than the F-table or if the $p - value < \alpha$, then H_0 is rejected and vice versa.

Hausman Test

This test is performed based on Fixed Effect Model contains an item trade off namely the loss of degrees of freedom due to the inclusion of dummy variables and Random Effect Model Care must be taken to ensure that there are no violations of the assumptions of each error component (Binkley et al., 2018). The hypothesis used is:

$H_0 : corr(X_{it}, U_{it}) = 0$ (Random Effect Model)
$H_1 : corr(X_{it}, U_{it}) \neq 0$ (Fixed Effect Model)

The basis of rejection H_0 The value derived from the Hausman statistic is formulated as follows (Greene, 2000).

$$\chi^2(K) = (b - \beta)'[Var(b - \beta)^{-1}(b - \beta)], \tag{10}$$

The test criteria for this test follow the chi-square distribution, which is if the value χ^2 greater than the value $\chi^2_{(K,\alpha)}$ or if the $p - value < \alpha$, then H_0 is rejected and vice versa.

Lagrange Multiplier Test

The Lagrange multiplier test (LM) is a test to find out whether the random effects model is better than the common effects model (Breusch & Pagan, 1980)

H_0 : The correct model is the common effects model
H_1 : The appropriate model is the random effects model.

The test statistics for the Lagrange multiplier test are as follows (Baltagi et al., 2012).

$$LM = \frac{KT}{2(T-1)} \left[\frac{\sum_{i=1}^{K} [\sum_{t=1}^{T} e_{it}]^2}{\sum_{i=1}^{K} \sum_{t=1}^{T} e_{it}^2} - 1 \right]^2 \sim \chi^2_{\alpha,1}, \tag{11}$$

where K is the number of sectors, T is the number of periods and e_{it} is the residual Common Effect Model. The test criteria used are: if the value $LM > \chi^2{}_{(\alpha,1)}$ or if the $p - value < \alpha$, then H_0 is rejected and vice versa.

Breusch Pagan Test
The Breusch-Pagan test was performed to see if there were single, temporal, or both effects on the fixed effects model and random Effect.

Model Selection

An optimal regression model yields unbiased linear estimates, known as the Best Linear Unbiased Estimator (BLUE). Meeting classical assumptions is crucial for this, especially in the context of combined cross-sectional and time-series data. Overcoming issues related to these assumptions, such as heteroscedasticity and autocorrelation, is vital to ensure the model is analyzable and delivers accurate results.

2.3 Cluster Analysis

Cluster analysis is a multivariate data analysis used to group objects/cases based on the similarity of the objects/cases' characteristics (Johnson & Winchern, 2007). K-Means is a non-hierarchical cluster analysis algorithm in which the clustering process is performed based on the nearest distance to the specified center (Bhattacharjee et al., 2017). One of the commonly used distances is the Euclidean distance. The formula for Euclidean distance is as follows.

$$d(x_i, x_j) = \sqrt{\left(\left|x_{i1} - x_{j1}\right|^2 + \left|x_{i2} - x_{j2}\right|^2 + \cdots + \left|x_{ip} - x_{jp}\right|^2\right)}, \tag{12}$$

where X_I, X_J are the two data calculated using the distance and p is the dimension of the data used. The determination of the cluster center can be seen from the following equation.

$$C_{m(q)} = \frac{1}{n_m} \sum_{i=1}^{n_m} x_{i(q)}, \tag{13}$$

The results of grouping each distance calculation can be checked for quality by performing a homogeneity test. This test is calculated using the Silhouette coefficient equation to be performed after convergence reaches 0, with the results of the last binning being identical to those of the previous binning (Erda et al., 2023).

The silhouette coefficient is determined by averaging the distance of the i-th data to all data in the same cluster. Here we assume that the i-th data is in cluster A. The formula of $a(i)$. Written in the following equation (Struyf et al., 1997).

$$a(i) = \frac{1}{|A| - 1} \sum_{j \in A, j \neq i} d(i, j), \tag{14}$$

where A is the amount of data in cluster A. Next, calculate the value $b(i)$, this is the minimum value of the ith data center distance with all data in different clusters. Now suppose that clusters other than A originate from cluster C. So, the calculation of the average distance between the it h data and all data in cluster C is as follows:

$$d(i, C) = \frac{1}{|C|} \sum_{j \in C} d(i, j), \tag{15}$$

After counting $d(i, C)$ for all clusters, $C \neq A$, then select the minimum distance value as the value $b(i)$.

$$b(i) = \min_{C \neq A} d(i, C), \tag{16}$$

If cluster B has a minimum distance value, then $d(i, B) = b(i)$ This is called the neighbor of the i-th data and is the second-best cluster for the i-th data after cluster A. After $a(i)$ and $b(i)$ is known, the final process of computing the silhouette coefficient is as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i) - b(i)\}}, \tag{17}$$

Table 2
The Silhouette Coefficient score criteria

Silhouette Coefficient	Suggested Interpretation
$0{,}71 - 1{,}00$	The resulting structure is strong
$0{,}51 - 0{,}70$	The resulting structure is good
$0{,}26 - 0{,}50$	The resulting structure is weak
$\leq 0{,}25$	Unstructured

Source: Struyf et al. (1997)

3. Findings and Discussion

This study utilized panel data collected from government educational websites and central statistical offices across diverse Indonesian regions. Its objective is to identify significant factors influencing school enrolment rates, offering a comprehensive overview of the present educational scenario. The dataset, with its longitudinal nature, enables the analysis of changes over time and variations across regions and educational aspects.

3.1. Multidimensional Scaling (MDS) Analysis

Dimensional Diagram Representation of MDS at *SD* Level

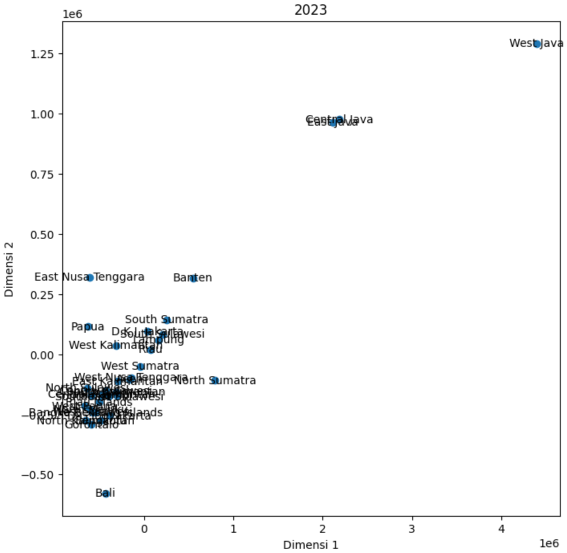


Fig.1. Visualization of the use of MDS at the *SD* level in 2023

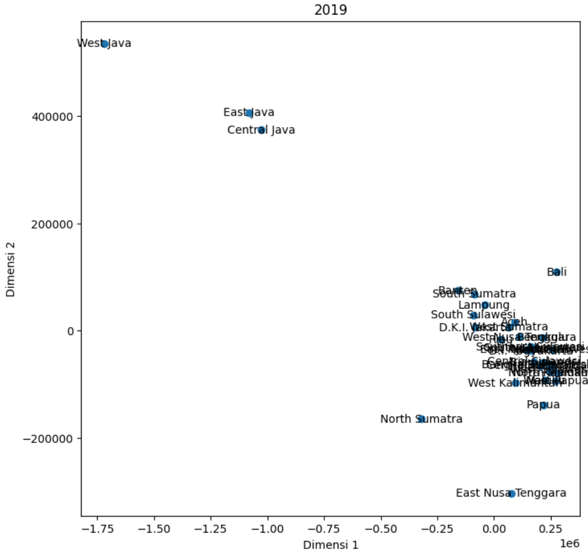


Fig.2. Visualization of the use of MDS at the *SMP* level in 2019

Fig. 1 displays the Multidimensional Scaling (MDS) analysis illustrating provincial groupings in Indonesia based on educational support at the elementary school level. The visualization showcases proximity between provinces, reflecting similarities in educational support aspects. The grouping remains consistent over the years, divided into four main groups. Notably, provinces like North Sumatra, East Java, Central Java, and West Java have isolated positions, with North Sumatra gradually aligning with the main cluster post-2019, signifying evolving educational characteristics. Central Java and East Java demonstrate convergence, indicating a similar approach to education development. Despite the geographic distance, certain provinces exhibit closeness in the visualization, emphasizing diverse government policies regarding educational support.

Dimensional Diagram Representation of MDS at SMP Level

Regarding the results in Fig.2 of the MDS visualization for the *SMP* level of education, it was found that there were similarities in the grouping pattern with the MDS visualization for the *SD* level. In the *SMP* MDS visualization, geographically close provinces reflect similarities in the characteristics of supportive education. For example, the provinces of West Java, Central Java and East Java appear to have a similar approach to educational development. Also bordering this group, the province of North Sumatra shows some similarities in its approach to education.

The MDS analysis consistently shows a stable grouping pattern for the *SMP* level from 2019 to 2023. Despite minor changes, particularly in East Java and Central Java's proximity, the overall grouping pattern remains constant. Notably, East Nusa Tenggara province has shifted away from the main cluster in the 2021 MDS visualization in Fig.3, indicating an alteration in educational characteristics or approach. However, in the subsequent year, the province returned to the main cluster. The *SMP* level analysis aligns with the previous *SD* level analysis, highlighting the consistent characteristics shared by pro-educationalists in certain Indonesian regions. Despite dynamics and shifts, the grouping pattern generally remains stable throughout the study period.

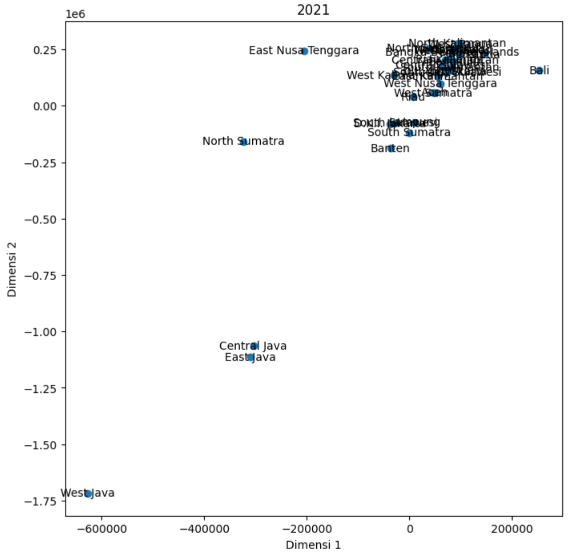


Fig. 3. Visualization of the use of MDS at the *SMP* level in 2021

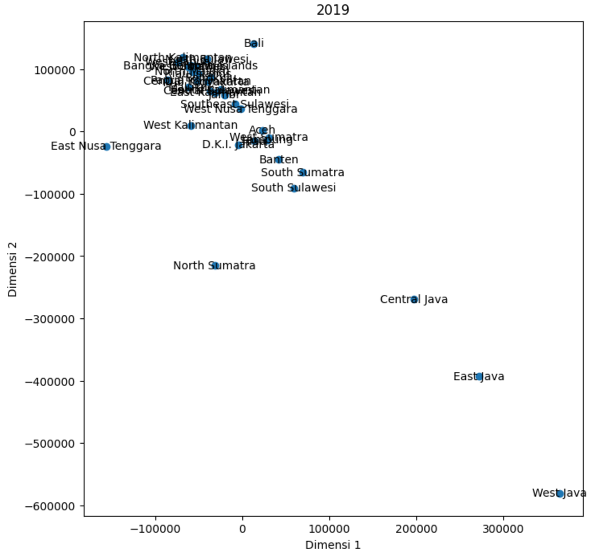


Fig.4. Visualization of the use of MDS at the *SMA* level in 2019

Dimensional Diagram Representation of MDS at SMA Level

In Fig. 4, as Multidimensional Scaling (MDS) plot for SMA level educational support in Indonesia, provinces in East Java, Central Java, and West Java cluster, showcasing varying education levels among them evident by point distances on the MDS plot. Additionally, Bali, East Nusa Tenggara, and North Sumatra stand isolated, indicating marked differences in educational support aspects. Over subsequent years (2020-2022), MDS analysis consistently forms seven similar groups, maintaining the isolation of regions like Central Java, East Java, West Java, North Sumatra, East Nusa Tenggara, and Bali, emphasizing notable differences in educational support characteristics within these regions.

The MDS analysis consistently showed 7 groups over 4 years, except for the last year 2023, in Fig.5, where only 6 groups formed. However, the characteristic disparities remained unchanged, hinting at a potential shift in the pattern formation or dynamics between regions. The decline in group numbers in recent years suggests a possible association or grouping of areas that were previously separate clusters. Despite fewer groups, the unchanged trait differences imply efforts to bridge educational support gaps across regions, indicating a shift in strategy or policy to promote uniform educational approaches, resulting in fewer but more similar groups.

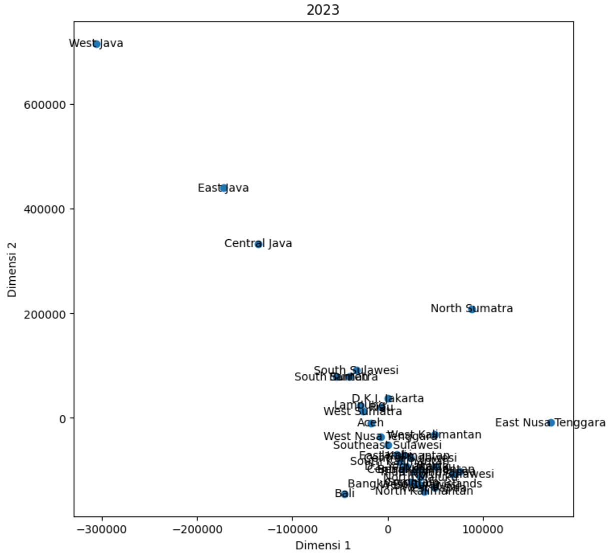


Fig. 5. Visualization of the use of MDS at the SMA level in 2023

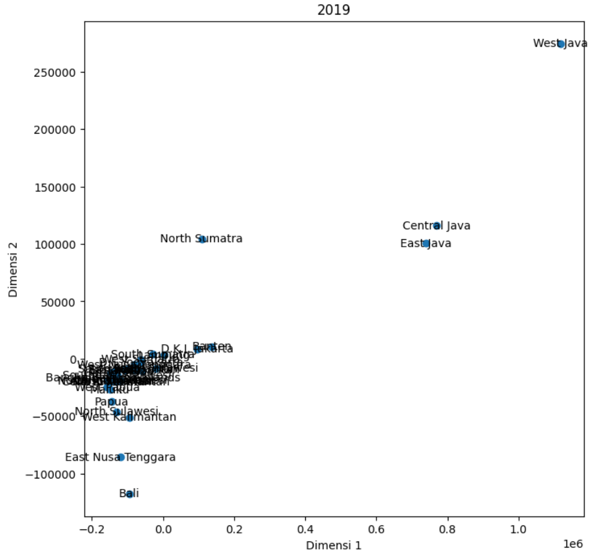


Fig.6. Visualization of the use of MDS at the SMK level in 2019

Dimensional Diagram Representation of MDS at SMK Level

Fig. 6 as a Multidimensional Scaling (MDS) analysis results illustrate the proximity between objects, helping identify objects with similar characteristics. Closer points denote higher similarity, while farther points signify differences. Bordering provinces like West Java, Central Java, and East Java, along with North Sumatra, DKI Jakarta, and Banten, are noticeable on the map. Conversely, provinces like Aceh, Bali, and Bengkulu exhibit relatively high affinity, indicating unique characteristics or specific educational approaches. Some provinces across Indonesia are more distant, reflecting the country's diverse education approaches. Despite this, proximity on the spatial map emphasizes similarities in educational support characteristics, especially among neighboring areas like Central Java and East Java. In subsequent years, MDS analysis consistently portrays similar grouping patterns, with certain provinces remaining isolated from the main cluster, while the proximity between East Java and Central Java suggests increasing similarity in characteristics.

In Fig.7, the multidimensional scaling analysis demonstrates a noteworthy deviation in object distribution compared to the visualizations of 2021, 2020, and 2019. The grouping pattern in 2021 shifted notably by 2022, although the major cluster count remained at three. Several points, including East Java, Central Java, and West Java regions, remained isolated from these clusters. The 2023 analysis shows a distribution pattern similar to 2022, implying stability in grouping patterns compared to the preceding years, 2021 and 2022. Three main clusters persist, with certain regions like East Java, Central Java, and West Java remaining isolated from the primary clusters. Overall, the MDS analysis at the SMK level shows greater variation in aspects of educational support. Consistent visualization from year to year indicates an ongoing pattern, with several provinces exhibiting closer traits that become more similar over time.

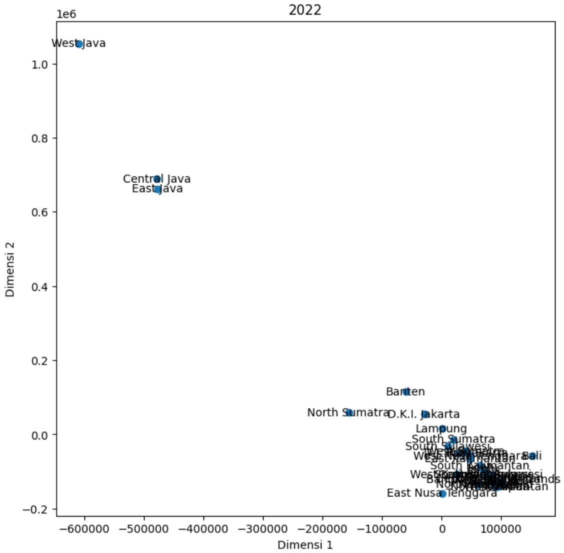


Fig.7. Visualization of the use of MDS at the SMK level in 2022

According to Faguet and Sanchez (2006), government spending in the education sector has a significant influence on school participation. In line with Dreher (2006) that government spending in the education sector is a supply factor that influences the quality of education and school participation. This is in line with the mapping that has been carried out with MDS at the elementary, middle school, high school and vocational school levels. These plots tend not to experience significant changes from the past 5 years, as it is known that the government continues to strive to increase the amount of the education budget, but the results obtained in the mapping MDS shows that there is no change in characteristics quite significant, so the addition and use of education funds must be given more attention and monitored closely by the government so that the policy on using these funds is appropriate and not misdirected so that Indonesian education continues to progress and develop and experiences significant changes every year.

Outliers

To detect outliers in regional clustering data regarding educational support, we observe points significantly distant or isolated from the main cluster in each MDS visualization. Notably, East Java, Central Java, West Java, Bali, D.K.I. Jakarta, Banten, East Nusa Tenggara, and North Sumatra stand out as isolated points, indicating substantial disparities in educational support. These outliers offer valuable insights *into areas necessitating special attention.*

3.2. Panel Data Regression Analysis

Identification of Variables Suspected of Having a Significant Impact

Table 3
Suspicious influencing variables

X1	Number of Schools
X2	Number of Muslim Students
X3	Number of Protestant Students
X4	Number of Catholic Students
X6	Number of Buddhist Students
X12	Number of Make Educators
X15	Number of Classrooms
X17	Number of Schools by Source of Packed Water
X21	Number of Schools by Source of Water Protected Well

In Table 3 the partial significance test indicates the significance of independent variables on the dependent variable in the regression model. If the p-value (Pr(>|z|)) in the Independent Variable column is < 0.05, it suggests a significant impact. Nine variables notably influence the proportion of students across educational levels.

Estimation Model

Chow test

Table 4
Chow test results

F Statistics	df	P-Value	Note
2.159	33, 637	2.20E-04	Fixed Effect

Related to Table 4 it is known that the probability of F Statistics is < 0.05. The probability value < 5%, so H0is rejected, the fixed effect model should be used in panel data regression modeling.

Hausman Test

Table 5
Hausman test results

Chi-Square	df	P-Value	Note
5.3355	9	0.8041	Random Effect

Related to Table 5 it is known that the chi-square probability is > 0.05. Probability value > 5%, so H0is accepted, the random effects model should be used in panel data regression modeling.

Breusch Pagan Test

Table 6
Breusch-Pagan Test Results

P-Value of Two Way Effect	P-Value of Individual Effect	P-Value of Time Effect	Note
< 2.2e-16	0.04236	0.1298	There is a two-way effect. However, after testing for cross section and time effects, there is only a cross section effect.

From the Hausman test and the Breusch-Pagan test, it can be concluded that the model to be estimated is a cross-sectional effect random effects data model.

Test Assumptions

Serial Correlation Test for Error Components

Table 7
Durbin-Watson Test Results

DW	P-Value
2.0585	0.7544

The Durbin-Watson test in Table 7 resulted in a DW value of 2.0585 with a corresponding p-value of 0.7544. Given the DW value's proximity to 2 and the high p-value (0.7544), there is insufficient evidence to reject the null hypothesis. Thus, the test does not suggest a significant serial correlation in idiosyncratic errors within this panel regression model.

Homoscedasticity Acceptance Test

Testing the assumption of homoscedasticity with Robust covariance estimator for heteroscedasticity

Based on the homoscedasticity acceptance test, it was found that there was no difference in the coefficients of the independent variables in the t-test with the covariance matrix. This satisfies the robust test results for the heteroscedasticity of the covariance matrix or the assumption that the residual variance-covariance structure is the same.

The results of the estimation are displayed in Table 8.

Table 8
Test Results of The Robust Covariance Estimator

Intercept	T test of coefficients (Estimate)	Coefficients (Estimate)
X1	7.01E-04	7.01E-04
X2	-1.92E-05	-1.92E-05
X3	3.64E-08	3.64E-08
X4	2.74E-08	2.32E-08
X6	2.32E-08	3.11E-08
X12	3.11E-08	8.93E-08
X15	2.62E-08	2.62E-06
X17	1.35E-04	1.35E-04
X21	8.42E-07	8.42E-07

Scoring Model

The panel regression results indicate a random effects model run on 34 individual units with 4 observation times, totaling 680 observations. The idiosyncratic component has a variance of 6.292e-05 and a standard deviation of approximately 7.932e-03. Individual component variances are zero, implying no explainable individual variation in the model.

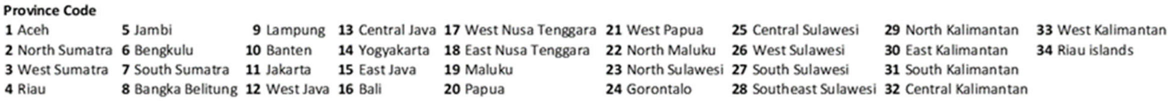
Regarding coefficient estimation, variables X1, X2, X3, X4, X6, X12, X15, and X17 significantly influence the response variable “proportion” with very low p-values (< 2.2e-16), denoting statistical significance. However, variable X21 has a p-value of 0.1928741, indicating insignificance at a given level of significance.

The coefficient of determination (R-squared) is approximately 0.92223, signifying that this model explains about 92.22% of the data variability. The best formula for the panel regression equation is as follows:

$$\begin{aligned}\boldsymbol{Proporsi_{it}} = {} & (7{,}01E-04) - (1{,}92E-05)\boldsymbol{X1_{it}} + (3{,}64E-08)\boldsymbol{X2_{it}} + (2{,}74E-08)\boldsymbol{X3_{it}} \\ & + (2{,}32E-08)\boldsymbol{X4_{it}} + (3{,}11E-08)\boldsymbol{X6_{it}} + (8{,}93E-08)\boldsymbol{X12_{it}} \\ & + (2{,}62E-06)\boldsymbol{X15_{it}} + (1{,}35E-04)\boldsymbol{X17_{it}} + (8{,}42E-07)\boldsymbol{X21_{it}}\end{aligned} \tag{22}$$

3.3. Cluster Analysis

Cluster analysis can help identify common characteristics between regions and support targeted interventions (Arisanti et al., 2023).



Province Code

1 Aceh
2 North Sumatra
3 West Sumatra
4 Riau
5 Jambi
6 Bengkulu
7 South Sumatra
8 Bangka Belitung
9 Lampung
10 Banten
11 Jakarta
12 West Java
13 Central Java
14 Yogyakarta
15 East Java
16 Bali
17 West Nusa Tenggara
18 East Nusa Tenggara
19 Maluku
20 Papua
21 West Papua
22 North Maluku
23 North Sulawesi
24 Gorontalo
25 Central Sulawesi
26 West Sulawesi
27 South Sulawesi
28 Southeast Sulawesi
29 North Kalimantan
30 East Kalimantan
31 South Kalimantan
32 Central Kalimantan
33 West Kalimantan
34 Riau islands

Fig. 8. Province code in the clustering visualization

The results in Fig.9 of regional grouping using cluster analysis shows that it consists of two clusters at *SD* level, with the first cluster containing West Java, Central Java and East Java, while the second cluster contains other provinces. The results in Fig.10 of regional grouping using cluster analysis show that it consists of two clusters at the *SMP* level, the first cluster being West Java, Central Java, East Java, East Nusa Tenggara, South Sulawesi, West Kalimantan and includes North Sumatra, while the second cluster includes clusters includes other provinces.

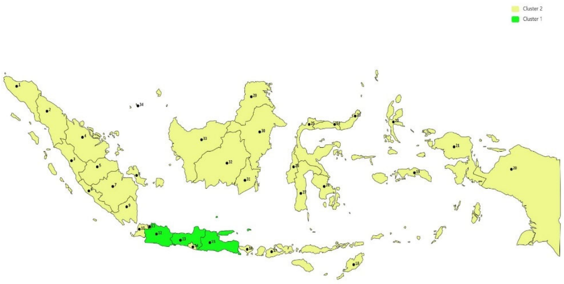


Fig. 9. Visualization clustering at *SD* level

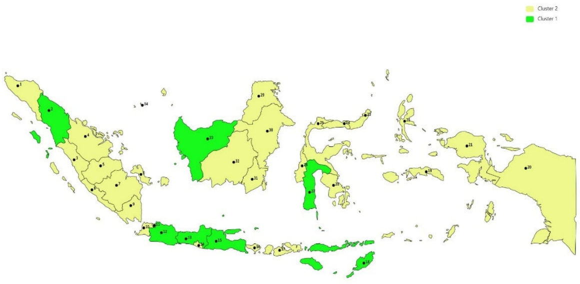


Fig.10. Visualization clustering at the *SMP* level

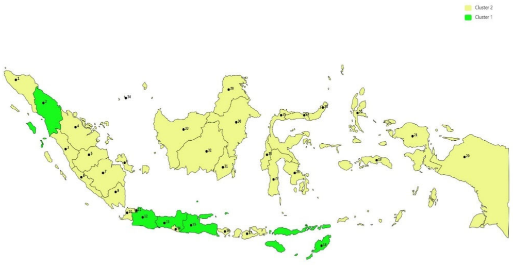


Fig. 11. Visualization clustering at the *SMA* level

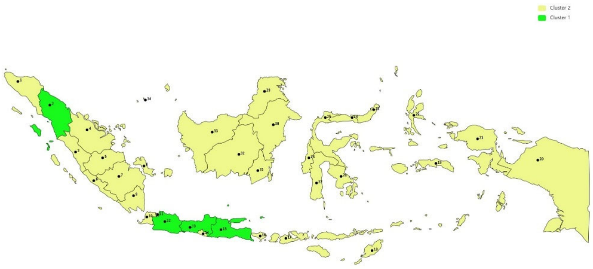


Fig.12. Visualization clustering at *SMK* level

The results in Fig.11 of regional grouping using cluster analysis show that the *SMA* level consists of two clusters, with the first cluster including West Java, Central Java, East Java, East Nusa Tenggara and North Sumatra, while the second cluster includes other provinces. The last results of regional grouping using cluster analysis show that it consists of two clusters at the *SMK* level in Fig.12, with the first cluster containing West Java, Central Java, East Java and North Sumatra, while the second cluster contains other Provinces. All clusters formed at each level show that they are only formed into 2 clusters, namely low and high levels of educational participation based on the silhouette method, but this is different from the results of research produced by Aryawwan et al. (2022) it is known that there are 5 clusters created, namely cluster 1 as a province with a high level of educational participation, cluster 2 as a province with a medium level of educational participation, cluster 3 as a province with a low level of educational participation, cluster 4 as a province with a very low level of educational participation, and cluster 5 as provinces with unknown educational participation rates. Observers concluded that the more clusters there are, the more detailed results will be provided about the real situation regarding the number of education participants in Indonesia, so it is hoped that detailed information will make it easier for the government to draw conclusions and make policies to increase education participants in a region.

4. Conclusion

The Multidimensional Scaling (MDS) evaluation revealed distinct educational promotion clusters in Indonesia. Certain areas, such as East Java, Central Java, and West Java, showed significant disparities in educational support at primary and upper-secondary levels. However, regional grouping patterns varied more within the *SMA* and *SMK* educational tiers.

Regression analysis on the panel data highlighted significant correlations between educational support factors and student distribution across academic levels. Positive effects were observed for various variables promoting educational growth, while certain factors in specific regions had adverse effects on student proportions. This emphasizes the need for public investment to elevate educational standards.

Cluster analysis outcomes identified zones with similar educational support patterns and unique clusters, providing insights into shared characteristics among regions. This information serves as a basis for strategic policy interventions, aiming to improve Indonesia's educational quality. With a deeper understanding of the determinants influencing education, there's an optimistic outlook for implementing effective measures to enhance educational quality and foster a brighter future for generations to come.

Author Contributions

Conceptualization, Restu Arisanti and Yuyun Hidayat; Methodology, Titi Purwandari and Irlandia Ginanjar; software, Arum Putri Juniarsih and Janatin.; validation, Restu Arisanti; formal analysis, Restu Arisanti and Arum Putri Juniarsih; investigation,

Restu Arisanti; resources, Yuyun Hidayat and Restu Arisanti; data curation, Arum Putri Juniarsih and Janatin.; writing—original draft preparation, Restu Arisanti.; writing—review and editing, Restu Arisanti; visualization, Janatin; supervision, Restu Arisanti; project administration,Yuyun Hidayat; funding acquisition, Yuyun Hidayat. All authors have read and agreed to the published version of the manuscript.

Acknowledgement

The authors would like to express their sincere gratitude to the Directorate of Research and Community Engagement of Padjadjaran University for their valuable support during the writing of this paper and for providing financial support for publication in this journal.

References

Adlakha, K., & Sharma, S. (2020). Brand positioning using Multidimensional Scaling technique: An application to herbal healthcare brands in Indian market. *Vision*, *24*(3), 345-355. DOI: https://doi.org/10.1177/0972262919850930.

Arisanti, R., Pontoh, R., Winarni, S., & Aini, S. (2023). Assessing service availability and accessibility of healthcare facilities in Indonesia: A spatially-informed correspondence analysis with visual approach. *Decision Science Letters*, *12*(3), 591-604. DOI: https://doi.org/10.5267/j.dsl.2023.4.005

Aryawwan, I. G. N., Sudiana, I., & Suardika, I. M. (2022). Clustering analysis of participation rate in education in Indonesia. *International Journal of Advanced Science and Technology, 28*(7), 505-516. DOI: https://doi.org/10.21834/ijast.2022.28.7.505

Bai, J., & Li, K. (2014). Theory and methods of panel data models with interactive effects. *The Annals of Statistics, 42*(1), 142–70. DOI: http://www.jstor.org/stable/43556275

Baltagi, B. H. (2008). *Econometric Analysis of Panel Data* (Vol. 4). Chichester, England: John Wiley & Sons Ltd.

Baltagi, B. H., Feng, Q., & Kao, C. (2012). A Lagrange Multiplier test for cross-sectional dependence in a fixed effects panel data model. *Journal of Econometrics*, *170*(1), 164-177. DOI: https://doi.org/10.1016/j.jeconom.2012.04.004

Bhattacharjee, P. S., Fujail, A. K. M., & Begum, S. A. (2017, December). A comparison of intrusion detection by K-means and fuzzy C-means clustering algorithm over the NSL-KDD dataset. In *2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)* (pp. 1-6). IEEE. DOI: https://doi.org/10.1109/ICCIC.2017.8524401

Binkley, J. K., & Young, J. (2018). The Chow Test with Time Series-Cross Section Data. *Available at SSRN 3212712*. SSRN. DOI: http://dx.doi.org/10.2139/ssrn.3212712

Breusch, T. S., & Pagan, A. R. (1980). The Lagrange multiplier test and its applications to model specification in econometrics. *The review of economic studies*, *47*(1), 239-253. DOI: https://doi.org/10.2307/2297111

Dreher, A. (2006). Does globalization affect growth? Evidence from a new index of globalization. *Applied economics*, *38*(10), 1091-1110. DOI: 10.1080/00036840500392078.

Erda, G., Gunawan, C., & Erda, Z. (2023). Grouping Of Poverty In Indonesia Using K-Means With Silhouette Coefficient. *Parameter: Journal of Statistics, 3*(1), 1-6. DOI: https://doi.org/10.22487/27765660.2023.v3.i1.16435

Faguet, J. P. & Sanchez, F. (2006). Decentralization's Effects on Educational Outcomes in Bolivia and Colombia. *World Development, 36*(7), 1294– 1316. DOI: https://doi.org/10.1016/j.worlddev.2007.06.021

Greene, W. H. (2000). *Solutions and Applications Manual: Econometric Analysis*. 6th ed., New Jersey: Prentice Hall.

Hsiao, C. (2003). *Analysis of Panel Data*. Cambridge University Press, Cambridge.

Ioan, B., Malar Kumaran, R., Larissa, B., Anca, N., Lucian, G., Gheorghe, F., ... & Mircea-Iosif, R. (2020). A panel data analysis on sustainable economic growth in India, Brazil, and Romania. *Journal of Risk and Financial Management*, *13*(8), 170. DOI: https://doi.org/10.3390/jrfm13080170

Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. (6th ed.). Upper Saddle River, NJ: Prentice-Hall.

Rabinowitz, G. B. (1975). An introduction to nonmetric multidimensional scaling. *American Journal of Political Science*, 19(2), 343-390. DOI: https://doi.org/10.2307/2110441.

Struyf, A., Hubert, M., & Rousseeuw, P. (1997). Clustering in an object-oriented environment. *Journal of Statistical Software*, *1*, 1-30.