Parameter: Journal of Statistics
Available at https://bestjournal.untad.ac.id/index.php/parameter
Vol. 5 No. 1 2025, 61 - 67
DOI: *https://doi.org/10.22487/27765660.2025.v5.i1.17440*
eISSN: 2776-5660

CLUSTER ANALYSIS OF HIGHEST EDUCATION COMPLETED IN EAST JAVA PROVINCE WITH SPHERICAL K-MEANS METHOD

Mohammad Dian Purnama[1]

[1]Universitas Negeri Surabaya, East Java, Indonesia

***e-mail**: [1*]mohammaddian.20053@mhs.unesa.ac.id

ABSTRACT

One of the key pillars of development that greatly aids in the social and economic advancement of civilization is education. The purpose of this study is to use the Spherical K-Means Clustering method to evaluate the distribution and degree of educational attainment in districts/cities in East Java Province. This approach was selected because it can group vector-based data according to directional similarity, making it appropriate for multidimensional data. Compared to traditional K-Means, this method uses cosine similarity, making it more suitable for data in proportion form where the focus is on patterns rather than magnitude. Based on education-related variables, including never attending school, not graduating from primary school, graduating from primary school, graduating from junior high school, graduating from senior high school, and graduating from university, this analysis groups regions. Based on the clustering results, three significant groups of districts and cities were identified. Cluster 1 is dominated by urban districts and cities surrounding Surabaya and shows strong performance in both secondary and tertiary education. Cluster 2 represents regions with a relatively balanced distribution between primary and secondary education, with moderate levels of higher education. Cluster 3 includes districts with a high proportion of basic education and lower levels of secondary and tertiary education. These results can help stakeholders develop more targeted and efficient education policies by providing insight into the educational disparities across East Java.

Keywords: Education, Cluster Analysis, Spherical K-Means Method.

Cite: Purnama, M. D. (2025). Cluster analysis of highest education completed in East Java Province with spherical K-means method. *Parameter: Journal of Statistics*, *5*(1), 61–67. https://doi.org/10.22487/27765660.2025.v4.i1.17440

[*]Received: May 2025; Reviewed: May 2025; Published: Jul 2025

INTRODUCTION

Education is one of the main pillars of a country's development, contributing directly to the social and economic progress of society (Lestari, 2013). In Indonesia, efforts to improve the quality and equity of education are one of the main focuses of the government. East Java province is the second most populous province in Indonesia after West Java province. With 14.90% of Indonesia's population, East Java has unique challenges and opportunities in achieving education goals (Badan Pusat Statistik, 2024).

Cluster analysis is a statistical analysis method that aims to place a set of objects into two or more groups based on their similarities and various other attributes(Wirawan & Prasetyawan, 2023). This study aims to analyze the distribution and level of educational attainment in districts/municipalities in East Java Province through a clustering approach. This study uses the Spherical K-Means Clustering method to cluster districts/municipalities based on education level indicators to identify patterns and differences in educational attainment.

Spherical K-Means Clustering is a method designed to cluster vector-based data by taking into account directional similarity, making it suitable for data that has a specific distribution in a multidimensional space (Schubert et al., 2021). In this context, the method will help in identifying clusters that have similar education patterns among districts/municipalities in East Java. Previous research related to Spherical K-Means Clustering was conducted by Rini et al., (2021) who discussed earthquakes in Bengkulu Province. Not only that, research related to clustering according to education aspect has been conducted using K-Means and K-Medoids (Tusyakdiah et al., 2023). The results of this clustering are expected to provide deeper insights into the distribution of education in various regions and assist stakeholders in formulating more targeted policies.

The urgency of this study lies in the increasing need for more accurate and nuanced regional education profiling, especially in East Java, which has significant variations in socioeconomic and demographic characteristics across its districts and municipalities (Ristanto, 2022). Traditional clustering methods often fail to capture the angular relationships among proportional data such as education levels, which can lead to less meaningful groupings (Sun & Sajda, 2024). By utilizing Spherical K-Means, this study aims to highlight latent structures in educational attainment patterns that are otherwise overlooked. In contrast to prior studies, which mainly employed Euclidean based approaches or focused on broader national-level analyses (Berry et al., 2010). This research offers a more localized and directionally aware clustering approach. Furthermore, this study contributes to the methodological expansion of educational data analysis by applying a clustering technique that aligns better with the compositional nature of categorical percentage data. It is expected that the findings will not only enrich academic discourse but also serve as a practical basis for policy formulation, planning, and intervention by local government and educational institutions (Resposo, 2024).

By understanding the characteristics of each cluster, stakeholders can more effectively design education interventions that suit the specific needs of each region (Aydemir, 2024). Through this research, it is expected that clusters with different educational strengths and challenges will be identified, providing a solid basis for better and more targeted education planning in East Java province. This research also contributes to the academic literature by offering a new approach in education data analysis through vector-based clustering.

MATERIALS AND METHODS

Materials

The Central Bureau of Statistics (BPS) of East Java Province provided the study's highest education completion data, which was taken from the East Java Youth Profile 2023. Additional education classifications include never having attended school (NS), not having completed elementary school (NE), having completed primary school (CE), having completed junior high school (CJ), having completed senior high school (CS), and having graduated from university (GU).

Methods

Spherical K-Means Clustering was used in this research process. The purpose of this analysis is to cluster districts/municipalities in East Java Province based on educational characteristics involving several variables, such as the level of never going to school, not finishing elementary school, finishing elementary school, finishing junior high school, finishing high school, and finishing college. Descriptive statistical analysis was conducted to provide an overview of the data used. After that, the Spherical K-Means method was applied to determine the resulting education clusters. This analysis process includes

evaluating the optimal number of clusters and interpreting the characteristics of each cluster. Finally, the clustering results will be applied to see the pattern of education in each district.

Descriptive Statistic

Descriptive statistical analysis, which entails interpreting the collected data, is one method of data examination. With the use of variables like mean, standard deviation, lowest and greatest values, and so on, the goal is to provide a comprehensive overview of the data. Descriptive statistics aid in the transformation of data into more easily understood information and provide an understanding of the relationship between the study's variables (Purnama, 2024).

Spherical K-Means Cluster Analysis

An object's classification based on shared characteristics is the primary goal of the multivariate approach known as cluster analysis. There is a high degree of similarity between an object's qualities inside a cluster, but a low degree of similarity between an object's characteristics within a cluster and those of other clusters. Stated differently, there is least diversity within a cluster and most diversity between clusters (Rini et al., 2021). In order to explain the technique for classifying an object into a specific cluster based on the closest distance to the center (means), MacQueen proposed using K-Means.

There are three steps in this method (Johnson & Wichern, 2002). First, Divide the objects into the first K cluster. Begin by logging the items and group the ones whose distances are closest to the center (mean). Euclid's distance is typically used to compute distance when utilizing standardized or non-standardized observations. For both the cluster that gained the new object and the cluster that lost it, recalculate the centroid. To determine the centroid of a group, compute the average value derived from the data in the equation (1) (Murphy et al., 2024):

$$\mu_{kj} = \frac{1}{n_k} \sum_{x_i \epsilon C_k} x_{ij} \; ; k = 1,2, \dots, K \; ; \; j = 1,2, \dots, p \tag{1}$$

where μ_{kj} represents the j-th component of the centroid vector for cluster C_k and x_{ij}. Let C_k is the set of n_k objects in cluster k, the number of groups $K \leq n$ is pre-specified by the practitioner and remains fixed C_k and x_{ij} is the j-th feature of observation i belonging to that cluster. The complete centroid vector $\boldsymbol{\mu}_k$ is then written in the equation (2):

$$\boldsymbol{\mu}_k = \left(\mu_{k1}, \mu_{k2}, \dots, \mu_{kp}\right)^{\top} \tag{2}$$

These centroids therefore correspond to the arithmetic mean vector of the observations in cluster C_k. Until there are no more moving items, step 2 is repeated. To evaluating a partition's result is a crucial factor in cluster analysis (Weatherill & Burton, 2009). The goal of the algorithm is to minimize the total within-cluster sum of squares (TWCSS). Formally, the defined in the equation (3):

$$TWCSS = \sum_{i=1}^{n} \sum_{k=1}^{K} I(x_i \in C_k) \left|\left|\mathbf{x}_i - \boldsymbol{\mu}_k\right|\right|_2^2 \tag{3}$$

where $\left|\left|\mathbf{x_i} - \boldsymbol{\mu}_k\right|\right|_2^2$ denotes the squared Euclidean distance for the centroid distance, $I(x_i \in C_k)$ is an indicator function that returns 1 if the observation x_i belongs to cluster C_k, and 0 otherwise (Murphy et al., 2024). The circle distance between two objects, if the data is in circular units, is $\cos(\alpha_1 = \alpha_2)$ where the angles are α_1 and α_2 (Jammalamadaka & SenGupta, 2001). For the hypersphere unit, the inner product of y_1 and y_2, indicated by $\langle y_1, y_2 \rangle$, is the standard definition of the cosine similarity between two vector units, y_1 and y_2. Assume that K clusters need to be created from n spherical data points. The spherical K-Means algorithm minimizes:

$$\sum_{k=1}^{K} \sum_{i=1}^{n} \mu_{ki} \langle y_i, p_k \rangle = \sum_{k=1}^{K} \sum_{i=1}^{n} \mu_{ki} (1 - \cos(y_i, p_k)) \tag{4}$$

$$= \sum_{k=1}^{K} \sum_{i=1}^{n} \mu_{ki} \left(1 - \frac{\langle y_i, p_k \rangle}{|y_i||p_k|}\right)$$
$$= \sum_{k=1}^{K} \mu_k \left(\sum_{i=1}^{n} \mu_{ki} - \left\langle \sum_{i=1}^{n} \mu_{ki} \frac{y_i}{|y_i|}, \frac{p_k}{|p_k|} \right\rangle\right)$$

For all clusters $c(i) \in \{1,2,\ldots,k\}$ and for all centroids p_k, where:

$$\mu_{ki} = \begin{cases} 1, \text{ if } c(i) = j \\ 0, \text{otherwise} \end{cases} \quad (5)$$

This means that μ_{ki} equals 1 if the *i*-th observation is assigned to the *j*-th cluster (i.e., $c(i) = j$), and 0 otherwise. It is a binary indicator representing whether observation *i* belongs to cluster *j*, and is equivalent to the latent variable $I(\cdot)$ in the K-means formulation. The process of determining the ideal cluster members and cluster centroids, the ideal cluster centroids for cluster members, and the ideal cluster members for cluster centroids make up the iterative optimization process. The following are the spherical K-Means steps (Hornik et al., 2012).

1. First, ascertain the number of clusters *K* and initialize the centroids. This initial step involves specifying the number of clusters *K*, then assigning the initial centroids either randomly or based on specific criteria. No specific mathematical formulation is directly applied at this point, but it sets the foundation for the iterative process.
2. Next, assign each object to a cluster based on the closest distance or highest similarity to the centroid. The cluster membership is determined by minimizing the squared Euclidean distance between each object and the centroid, as formulated in equation (3). In this context, the objective is to reduce the total within-cluster sum of squares (TWCSS).
3. Determine the cluster membership indicator based on the assignment. An object is marked as a member of a specific cluster using a binary indicator variable, as expressed in equation (5). This indicator takes the value 1 if an object belongs to a given cluster and 0 otherwise.
4. Recalculate the centroid of each cluster using the newly assigned members. The centroid of a cluster is updated by computing the arithmetic mean of all observations within that cluster, as shown in equation (1). The complete centroid vector is then represented in equation (2).
5. For data on the hypersphere, update the centroid using cosine similarity. In the case of spherical K-Means, the centroid is redefined to maximize the cosine similarity between objects and centroids, as described in equation (4). This ensures that the updated centroid lies on the unit hypersphere.
6. Repeat steps 2 through 5 iteratively until convergence is achieved. The algorithm proceeds by alternating between object reassignment in equation (3) and (5) and centroid updating in equation (1), (2), and (4) until the cluster assignments no longer change significantly.

RESULTS AND DISCUSSION

The data on the highest level of education attained were obtained from the East Java Youth Profile 2023 published by the Central Bureau of Statistics (BPS) of East Java Province. The dataset includes all 38 districts and municipalities in East Java Province, with each region representing a single observation. The education categories include: Never attended school (NS), Not completed elementary school (NE), Completed elementary school (CE), Completed junior high school (CJ), Completed senior high school (CS), and Graduated from university (GU). Each district or municipality has complete data across these six educational attainment categories. Table 1 presents the descriptive statistics (mean, minimum, maximum, and standard deviation) for each educational category across the 38 regions in the dataset.

Table 1. Descriptive Statistics Every Variable

	NS	NE	CE	CJ	CS	GU
Mean	0.53	1.02	9.90	38.23	39.67	10.64
Standard Deviation	0.65	0.87	6.31	5.44	6.39	3.96
Minimum	0.00	0.00	2.57	26.72	24.80	3.00
Maximum	2.97	3.42	27.96	51.89	51.65	19.61

The value of K must be ascertained before applying the spherical K-Means algorithm. To facilitate the representation of the number of clusters containing the highest education data completed in the Province of East Java, K = 3 is chosen in this study. Figure 1 displays the findings of the spherical K-Means cluster analysis with K = 3:

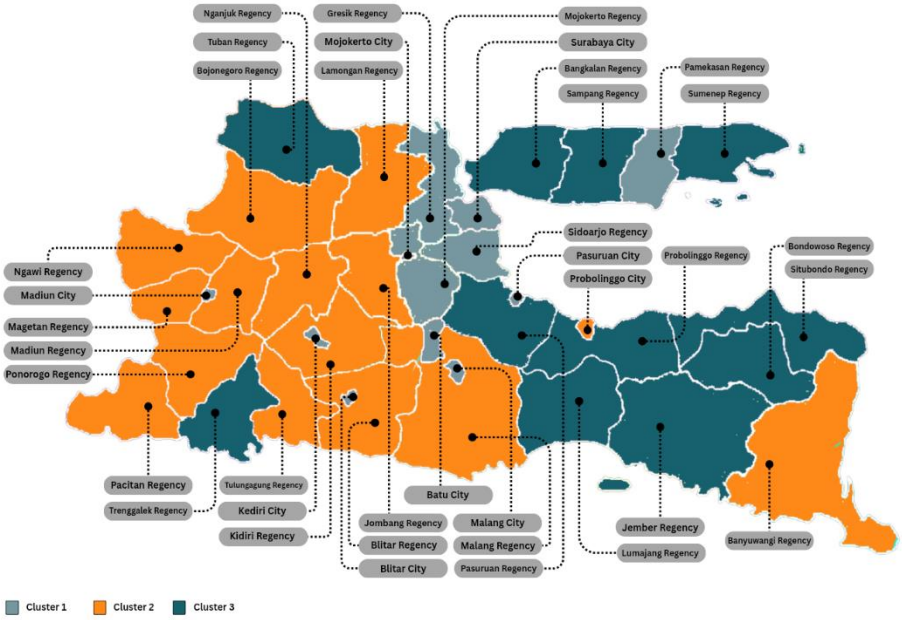

Figure 1. Map of Cluster

To identify the clusters formed based on the similarity of education patterns in each region, the Spherical K-Means approach is being applied to education data in districts and municipalities within East Java Province. Each district or city within a cluster is chosen based on the similarity of education distribution or similar levels of educational attainment because a number of education variables—such as the degree of never attending school, not graduating from elementary school, graduating from elementary school, graduating from junior high school, graduating from high school, and graduating from college—are used as cluster determinants. According to Figure 1, regions that are part of a cluster typically have comparable educational traits and exhibit a particular pattern that is represented by a distinct color for each district or municipality.

Most of the education challenges in East Java province are related to variations in education levels across districts/cities. This is due to the diverse social, economic and geographical factors in each region. Some areas tend to have higher levels of educational attainment, especially in big cities, while rural or remote areas show greater challenges in terms of access and quality of education. This clustering of educational attainment levels uses the Spherical K-Means method, which is grouped into 3 clusters. Based on the Spherical K-Means cluster analysis, the following table of education cluster characteristics was produced:

Table 2. Cluster characteristic with K=3

Cluster	NS	NE	CE	CJ	CS	GU
1	0.29	0.64	5.08	32.62	46.25	15.12
2	0.24	0.63	7.70	42.21	39.32	9.91
3	1.18	1.97	18.16	38.94	32.98	6.77

In the clustering analysis using the Spherical K-Means method on education data from districts/municipalities in East Java Province, three main clusters were identified based on the average value of education levels achieved. The following is a summary of the average distribution of education in each cluster.

Cluster 1 is the Education Cluster with High Proportions of Secondary and Primary Education. This cluster reflects districts that have a high proportion of secondary and primary education levels, with the highest average scores in the completion of junior high school and completion of senior high school categories. While there is a significant contribution from tertiary education completion, basic education categories such as primary education completion show lower values compared to secondary education. This cluster includes Sidoarjo district, Mojokerto district, Gresik district, Pamekasan district, Kediri city, Blitar city, Malang city, Pasuruan city, Mojokerto city, Madiun city, Surabaya city and Batu city.

This pattern suggests that while access to secondary education is relatively successful, efforts should be directed toward strengthening early education quality and increasing transition rates to higher education. Interventions might include scholarship support, improved infrastructure, and teacher training programs targeting early and tertiary levels.

Cluster 2 is the Education Cluster with a Focus on Primary and Secondary Education. This cluster shows a relatively high distribution in the primary and secondary education categories, with high mean values in the junior secondary and senior secondary categories. The Higher Education Graduation category shows a lower mean value, indicating that provinces in this cluster may have strengths in secondary education with lower contributions to further education. This cluster includes Pacitan District, Ponorogo District, Tulungagung District, Blitar District, Kediri District, Malang District, Banyuwangi District, Jombang District, Nganjuk District, Madiun District, Magetan District, Ngawi District, Bojonegoro District, Lamongan District, and Probolinggo City.

Given this composition, policies should focus on building pathways from secondary to tertiary education, such as career counseling, university outreach programs, and improving access to higher education institutions. It is also crucial to ensure that vocational and technical education options are available to meet the needs of students who may not pursue academic tertiary paths.

Cluster 3 is the Education Cluster with High Proportion in Primary Education and Low in Secondary Education. This cluster stands out with the highest mean scores in the categories of Never been to school, Not completed primary school, and Completed primary school, indicating a large proportion in primary education. However, secondary education levels such as High School Graduation and College Graduation show lower mean values, indicating that districts in this cluster have a greater focus on primary education compared to secondary education. This cluster includes Trenggalek district, Lumajang district, Jember district, Bondowoso district, Situbondo district, Probolinggo district, Pasuruan district, Tuban district, Bangkalan district, Sampang district and Sumenep district.

This suggests a pressing need to reduce school dropouts and improve transition to secondary education. Policymakers should consider conditional cash transfers, school feeding programs, and targeted outreach to marginalized communities to encourage school continuation. Moreover, enhancing the quality and relevance of primary education may help foster better long-term outcomes.

CONCLUSION

According to the results of the analysis, there are significant disparities in educational attainment levels among districts and municipalities in East Java Province. Cluster 1 is dominated by urban districts surrounding Surabaya and is characterized by high levels of both secondary and tertiary education. Cluster 2 includes inland and transitional regions that show a stronger focus on primary and secondary education, with moderate progression to higher education. Cluster 3 consists largely of more rural and eastern areas, where educational attainment is concentrated at the primary level, with limited access to secondary and higher education. These findings provide valuable insights for regional education policy, particularly in targeting interventions based on local education profiles.

REFERENCES

Aydemir, R. (2024). Examining the Cluster Life Cycle in the Process of Economic Development. *Journal of Policy Options*, *7*(1), 18–26.

Badan Pusat Statistik. (2024). *Statistik Indonesia 2024*. https://www.bps.go.id/id/publication/2024/02/28/c1bacde03256343b2bf769b0/statistik-indonesia-2024.html

Berry, H., Guillén, M. F., & Zhou, N. (2010). An institutional approach to cross-national distance. *Journal of International Business Studies*, *41*(9), 1460–1480. https://doi.org/10.1057/jibs.2010.28

Hornik, K., Feinerer, I., Kober, M., & Buchta, C. (2012). Spherical k -Means Clustering . *Journal of Statistical Software*, *50*(10). https://doi.org/10.18637/jss.v050.i10

Jammalamadaka, S. R., & SenGupta, A. (2001). *Topics in Circular Statistics* (Vol. 5). WORLD SCIENTIFIC. https://doi.org/10.1142/4031

Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis*.

Lestari, W. S. (2013). Kendala Pelaksanaan Pembelajaran Jarak Jauh (PJJ) Dalam Masa Pandemi Ditinjau dari Media Pembelajaran. *Journal of Chemical Information and Modeling*, *53*(9), 1689.

Murphy, K., López-Pernas, S., & Saqr, M. (2024). Dissimilarity-Based Cluster Analysis of Educational Data: A Comparative Tutorial Using R. In *Learning Analytics Methods and Tutorials* (pp. 231–283). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-54464-4_8

Purnama, M. D. (2024). An Implementation of Ordinal Probit Regression Model on Factor Affecting East Java Human Development Index. *Engineering, MAthematics and Computer Science Journal (EMACS)*, *6*(3), 219–225. https://doi.org/10.21512/emacsjournal.v6i3.12094

Resposo, R. B. (2024). Exploring the Implementation of Strategic Intervention Materials: Basis for Policy Development. *International Journal For Multidisciplinary Research*, *6*(4). https://doi.org/10.36948/ijfmr.2024.v06i04.25922

Rini, D. S., Sriliana, I., Novianti, P., Nugroho, S., & Jana, P. (2021). Spherical K-Means method to determine earthquake clusters. *Journal of Physics: Conference Series*, *1823*(1), 012043. https://doi.org/10.1088/1742-6596/1823/1/012043

Ristanto, A. D. (2022). Effect of Socio-Economic Characteristics and Cultural Areas on the Educated Poor in East Java Province. *Journal of International Conference Proceedings*, 150–159. https://doi.org/10.32535/jicp.v5i4.1930

Schubert, E., Lang, A., & Feher, G. (2021). *Accelerating Spherical k-Means* (pp. 217–231). https://doi.org/10.1007/978-3-030-89657-7_17

Sun, X., & Sajda, P. (2024). Circular Clustering With Polar Coordinate Reconstruction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *21*(5), 1591–1600. https://doi.org/10.1109/TCBB.2024.3406341

Tusyakdiah, H., Hasanah, I., Panggol, S. A., Ramdhanti, T., Permatasari, R., Cusanti, C., & Widodo, E. (2023). Implementasi metode K-Means dan K-Medoids Pada Pengelompokan Provinsi Indonesia Berdasarkan Aspek Pendidikan Pemuda. *Community Services and Social Work Bulletin*, *3*(1), 1–10.

Weatherill, G., & Burton, P. W. (2009). Delineation of shallow seismic source zones using K -means cluster analysis, with application to the Aegean region. *Geophysical Journal International*, *176*(2), 565–588. https://doi.org/10.1111/j.1365-246X.2008.03997.x

Wirawan, A., & Prasetyawan, D. (2023). Analisis cluster data latar belakang ekonomi mahasiswa untuk rekomendasi penentuan uang kuliah tunggal dengan model K-Modes. *INFOTECH : Jurnal Informatika & Teknologi*, *4*(2), 234–246. https://doi.org/10.37373/infotech.v4i2.898