CAUCHY: Jurnal Matematika Murni dan Aplikasi
Volume 10(1) (2025), Pages 72-81
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Integration of Dbscan Cluster Analysis with Multigroup Moderation Path Analysis

Hafizh Syihabuddin Al Jauhar*, Solimun, Rahma Fitriani

Brawijaya University, Indonesia

Email: aljauhar.hafizh@student.ub.ac.id

ABSTRACT

Path analysis is a development of regression analysis that can handle multiple structural equations at once. However, path analysis has limitations, especially when the sample being analysed has uniform characteristics, which can lead to less than optimal modelling results. Therefore, before conducting path analysis, a more in-depth preliminary analysis is required. One approach that can be used to improve this is to integrate cluster analysis with path analysis. This research examines the application of integration between DBSCAN cluster analysis and multigroup moderation path analysis to analyse patterns of waste management behaviour in Batu City. DBSCAN was used to cluster the data based on density, resulting in two main clusters and also some noise data. The first cluster consisted of 189 respondents, while the second cluster consisted of 196 respondents, with the remaining 10 data identified as noise. The DBSCAN clustering results show a silhouette index of 0.664, which indicates good clustering quality in terms of compactness and separation between clusters. After the data was clustered, each cluster was analysed using multigroup moderation path analysis to assess the relationship between environmental quality, understanding of 3R-based waste management, and economic usefulness of waste with facilities and infrastructure variables as moderators. The results showed that clusters with good quality facilities had a stronger understanding of 3R-based waste management and its economic benefits. This finding underscores the importance of facilities and infrastructure in influencing community waste management behaviour patterns.

Keywords: DBSCAN; path analysis; multigroup moderation; waste management; facilities and infrastructure; Batu City

INTRODUCTION

Path analysis is an extension of regression analysis designed to accommodate multiple structural equations. One key characteristic of structural equations in path analysis is the presence of at least one exogenous variable, one intervening endogenous variable, and one pure endogenous variableus variables influence other variables but are not affected themselves, whereas pure endogenous variables are influenced by other variables without affecting any others [1]. Intervening endogenous variables, on the other hand, both influence and are influenced by other variables.

Submitted: November 14, 2024 Reviewed: December 26, 2024 Accepted: January 16, 2025
DOI: http://dx.doi.org/10.18860/ca.v10i1.29847

Another important concept in path analysis is the moderating variable. Moderating variables either strengthen or weaken the relationship between exogenous and endogenous variables [2]. They work alongside exogenous variables to affect endogenous variables. There are two primary approaches to analyzing moderating variables: the moderation regression analysis approach (interaction method) and the multigroup approach [3].

Path as some limitations, especially when the sample displays grouped characteristics, which can result in suboptimal modeling outcomes [4]. To address this, it is beneficial to perform a more comprehensive preliminary analysis before applying path analysis. One such approach involves integrating cluster analysis with path analysis. This integration uses clustering results to divide the dataset into groups, after which each subset is analyzed using path analysis techniques [5].

Cluster analysis is a type of multivariate analysis under the interdependency method, where independent (explanatory) and dependent (response) variables are not distinguished [6]. Broadly, cluste methods are divided into hierarchical and non-hierarchical techniques. Hierarchical methods include Single Linkage, Average Linkage, Complete Linkage, Centroid Linkage, and Ward's method, while non-hierarchical methods are exemplified by the K-Means method [7].

A common issue in cluster analysis is that not all objects in a study can be easily classified into a particular cluster. Some objects, having unique characteristics that differ substantially from others, are identified as noise. Including such noise in clusters can distort character identification and reduce clustering accuracy [8].

Noise data can sometimevaluable information that the main data set does not offer, often due to unique combinations of circumstances that warrant further investigation. Algorithms that can handle noise effectively include Density-Based Spatial Clustering of Applications with Noise (DBSCAN), K-Medoids, Self-Organizing Maps, and Fuzzy C-Means. According to [8], the DBSCAN algorithm is particularly effective in detecting and managing noise while clustering data based on density. This non-parametric clustering method identifies and categorizes points that do not belong to any cluster as noise, grouping only points that exhibit high similarity.

Introduced by Ester, DBSCAN was awarded the Test of Time Award in 2014 by the Association for Computing Machinery (ACM) at a data mining conference [9]. DBSCAN does not require specifying the numsters beforehand, as it creates clusters based on data density rather than presuppositions (Mahesh & Rama, 2016). This feature is advantageous, as noise data differ from the overall dataset are excluded from clusters, thereby maintaining clustering integrity [11].

This study explores the integration of DBSCAN cluster analysis with multigroup moderation path analysis. This combined approach aims to yield novel insights and potentially different results by leveraging the strengths of both techniques.

METHODS

Data

This study uses secondary data taken from research conducted by [12]. The population in the study included the people of Batu City, totalling 217,871 families. The sample of this study consisted of people living in Batu sub-district and Bumiaji sub-district, with a total sample determined using qouta sampling of 395 respondents.

Table 1. Population and Sample Details for Each Village.

District	Village	Total Population	Sample
Batu sub-district	Ngaglik	12.698	31
	Oro Oro Ombo	10.717	26
	Pesanggrahan	13.780	34
	Sidomulyo	8.385	20
	Sisir	21.082	51
	Songgokerto	7.402	18
	Sumberejo	7.631	19
	Temas	17.911	44
	Total	**99.606**	**242**
Bumiaji sub-district	Bulukerto	6.708	16
	Bumiaji	7.317	18
	Giripurno	11.087	27
	Gunungsari	7.240	18
	Pandanrejo	6.123	15
	Punten	5.389	13
	Sumberbrantas	4.824	12
	Sumbergondo	4.263	10
	Tulungrejo	9.825	24
	Total	**62.776**	**153**
Grand Total		**162.382**	**395**

Research Model

The proposed research model is presented in Figure 1.

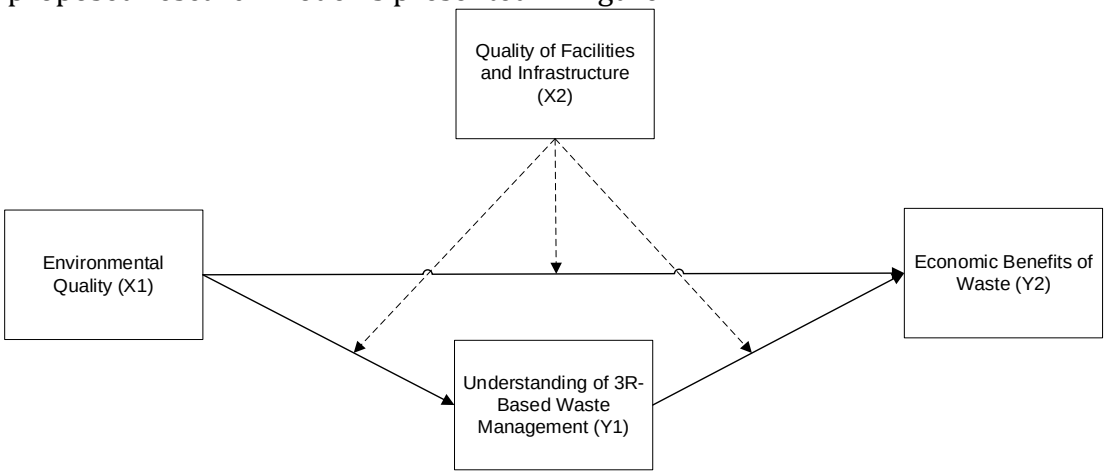

Figure 1. Research Model

DBSCAN Algorithm

The DBSCAN algorithm requires two important parameters, namely Eps (ε) and MinPts. Minimum points is the minimum number of objects included in a group [13]. while Epsilon is a distance that measures the closeness between an object and other objects around it [14]. The DBSCAN algorithm is as follows (Devi et al., 2015):

1. Randomly select a starting point p.
2. Initialise the input parameters MinPts and Eps.

3. Calculate Eps or all affordable density distances using euclidean distance

$$d_{ij} = \sqrt{\sum_{a}^{p} \left(x_{ia} - x_{ja}\right)^2} \quad (1)$$

Where x_{ia} is the a-th variable of object i $(i = 1, \ldots, n;\ a = 1, \ldots, p)$ and d_{ij} is the Euclidean Distance value.

4. If the number of points that satisfy Eps is more than MinPts then point p is the centre point (core) and the cluster is formed.
5. Repeat steps 3 - 4 until all points are processed.

According to [15] there are several advantages possessed by the DBSCAN algorithm, among others:

1. Recognises non-convex clusters.
2. Partition with the most appropriate number of clusters is obtained automatically.
3. There is no need to use an index to determine the appropriate number of clusters in a partition.

K-Nearest Neighbour Algorithm

There are two main parameters in clustering data using DBSCAN, namely epsilon and MinPts [16]. Determination of epsilon and MinPts can be done with the help of the K-Nearest Neighbour (KNN) algorithm. The KNN algorithm is an algorithm for classifying objects against learning data that is closest to the object [17]. To determine epsilon and MinPts can use the help of k-dist graph.

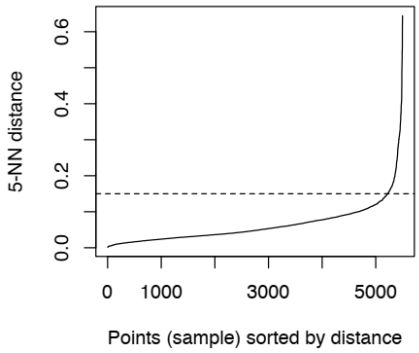

Figure 2. Example of k-dist Graph
Source: Kassambara (2017)

The illustration above is a k-dist graph with a k value of 5. The sharp (elbow-shaped) changes that occur in the graph will be considered as epsilon and the k value is used as MinPts. The vertical line on the graph is a threshold that is determined based on a significant increase in the overall k-dist graph [19]. The threshold line aims to prevent errors in determining the epsilon value where if the eps value is too large it results in outliers joining the group. The steps in determining epsilon and MinPts are as follows [20].

1. Calculate the k-dist value for all points at each k value. The k-dist value is obtained from the average value of the euclidean distance of an object with the k nearest points.

2. Sort the k-dist values in increasing order from least to greatest average.
3. Observe the point with the sharpest change on each k-dist graph, and mark it as the threshold.
4. The change in the k-dist value or the point forming the elbow will be used as the epsilon value and the k value as the corresponding MinPts.
5. If the value of k is too large, the groups formed will be smaller and many points will be identified as noise.

Multigroup Moderation Path Analysis

Moderating variables are variables that strengthen or weaken the effect of exogenous variables (predictor or independent) on endogenous variables (response or dependent) [3].

Graphically the path diagram for moderation of the multigroup method can be seen in Figure 2.

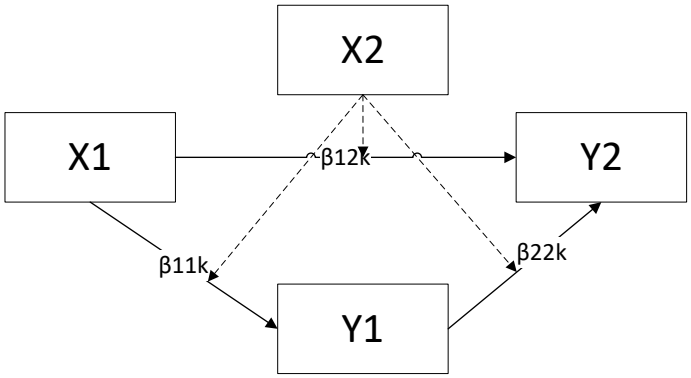

Figure 3. Path Diagram for Multigroup with Moderating Variables

Based on Figure 2, the equation that can be formed is

$$Z_{Y_{1ki}} = \beta_{11k} Z_{X_{1ki}} + \varepsilon_{1ki} \quad (2)$$
$$Z_{Y_{2ki}} = \beta_{12k} Z_{X_{1ki}} + \beta_{22k} Z_{Y_{1ki}} + \varepsilon_{2ki}$$

Group 1 ($k = 1$)

$$Z_{Y_{11i}} = \beta_{111} Z_{X_{11i}} + \varepsilon_{11i} \quad (3)$$
$$Z_{Y_{21i}} = \beta_{121} Z_{X_{11i}} + \beta_{221} Z_{Y_{11i}} + \varepsilon_{21i}$$

Group 2 ($k = 2$)

$$Z_{Y_{12i}} = \beta_{112} Z_{X_{12i}} + \varepsilon_{12i} \quad (4)$$
$$Z_{Y_{22i}} = \beta_{122} Z_{X_{12i}} + \beta_{222} Z_{Y_{12i}} + \varepsilon_{22i}$$

System Design

In the DBSCAN model, the number of clusters is not determined at the beginning; the number of clusters is known only after the Epsilon (ε) value and MinPts threshold are determined. The selection of these two parameters is very important as it directly affects the quality and number of clusters formed. After the DBSCAN cluster analysis is completed, the dataset is divided based on the clusters formed, and each cluster is then

further analysed using path analysis. The workflow of the integration algorithm between DBSCAN cluster analysis and multigroup moderation path analysis can be seen in Figure 3.

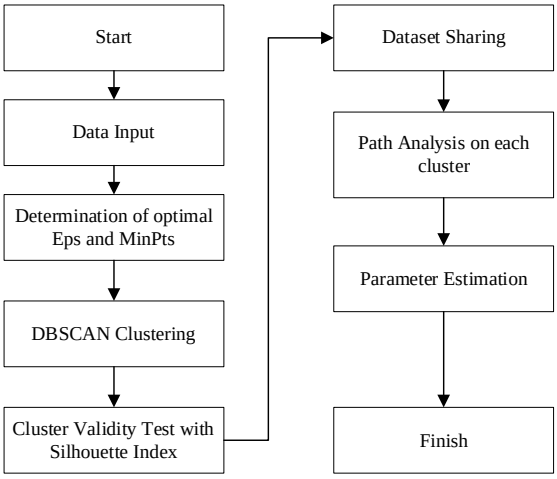

Figure 4. Algorithm Flow of Integration of DBSCAN Cluster Analysis and Multigroup Moderation Path Analysis

RESULTS AND DISCUSSION

DBSCAN Cluster Analysis

In this study to determine the optimal epsilon (ε) and MinPts values using the KNN method by calculating the distance in the point matrix to KNN. The value of k is shown in a rising curve where there is a sharp change in the curve is the optimal distance k as shown in Figure 4.

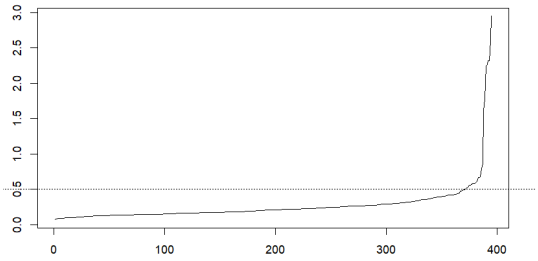

Figure 5. KNN Curve Determination of Optimal Eps Value

After obtaining the optimal epsilon (ε) value and MinPts with a value of ε = 0.5 and MinPts = 8 based on the highest Silhouette Coefficient value, the clustering process is then carried out with these two parameters as input. The clustering results are shown in Table 1.

Table 2. Cluster Analysis Results

MinPts	epsilon	Cluster	Many Members	Silhouette Index
8	0,5	Noise	10	0,664
		Cluster 1	189	
		Cluster 2	196	

Based on Table 1, the DBSCAN method with an epsilon value of 0.5 and MinPts value of 8 resulted in two clusters. The first cluster consists of 189 respondents, while the second cluster includes 196 respondents. In addition, there are 10 respondents who are categorised as noise. This noise refers to data that does not follow the main pattern in the

cluster and is therefore excluded to maintain the quality of the clustering. The presence of noise indicates a significant variation in the data distribution, which may result from outliers or pattern mismatch. The DBSCAN clustering results showed a silhouette index of 0.664, indicating a clustering quality with a good degree of compactness and separation.

Estimation of Path Analysis Parameters

Estimating the path coefficient in this study using the least squares method. The results of estimating the path coefficient in group 1 based on DBSCAN grouping can be seen in equation (5).

$$\hat{Z}_{Y_1} = 0{,}977 Z_{X_1} \qquad (5)$$

$$\hat{Z}_{Y_2} = 0{,}439 Z_{X_1} + 0{,}519 Z_{Y_1}$$

Based on Equation (5), it can be seen that the environmental quality variable has a positive relationship with the understanding of 3R-based waste management. This means that when environmental quality increases, understanding of 3R-based waste management will also be higher. In the second path model, environmental quality and understanding of 3R-based waste management also have a positive relationship with the economic usefulness of waste. In other words, improving environmental quality and understanding of 3R-based waste management will have an impact on increasing the economic usefulness of waste.

The estimation results of the group 2 path coefficient can be seen in equation (6).

$$\hat{Z}_{Y_1} = 0{,}988 Z_{X_1} \qquad (6)$$

$$\hat{Z}_{Y_2} = 0{,}614 Z_{X_1} + 0{,}349 Z_{Y_1}$$

Based on Equation (6), it is known that environmental quality has a positive relationship with the understanding of 3R-based waste management. With the increase in environmental quality, understanding of 3R-based waste management will also increase. In the second path model, it can be seen that environmental quality and understanding of 3R-based waste management are positively related to the economic usefulness of waste. This means that improving environmental quality and understanding of 3R-based waste management will have an impact on increasing the economic benefits of waste.

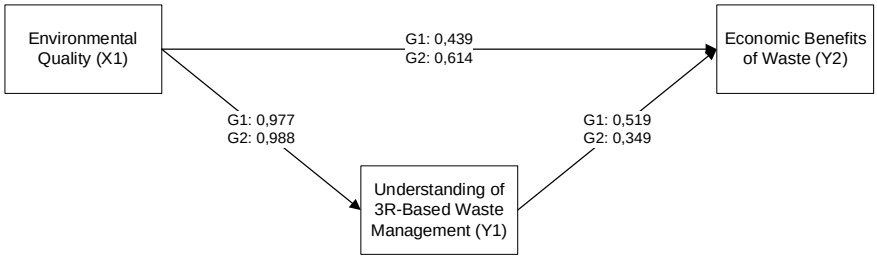

Figure 6. Group 1 and Group 2 Path Diagram based on DBSCAN clustering

Hypothesis Testing

Based on the structural equation model in Figure 5, hypothesis testing is carried out with the Jackknife resampling approach. The results of parameter significance testing based on DBSCAN clustering are presented in Table 2.

Table 3. Hypothesis Testing Results of Multigroup Direct Effect

Relationship	Group 1		Group 2		Difference Test (p-value)
	Coefficient	p-value	Coefficient	p-value	
X1 → Y1	0,974	<0,001	0,987	<0,001	<0,001
X1 → Y2	0,446	<0,001	0,623	<0,001	<0,001
Y1 → Y2	0,512	<0,001	0,339	<0,001	<0,001

Table 2 shows the results of the analysis of relationships between variables in two different groups, namely Group 1 (good facilities and infrastructure) and Group 2 (excellent facilities and infrastructure). In Group 1, the analysis results show that all relationships between variables have positive significance. In particular, Environmental Quality (X1) has a significant positive influence on the Understanding of 3R-Based Waste Management (Y1) and Economic Benefits from Waste (Y2). This finding indicates that the better the environmental quality, the higher the community's understanding of 3R-based waste management, which in turn increases the economic benefits that can be obtained from waste. This reinforces the importance of environmental quality as a supporting factor in increasing public awareness of sustainable waste management and the utilisation of waste for more optimal economic benefits.

In addition, understanding of 3R-Based Waste Management (Y1) is proven to have a significant positive influence on the Economic Benefits of Waste (Y2), which indicates that the higher the community's understanding of waste management, the greater the potential economic benefits that can be generated. This finding shows the important role of education and understanding in encouraging the creation of economic benefits from more effective waste management.

On the other hand, in Group 2, which has better facilities and infrastructure, the analysis results also show positive significance in all relationships between variables. In this group, Environmental Quality (X1) not only influences Understanding of 3R-Based Waste Management (Y1), but also has a positive influence on Economic Benefits from Waste (Y2). This finding further reinforces that good environmental conditions, supported by more optimised facilities and infrastructure, contribute significantly to increasing people's awareness and understanding of 3R-based waste management and its economic benefits.

In addition, the table also shows that the excellent facilities and infrastructure in Group 2 function as moderating factors that strengthen the three main relationships, namely between Environmental Quality (X1) and Understanding of 3R-Based Waste Management (Y1), Environmental Quality (X1) and Waste Economic Benefits (Y2), and Understanding of 3R-Based Waste Management (Y1) and Waste Economic Benefits (Y2). In other words, better quality of facilities and infrastructure strengthens the relationship between environmental factors and 3R-based waste management, as well as the relationship between understanding of waste management and economic benefits from waste, indicating the important role of facilities and infrastructure in supporting the success of sustainable waste management and its economic utilisation.

CONCLUSIONS

The results of the DBSCAN cluster analysis show that there are two main clusters in the analysed data, namely Cluster 1 with 189 members who have good facilities and infrastructure, and Cluster 2 with 196 members who have excellent facilities and infrastructure. The silhouette index value of 0.664 indicates that the separation between clusters is quite good, with 10 members clustered as noise. This separation indicates that the quality of facilities and infrastructure plays an important role in grouping communities based on their understanding and implementation of 3R-based waste management.

Furthermore, the results of the multigroup moderated path analysis provide a deeper insight into the differences between groups with good and inadequate facilities and infrastructure. This finding shows that groups with better facilities and infrastructure, such as those in Cluster 2, strengthen the relationship between Environmental Quality (X1) and Understanding of 3R-based Waste Management (Y1), as well as between Environmental Quality (X1) and Waste Economic Benefits (Y2). In this context, adequate facilities and infrastructure not only support increased public awareness of waste management, but also play an important role in encouraging the economic utilisation of waste.

Overall, these findings underscore the importance of good facilities and infrastructure in supporting more effective waste management behaviour in Batu city. Adequate infrastructure strengthens people's understanding of the 3Rs-based waste management concept and increases the potential economic benefits of waste management. This suggests that the development of better facilities and infrastructure can have a significant impact in improving sustainable waste management and encouraging the achievement of more optimal economic goals from waste.

REFERENCES

[1] A. A. R. Fernandes and Solimun, *Analisis Regresi dalam Pendekatan Fleksibel: Ilustrasi dengan Paket Program R*. UB Press, 2021.

[2] G. Brunet, A. Girona, G. Fajardo, and G. Ares, “Moderators of the Effect of Household Food Insecurity on Food Consumption Among Uruguayan Children and Adolescents,” *Sage Open*, vol. 14, no. 4, Oct. 2024, doi: 10.1177/21582440241281843.

[3] Solimun, A. A. E. Fernandes, and Nurjannah, *Metode Statistika Multivariat Pemodelan Persamaan Struktural (SEM) Pendekatan WarpPLS*. UB Press, 2017.

[4] A. Sadeghpour, V. D. Badal, D. L. Pogge, E. O’Donoghue, T. Bigdeli, and P. D. Harvey, “Using machine learning modeling to identify childhood abuse victims on the basis of personality inventory responses,” *J Psychiatr Res*, vol. 180, pp. 8–15, Dec. 2024, doi: 10.1016/j.jpsychires.2024.09.046.

[5] X. He and Y. Liu, “Knowledge evolutionary process of Artificial intelligence in E-commerce: Main path analysis and science mapping analysis,” *Expert Syst Appl*, vol. 238, p. 121801, Mar. 2024, doi: 10.1016/j.eswa.2023.121801.

[6] K. Mardani, K. Maghooli, and F. Farokhi, “Segmentation of coronary arteries from X-ray angiographic images using density based spatial clustering of applications with noise (DBSCAN),” *Biomed Signal Process Control*, vol. 101, p. 107175, Mar. 2025, doi: 10.1016/j.bspc.2024.107175.

[7] S. Wu, Y. Liu, J. Wang, and Q. Li, "Sentiment Analysis Method based on Kmeans and Online Transfer Learning," *Computers, Materials & Continua*, vol. 60, no. 3, pp. 1207–1222, 2019, doi: 10.32604/cmc.2019.05835.

[8] I. D. Id, A. Astrid, and E. Mahdiyah, "Modifikasi DBSCAN (Density-Based Spatial Clustering With Noise) pada Objek 3 Dimensi," *Jurnal Komputer Terapan*, vol. 3, no. 1, pp. 41–52, 2017, [Online]. Available: http://jurnal.pcr.ac.id

[9] J. Kim, H. Lee, and Y. M. Ko, "Constrained Density-Based Spatial Clustering of Applications with Noise (DBSCAN) using hyperparameter optimization," *Knowl Based Syst*, vol. 303, p. 112436, Nov. 2024, doi: 10.1016/j.knosys.2024.112436.

[10] K. Mahesh Kumar and A. Rama Mohan Reddy, "A fast DBSCAN clustering algorithm by accelerating neighbor searching using Groups method," *Pattern Recognit*, vol. 58, pp. 39–48, Oct. 2016, doi: 10.1016/j.patcog.2016.03.008.

[11] S. U. Rehman, S. Asghar, S. Fong, and S. Sarasvady, "DBSCAN: Past, present and future," in *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*, IEEE, Feb. 2014, pp. 232–238. doi: 10.1109/ICADIWT.2014.6814687.

[12] S. Solimun, A. A. R. Fernandes, I. Rahmawati, R. Isaskar, L. Muflikhah, and F. L. N. Rasyidah, "Cluster Integration Path Analysis to Model PT Pelindo II's Market Mapping," *International Journal of Circuits, Systems and Signal Processing*, vol. 15, pp. 1833–1841, Jan. 2022, doi: 10.46300/9106.2021.15.198.

[13] B. Hou, R. Ding, W. Shao, S. Liu, and L. Wang, "Pattern clustering method of magnetic near-field radiation emissions based on DBSCAN algorithm," *IET Science, Measurement & Technology*, vol. 18, no. 8, pp. 385–398, Oct. 2024, doi: 10.1049/smt2.12182.

[14] D. Cheng *et al.*, "GB-DBSCAN: A fast granular-ball based DBSCAN clustering algorithm," *Inf Sci (N Y)*, vol. 674, p. 120731, Jul. 2024, doi: 10.1016/j.ins.2024.120731.

[15] S. Scitovski, "A density-based clustering algorithm for earthquake zoning," *Comput Geosci*, vol. 110, pp. 90–95, Jan. 2018, doi: 10.1016/j.cageo.2017.08.014.

[16] L. Li, X. Chen, and C. Song, "A robust clustering method with noise identification based on directed K-nearest neighbor graph," *Neurocomputing*, vol. 508, pp. 19–35, Oct. 2022, doi: 10.1016/j.neucom.2022.08.029.

[17] X. Zhang, Y. Chen, J. Jia, K. Kuang, Y. Lan, and C. Wu, "Multi-view density-based field-road classification for agricultural machinery: DBSCAN and object detection," *Comput Electron Agric*, vol. 200, p. 107263, Sep. 2022, doi: 10.1016/j.compag.2022.107263.

[18] A. Kassambara, *Practical guide to cluster analysis in R: Unsupervised machine learning*, vol. 1. Sthda, 2017.

[19] X. Yuan, H. Yu, J. Liang, and B. Xu, "A novel density peaks clustering algorithm based on K nearest neighbors with adaptive merging strategy," *International Journal of Machine Learning and Cybernetics*, vol. 12, no. 10, pp. 2825–2841, Oct. 2021, doi: 10.1007/s13042-021-01369-7.

[20] P. Kaliyaperumal, S. Periyasamy, M. Thirumalaisamy, B. Balusamy, and F. Benedetto, "A Novel Hybrid Unsupervised Learning Approach for Enhanced Cybersecurity in the IoT," *Future Internet*, vol. 16, no. 7, p. 253, Jul. 2024, doi: 10.3390/fi16070253.