WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Grouping of District Based on Poverty Indicators Using The K-Medoids Cluster Method

Rita Rahmawati [a*]

[a] *Department of Statistics, Faculty of Science and Mathematics, Diponegoro University, Semarang, Indonesia*
[*] *e-mail: ritarahmawati@gmail.com*

A B S T R A C T

Poverty remains a major challenge to sustainable development in Indonesia, with significant disparities across regions. This study applies the K-Medoids clustering algorithm to classify 26 districts in Majalengka Regency based on five poverty-related indicators: the number of poor residents, stunting prevalence, school dropout rate, unemployment rate, and number of extreme poor households. Robust standardization was used to address data variability and outliers. The analysis revealed two main clusters: one with relatively low poverty and unemployment rates but high stunting prevalence, and another with universally poor socio-economic conditions. Further sub-clustering highlighted internal heterogeneity, although with moderate silhouette values. The findings offer a nuanced spatial understanding of poverty distribution and provide a data-driven foundation for targeted poverty alleviation policies.

Keywords: poverty indicators, clustering, K-Medoids, silhouette coefficient, Majalengka, socio-economic analysis

1. Introduction

Poverty remains a major challenge in sustainable development. This is because poverty encompasses many dimensions, from low income to limited access to basic services, such as education, health and employment. Although in 2024, Indonesia is targeting a reduction in poverty rates below 7%, the latest data from the Central Statistics Agency shows that the national poverty rate is still 9.03%. In West Java Province, the poverty rate was recorded at 7.46%, while in Majalengka Regency it reached 10.82%, or around 134 thousand residents. This shows that this district is facing serious problems in poverty alleviation efforts.

The high poverty rate is also accompanied by disparities between regions within Majalengka Regency, which shows that each sub-district has different socio-economic conditions. For this reason, an analytical approach is needed that is able to group sub-districts based on similarities in poverty characteristics in order to understand poverty distribution patterns in a more focused manner.

This study uses the K-Medoids method to cluster 26 sub-districts in Majalengka Regency based on 5 (five) indicators, namely the number of poor people, stunting prevalence, school dropout rates, unemployment, and the number of extremely poor families. The K-Medoids method was chosen because it is able to produce representative clusters and is more resistant to outliers, making it suitable for heterogeneous socio-economic data. The purpose of this study was to determine the results of sub-district grouping based on poverty indicators using K-Medoids, identify the characteristics of each cluster formed, evaluate the quality of clustering through the Silhouette index, and explore the specific conditions of each cluster as a basis for recommendations for more targeted regional development policies.

2. Literature Review

Poverty is a condition of inability of individuals or groups to meet basic needs, both food and non-food, which is usually measured through an expenditure approach (Central Bureau of Statistics, 2023). Suharto (2005) emphasized that poverty is not only an economic issue, but also includes social aspects, such as limited access to power, information, and justice. Therefore, poverty alleviation strategies must be comprehensive and coordinated across sectors.

In spatial and socio-economic analysis, clustering techniques are often used to group units of analysis based on similar characteristics. According to Hair et al. (2010), cluster analysis aims to form groups with high internal homogeneity and maximum external heterogeneity. There are two main approaches, namely hierarchical and non-hierarchical methods. Non-hierarchical methods, such as K-Means and K-Medoids, require the number of clusters to be determined in advance and are more efficient for big data (Triaynto, 2015). The K-Medoids method is a non-hierarchical algorithm similar to K-Means, but uses real objects (medoids) as cluster centers so that it is more resistant to outliers and noise (Febrianti et al., 2019). This advantage makes K-Medoids ideal in socio-economic contexts that often contain outlier data.

Outliers are extreme values that can distort the results of distance-based analysis such as clustering. Outlier detection can be done using a boxplot that refers to the Interquartile Range limit (Triola, 2018). To reduce the influence of variables with large scales and outliers, the robust standardization technique is used, namely data transformation based on medians and IQRs that are not significantly affected by extreme values (Proclus Academy, 2022).

The validity of clustering results also depends on the fulfillment of basic assumptions. Hair et al. (2010) stated that the data must be representative, which is tested through Kaiser-Meyer-Olkin (KMO). In addition, the data should not have multicollinearity, which can be tested using the Variance Inflation Factor (VIF) (Gujarati, 2009). If this assumption is met, the clustering results will be more reliable. To handle multicollinearity and simplify high-dimensional data structures, the Principal Component Analysis (PCA) method is used. PCA changes correlated variables into a number of independent principal components, thereby increasing the accuracy of the clustering model (Rosyada & Utari, 2024).

In the implementation of the Partitioning Around Medoids (PAM) algorithm, clustering begins with random selection of initial medoids, distance calculations using Euclidean Distance, and iterative evaluation until the optimal medoids are obtained (Kaufman & Rousseeuw, 1987). This process is validated using the Silhouette Coefficient, which is an internal evaluation index that measures how well an object is in its cluster compared to other clusters (Dewi & Pramita, 2019). Index values between 0 and 1 indicate good cluster quality.

3. Methodology

This study uses secondary data covering five poverty indicators in 26 sub-districts in Majalengka Regency in 2024, namely the number of poor people (X1), stunting prevalence (X2), school dropout rates (X3), number of unemployed (X4), and number of extremely poor families (X5). Data were obtained from the Regional Development Planning Agency, Research and Development of Majalengka Regency, and from the Regional Poverty Alleviation Coordination Team.

Data analysis was carried out using the K-Medoids clustering approach and Silhouette Method validation, with the help of R-Studio software. The stages of analysis carried out include:

1. Collection of poverty indicator data,
2. Initial descriptive statistical analysis,
3. Outlier detection through boxplots,
4. Data standardization using the robust standardization method to equalize the scale between variables,
5. Sample representativeness assumption test using Kaiser-Meyer-Olkin (KMO),
6. Multicollinearity test between variables using the Variance Inflation Factor (VIF); if multicollinearity is detected, dimension reduction is performed using Principal Component Analysis (PCA),
7. Determination of the optimal number of clusters based on the Silhouette Coefficient,
8. Sub-district clustering using the K-Medoids algorithm,
9. Interpretation of cluster results based on the characteristics of each group.

This method allows the formation of more stable and robust clusters against outliers, and produces more representative regional segmentation in the context of poverty alleviation policies.

4. Results and Discussion

Descriptive analysis shows that poverty indicators in Majalengka Regency have high diversity between sub-districts, especially in the variable of the number of extremely poor households which has a coefficient of variation of 125.34%. Outlier detection through boxplots also revealed that most variables contain extreme values. Therefore, robust standardization was carried out to equalize the scale between variables.

The KMO test shows a value of 0.75, which indicates that the data is representative. All VIF values <10, indicating no multicollinearity, so dimension reduction through PCA is not required. Based on the Silhouette Coefficient, the optimal number of clusters is two, with an average silhouette value of 0.6386 indicating a fairly good cluster structure.

The K-Medoids method forms two clusters, namely:

- Cluster 1, consisting of 23 sub-districts. The characteristics of sub-districts in cluster 1 generally have relatively low levels of poverty, unemployment, and extremely poor households, but the prevalence of stunting is quite high.
- Cluster 2, consists of 3 sub-districts, namely Cingambul Sub-district, Lemahsugih Sub-district, and Malausma Sub-district. All three show worse socio-economic conditions overall.

Validation shows that Cluster 1 has a better silhouette value (0.66) than Cluster 2 (0.49). To deepen the analysis on Cluster 1 which has quite a lot of members, re-clustering was carried out on Cluster 1 using the iterative refinement approach.

The results of the re-clustering identified two new sub-clusters:

- Sub-cluster 1, consists of 8 sub-districts with the following characteristics: showing the best socio-economic conditions, with low values on all indicators.
- Subcluster 2, consisting of 15 sub-districts with the following characteristics: higher levels of poverty, unemployment, school dropouts, and stunting compared to subcluster 1.

The average value of the re-cluster silhouette decreased to 0.3181, indicating moderate to weak cluster separation quality. However, this step remains relevant because it reveals heterogeneity within the initial cluster.

Visualization using ArcGIS strengthens the analysis results. The map shows the spatial distribution of clusters, with Cluster 2 (high poverty) located in the southern part of Majalengka, marked in dark red. Meanwhile, the re-clustered subclusters are visualized with color gradations, reflecting differences in the level of socio-economic vulnerability between regions.

Fig. 1 – Districts Visualization using ArcGIS

5. Conclusion

This study successfully grouped 26 sub-districts in Majalengka Regency into two clusters based on poverty indicators using the K-Medoids method. Cluster 1 includes 23 sub-districts with lower levels of poverty, unemployment, and extremely poor household heads, but the prevalence of stunting is still high. Cluster 2 consists of 3 sub-districts with worse socio-economic conditions overall.

The validity of the clusters shows that Cluster 1 is quite good (silhouette 0.66), while Cluster 2 is weaker (0.48). Re-clustering of Cluster 1 revealed internal heterogeneity, resulting in two new sub-clusters, although with lower silhouette values (0.40 and 0.17), but still providing a deeper understanding of the socio-economic structure of the region to prioritize poverty management in Majalengka Regency.

References

Central Bureau of Statistics (Badan Pusat Statistik). (2023). Profil Kemiskinan di Indonesia. BPS RI.

Dewi, N. L. P. S., & Pramita, D. A. (2019). Evaluasi hasil klasterisasi menggunakan Silhouette Coefficient. Jurnal Ilmiah Teknologi dan Informasi, 6(2), 115–123.

Febrianti, E., Syafi'i, M. F., & Prasetyo, E. (2019). Perbandingan metode K-Means dan K-Medoids dalam pengelompokan data kemiskinan. Jurnal Sains dan Informatika, 5(1), 35–42.

Gujarati, D. N. (2009). Basic Econometrics (5th ed.). McGraw-Hill.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). Multivariate Data Analysis (7th ed.). Pearson Education.

Kaufman, L., & Rousseeuw, P. J. (1987). Clustering by means of medoids. North-Holland.

Proclus Academy. (2022). Robust data transformation in social research.

Rosyada, A., & Utari, D. A. (2024). Penerapan PCA pada analisis klaster sosial ekonomi. Jurnal Statistika Terapan, 9(1), 45–52.

Suharto, E. (2005). Analisis Kemiskinan dan Strategi Penanggulangannya. Rajawali Pers.

Triaynto, A. (2015). Metode Klasterisasi dalam Data Mining. Penerbit Informatika.

Triola, M. F. (2018). Elementary Statistics (13th ed.). Pearson.