# Comparison of DBSCAN Algorithm and Resistivity Data Inversion, Case Study of Identification of Granite Distribution in 'Zs' Area

Muhammad Azis Albar[1], Y Yatini[1*]

[1]*Department of Geophysical Engineering, Faculty of Mineral Technology, Universitas Pembangunan Nasional Veteran Yogyakarta, SWK No. 104 North Ring*

[*]*Email: jeng_tini@upnyk.ac.id*

***Article Information:***

Received:
*6 March 2025*

Received in revised form:
*1 May 2025*

Accepted:
*20 May 2025*

Volume 7, Issue 1, June 2025
pp. 1 – 7

http://doi.org/10.23960/jesr.v7i1.168

***Abstract***

*The emergence of automatic data analysis techniques based on data mining algorithms can be applied in various fields including geophysics, thereby improving the quality of interpretation results. Wanner-Schlumberger configuration resistivity data were used in this study. Geoelectric data processing usually uses inversion methods, to determine the true resistivity distribution below the surface. This study proposes the use of the DBSCAN algorithm. Inversion comparison with linearization and clustering with the DBSCAN algorithm is carried out to identify granite dispersal in the 'ZS' area. The clustered cross-section has a clearer picture than the inversion cross-section. Based on the interpretation results on the resistivity cross-section that has an RMS error value of 0.68%, three ranges of resistivity values can be interpreted, namely low resistivity values < 100 ohms.m are indicated as soil, medium resistivity values of 100 – 500 ohms.m are indicated as granite gravel aquifers, and resistivity values > 500 ohms.m are indicated as pink basement granite. The cross-section of clustering results with input parameters eps 0.22 and mints 7 is interpreted. Namely, cluster 1 is an unsaturated granite gravel aquifer, cluster 2 is a saturated granite gravel aquifer, cluster 3 pink granite bedrock, cluster 4 soil and noise of 5.54%*

***Keywords:*** *Clustering, DBSCAN, Inversion, Resistivity, Granite.*

## I. INTRODUCTION

The resistivity method is one of the geophysical methods that is widely used in various fields, such as in the field of hydrogeology to find groundwater sources, the field of archeology to investigate parts of historical sites that are still buried and the mining field to find the presence of metallic and non-metallic minerals [1].

The emergence of automated data analysis techniques based on data mining algorithms can be applied in various fields including geophysics, thus improving interpretation for the better [2]. Clustering is a part of data mining that explores patterns in unlabeled data (unsupervised learning) and separates them into groups based on similarities [3]. The use of clustering algorithms is proven to improve the quality of processing and interpretation of geophysical data. Ward [4] conducted fuzzy clustering to determine lithological boundaries in resistivity data. Then, Sabor, et al. [2] used the DBSCAN algorithm to improve the interpretation of resistivity data.

This study aims to apply the clustering method using the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm to resistivity data. The main objectives include determining the optimal parameters for the DBSCAN algorithm in clustering resistivity data, and comparing the data distribution patterns for identifying granite rocks between DBSCAN clustering cross sections and conventional resistivity cross sections. Additionally, the study evaluates the effectiveness of using the DBSCAN clustering method on resistivity data.

## II. MATERIALS AND METHODS

Cluster analysis is one of the data mining techniques included in unsupervised learning, which aims to

identify a group of objects that have certain similar characteristics that can be separated from other groups of objects, so that objects in the same group are relatively more homogeneous than objects in different groups. The number of groups that can be identified depends on the amount and variety of object data. The purpose of grouping a set of object data into several groups that have certain characteristics and can be distinguished from one another is for further analysis and interpretation following the objectives of the research being carried out [5].

### *A. DBSCAN Algorithm*

DBSCAN (Density-Based Spatial Clustering of Application with Noise) is a density-based clustering algorithm. DBSCAN determines clusters based on connected regions with high density. The concept of clustering in DBSCAN is very simple [3]. First, DBSCAN looks for core objects, which are objects that have dense neighbors. Second, DBSCAN connects these core objects with their neighboring objects to form a dense region. The dense region is expressed as a cluster [6]. The shape of the cluster produced by DBSCAN depends on the density, so with this algorithm it is possible to produce arbitrary cluster shapes.

The DBSCAN algorithm uses two parameters that must be determined appropriately, namely: the minimum number of samples that become the density threshold to determine whether a region is dense or not, which is symbolized MinObj and the radius of neighborliness, which is symbolized ε. These parameters define the outlier or noise tolerance level of the algorithm [7]. DBSCAN is included in unsupervised clustering because the number of clusters generated is determined by the shape of the data distribution itself, without knowing the class label. DBSCAN can be implemented simply using the pseudocode below [3].

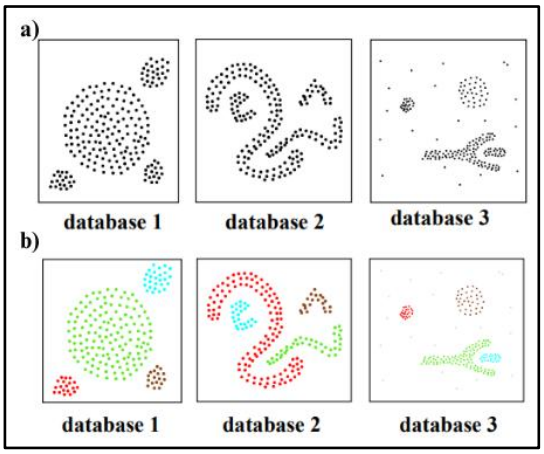


**Figure 1**. Clustering using the DBSCAN algorithm. a). Sample Databases. b). Clustering Results [8]

**Algorihm 1.** DBSCAN Algorithm Pseudocode

```
Cluster Set = DBSCAN (ε, MinObj)
Mark all object as unvisited
repeat
    Randomly select an object p from all objects labeled as unvisited
    Mark p as visited
    if within radius ε object p from all object labelled as univited then
        Create a new cluster C
        Add p to C
        Put all object that are neighbors of p into N
        for each object p' in N do
            if p' is labeled unvisited then
                Mark p' as visited
                if within radius ε object p' has at least MinObj object
                then add all object in radius ε to N
            if p' not a member of any cluster
            then add p to C
        end
    Output C as an output cluster
    else mark p as noise
until there are no objects labelled unvisited
```

### *B. Euclidean Distance*

Calculation of the distance between data when creating a cluster with the DBSCAN algorithm uses the Euclidean Distance function. Euclidean Distance is a distance measurement method that is widely used to measure the distance of two points in euclidean space. This method, commonly referred to as straight line distance, uses the formula [3].

$$d(x,y) = \sqrt{\sum_{t=1}^{n}(X_i - y_i)^2} \quad (1)$$

### *C. Silhouette Coefficient*

Silhouette Coefficient is a test model to determine how close the relationship between objects in the cluster and how far the cluster is separated from other clusters. To calculate the silhouette coefficient value, it is necessary to calculate the silhouette index value of the i-th data. The silhouette coefficient value is obtained by finding the maximum value of the global silhouette index value from the number of clusters two to the number of clusters n [9], as in the equation (2)

$$SI_i^j = \frac{b_i^j - a_i^j}{\max\{b_i^j, a_i^j\}} \quad (2)$$

### *D. Resistivity Method*

The resistivity method is one of the geophysical

methods used to describe the subsurface state with the resistivity distribution of the rock layer in the earth, where the earth is composed of rocks that have different resistivity values. The basis of the resistivity method is Ohm's Law, which is by conducting current into the earth through a current electrode and measuring its potential at the earth's surface using a potential electrode [10]. Mathematically, Ohm's Law can be written with the equation (3)

$$V = I.R \tag{3}$$

The resistivity method is one of the active geoelectric methods which uses an artificial electric current source injected into the subsurface. In this method, an electric current is injected into the earth layer through two current electrodes. So that the potential current price is known, the resistivity value can be determined [11]. The longer the current electrode distance will cause the electric current to penetrate deeper into the rock layer [1]. Resistivity geoelectric measurements can be used in various fields, such as in hydrogeology, mining, and geotechnical investigations [11].

### *E. Apparent Resistivity*

Geoelectric measurements aim to obtain resistivity values below the ground surface. In this resistivity method, it is assumed that the earth is isotropically homogeneous. With this assumption, the measured resistivity is true and does not depend on the electrode spacing. However, in reality, the earth is heterogeneous and composed of layers with different compositional and physical variations, so the measured potential is the influence of these layers. The resistivity of geoelectric measurements is the apparent resistivity (ρa) which depends on the electrode spacing [10]. The apparent resistivity value is formulated in the following equation:

$$\rho_a = K\frac{\Delta V}{I} \tag{4}$$

## III. RESULTS AND DISCUSSIONS

### *A. Application of DBSCAN on Synthetic Data*

The resistivity model is shown in Figure 2 (a). is data from forward modeling to describe the condition of the subsurface rock layer, wherein in the model, two different resistivity values are described by each color. At a depth of 0-25 meters is the first field, which has a resistivity value of 500 Ωm marked in blue, then the second field is at a depth of 25-40 meters, which has a resistivity value of 2500 Ωm marked in green, between the two layers there is an increase in the boundary field at a track length of 95 meters. The contrast of large resistivity values in the two layers in the resistivity model is made to provide clear results on the resistivity cross-section when the inversion process has been carried out.

Figure 2 (b). (c). is a comparison between the resistivity cross-section of the inversion results and the DBSCAN algorithm clustering cross-section. In Figure 2 (b) is a resistivity cross section that shows an RMS error value of 0.68% with a range of resistivity values from 440 - 2065 Ωm. Quantitatively, the cross section can be grouped into 3 ranges of resistivity values, namely low, medium and high. In areas with low resistivity values have a range of 440 - 684 Ωm marked in blue at a depth of 0 - 19.2 meters, and areas with medium resistivity values have a range of 854 - 1328 Ωm marked in green at a depth of 19.2 - 36.5 meters at a track length of 0 - 95 meters, then areas with high resistivity values have a range of 1656 - 2065 Ωm marked in red at a depth of 19.2 - 36.5 meters at a track length of 95 - 170 meters.

In Figure 2 (c). is a clustering cross-section consisting of two clusters accompanied by noise. The cross-section results from clustering resistivity data using the DBSCAN algorithm. The DBSCAN algorithm works by finding the nearest neighbor radius commonly called epsilon "ε" in resistivity data with the minimum number of neighbors or commonly called minPts as a condition for the formation of a cluster [12]. Based on the clustering results, cluster 1 is marked in blue, which is the first rock layer at a depth of 0 - 20 meters, then cluster 2 is marked in green at a depth of 20 - 35 meters, then based on the results of the clustering data it is also illustrated that there is an increase in the boundary layer field at a track length of 95 meters. In the clustering cross section, there is noise of 0.80%, which is outlier data from the inversion results that cannot be included in any cluster criteria. Clusters formed based on the DBSCAN algorithm can be validated by measuring the similarity or dissimilarity between data in one cluster and other clusters resulting from clusterization using the silhouette coefficient [13]. The clusterization cross section has a silhouette coefficient value of 0.307 which indicates that the data is included in the intermediate case.

Based on Figure 2, it can be seen that there is a difference in the shape of the pattern between the resistivity cross-section and the clustering cross-section. The difference in data distribution patterns occurs because, in Figure 2 (c), data clustering has been carried out so that data with relevant values are included in the same cluster. The clustering cross-section can more clearly show the boundary plane of the rock layer compared to the resistivity cross-section, usually marked by the presence of a layer increase plane at a

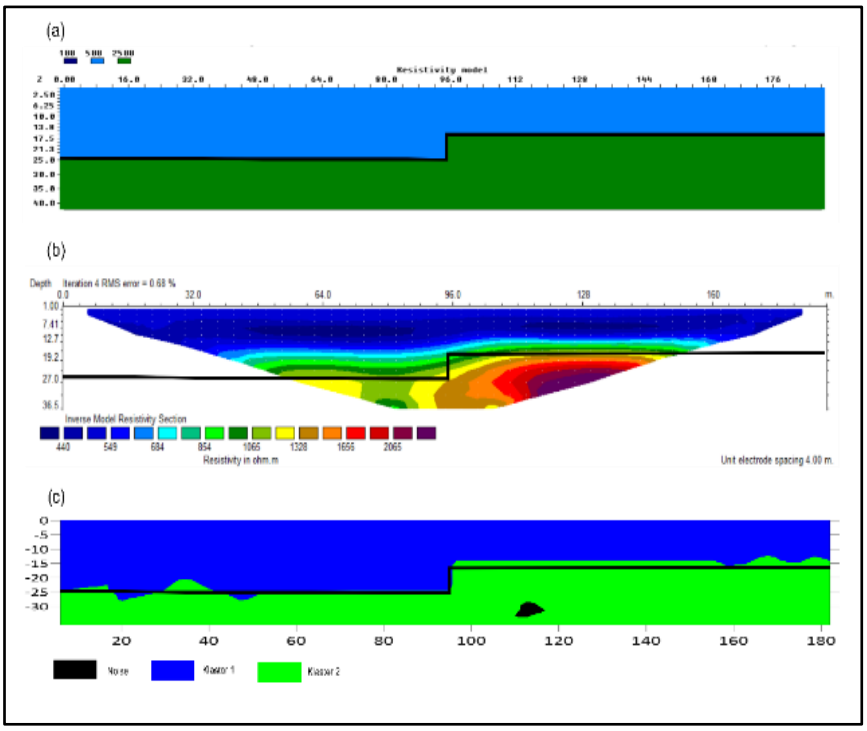


**Figure 2.** (a) Resistivity model used, (b) Resistivity cross section, (c) DBSCAN clustering cross section

track length of 95 meters, which is similar to the resistivity model used in this study.

### *B. Optimal Parameter Selection*

The DBSCAN algorithm is an algorithm that uses the epsilon parameter "ε" and minPts as input, based on these parameters the algorithm is run to perform clustering with the output in the form of the number of clusters and the presence of noise in the data. To get optimal results it is necessary to test with a combination of the two parameters, the following are the test results shown in Table 2.

**Table 2**. DBSCAN Algorithm Testing Result

| No | Eps | MinPts | Clusters Formed | Silhouette Index | Noise Percentage |
|---|---|---|---|---|---|
| 1 | 0.15 | 6 | 10 | 0.176 | 7.03% |
| 2 | 0.15 | 5 | 12 | 0.174 | 4.82% |
| 3 | 0.15 | 4 | 12 | 0.184 | 3.41% |
| 4 | 0.15 | 3 | 14 | 0.098 | 1.61% |
| 5 | 0.15 | 2 | 15 | 0.098 | 1.20% |
| 6 | 0.2 | 6 | 5 | 0.226 | 1.00% |
| 7 | 0.2 | 5 | 5 | 0.226 | 1.00% |
| 8 | 0.2 | 4 | 3 | 0.307 | 0.80% |
| 9 | 0.2 | 3 | 3 | 0.307 | 0.00% |
| 10 | 0.2 | 2 | 3 | 0.307 | 0.00% |
| 11 | 0.25 | 6 | 2 | 0.16 | 0.60% |
| 12 | 0.25 | 5 | 2 | 0.16 | 0.60% |
| 13 | 0.25 | 4 | 2 | 0.16 | 0.60% |
| 14 | 0.25 | 3 | 2 | 0.16 | 0.00% |
| 15 | 0.25 | 2 | 2 | 0.16 | 0.00% |

The best cluster results are obtained with a combination of eps input with a value of 0.2 and minPts of 4, based on these inputs resulting in a silhouette coefficient value of 0.307. These results indicate that the data is included in the intermediate case. This means that the similarity of the data with other data in the same cluster is relatively the same compared to the similarity of the data with data from other adjacent clusters [13].

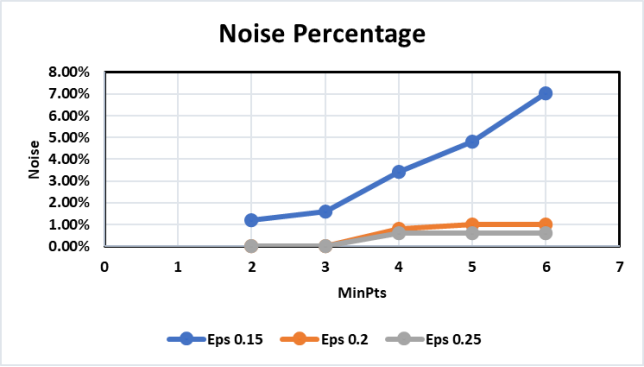


**Figure 2**. Graph of the Effect of Eps and MinPts Values on Noise Percentage

Noise or outliers are objects that have different characteristics compared to most objects in the dataset. One of the advantages of the DBSCAN algorithm is that it is able to properly detect noise because the concept of the DBSCAN algorithm is to cluster objects based on their density (density-based) with other objects, so it will ignore objects with characteristics that are not similar to the surrounding objects. Based on the

clustering results of resistivity data in Algorithm 1 before, the analysis of the effect of ε and minPts values on the amount of noise can be seen in Figure 3.

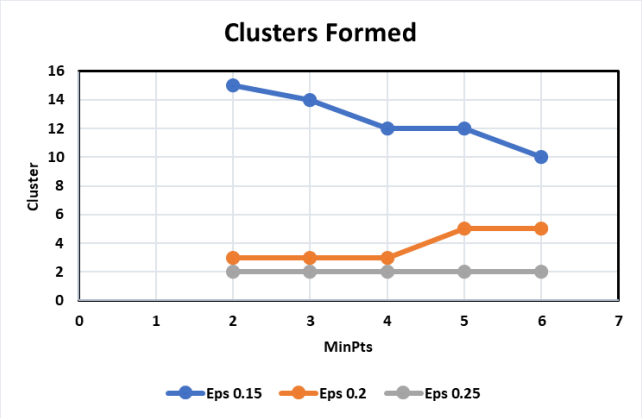


**Figure 3.** Graph of the Effect of Eps and MinPts Values on Clusters Formed

The clusters formed are influenced by the eps and minPts values entered. The larger eps value illustrates the wider range of density coverage of an object with other objects. Meanwhile, the greater the minPts value, the less likely the formation of a cluster. In the resistivity data, the cluster formed describes the distribution of rocks in the subsurface. Thus, the clusters formed can be used as a reference based on the geological conditions in the research area.

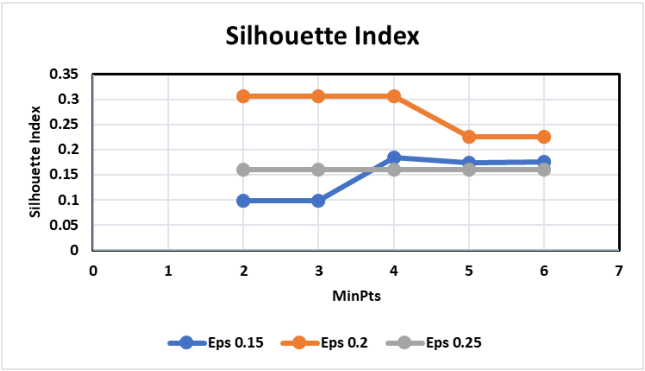


**Figure 4.** Graph of the Effect of Eps and MinPts Values on Silhouette Index

Based on the clustering results of the DBSCAN algorithm in Table 2, it can be seen that changes in the input values of eps and minPts greatly affect the quality of the clustering results. This is because the effect of the greater the value of eps, the wider the coverage of the density of a cluster. While the effect of the minPts value on the cluster results is that the greater the minPts value, the more difficult it will be for an object even though it is close to each other.

Figure 6 is a cross-section of the results of the combination of input parameters of the DBSCAN algorithm in accordance with Table 2 conducted in 6 experiments. Based on Figure 6 (a), we can see the test results of using the DBSCAN algorithm with a parameter value of epsilon = 0.15 and MinPts = 6, resulting in 8 clusters with 7.03% noise. Meanwhile, in Figure 6 (b), we can see the test results of using the DBSCAN algorithm with a parameter value of epsilon = 0.15 and MinPts = 5, resulting in 10 clusters with 4.82% noise. In Figure 6 (c), we can see the test results of using the DBSCAN algorithm with a parameter value of epsilon = 0.2 and MinPts = 5, resulting in 5 clusters with 1% noise. Figure 6 (d) shows the test results of using the DBSCAN algorithm with a parameter value of epsilon = 0.2 and MinPts = 4, resulting in 3 clusters accompanied by 0.80% noise. Then, Figure 6 (e) shows the test results of using the DBSCAN algorithm with a parameter value of epsilon = 0.25 and MinPts = 4, resulting in 2 clusters accompanied by 0.60% noise. Meanwhile, Figure 6 (f) shows the test results of using the DBSCAN algorithm with a parameter value of epsilon = 0.25 and MinPts = 3, resulting in only 2 clusters without noise

### *C. Application of DBSCAN oo Field Data*

Inversion is done to find the distribution of resistivity values in the research area. This study's interpretation of resistivity values is divided into 3 categories: low resistivity values, medium resistivity values, and high resistivity values. Figure 7 (a) is the result of the inversion cross-section of Track 1. Track 1 has a length of 180 meters with a maximum depth obtained of ± 50 meters. In the process of inversion of resistivity data, the data was iterated 4 times, and the final result had an error value of 7.7%. The resistivity cross-section can be divided into three, namely high, medium, and low resistivity. High resistivity has a value of more than 500 Ωm, medium value between 100 - 500 Ωm, and low value < 100 Ωm. High resistivity values with a value range of > 500 Ωm are interpreted as pink granite bedrock lithology, medium resistivity values with a value range of 100 - 500 Ωm are interpreted as granite gravel aquifer, low resistivity values with a value range of < 100 Ωm are interpreted as soil.

Figure 7 (b) is the result of clustering cross section using DBSCAN algorithm. In the clustering results, six clusters and noise were formed, and then the cross-section was regrouped into four clusters, and noise of 5.54% of the overall resistivity data used marked with each color; cluster 1 is marked with pink, which is interpreted as unsaturated granite gravel, cluster 2 is marked with light blue which is interpreted as saturated granite gravel, cluster 3 is marked with blue which is interpreted as pink granite bedrock, cluster 4 is marked with brown which is interpreted as soil and noise which is marked with black. The noise in the clustering cross-section is an outlier value resulting from the inversion

results that cannot be grouped in any of the clusters formed. This is an advantage of applying the DBSCAN algorithm because it can detect the presence of noise in the data used.

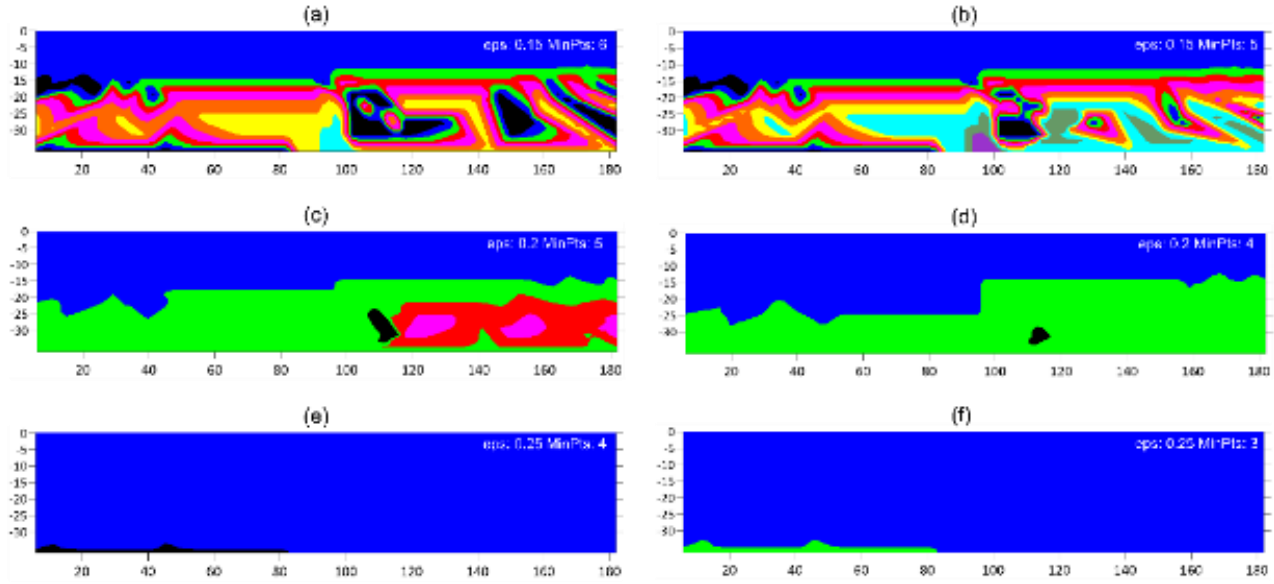


**Figure 5.** Testing Results of DBSCAN Algorithm in selecting Optimal Parameters. (a). eps = 0.15 minPts = 6, (b) eps = 0.15 minPts = 5, (c) eps = 0.2 minPts = 5, (d) eps = 0.2 minPts = 4, (e) eps = 0.25 minPts = 4, (f) eps = 0.25 minPts = 3

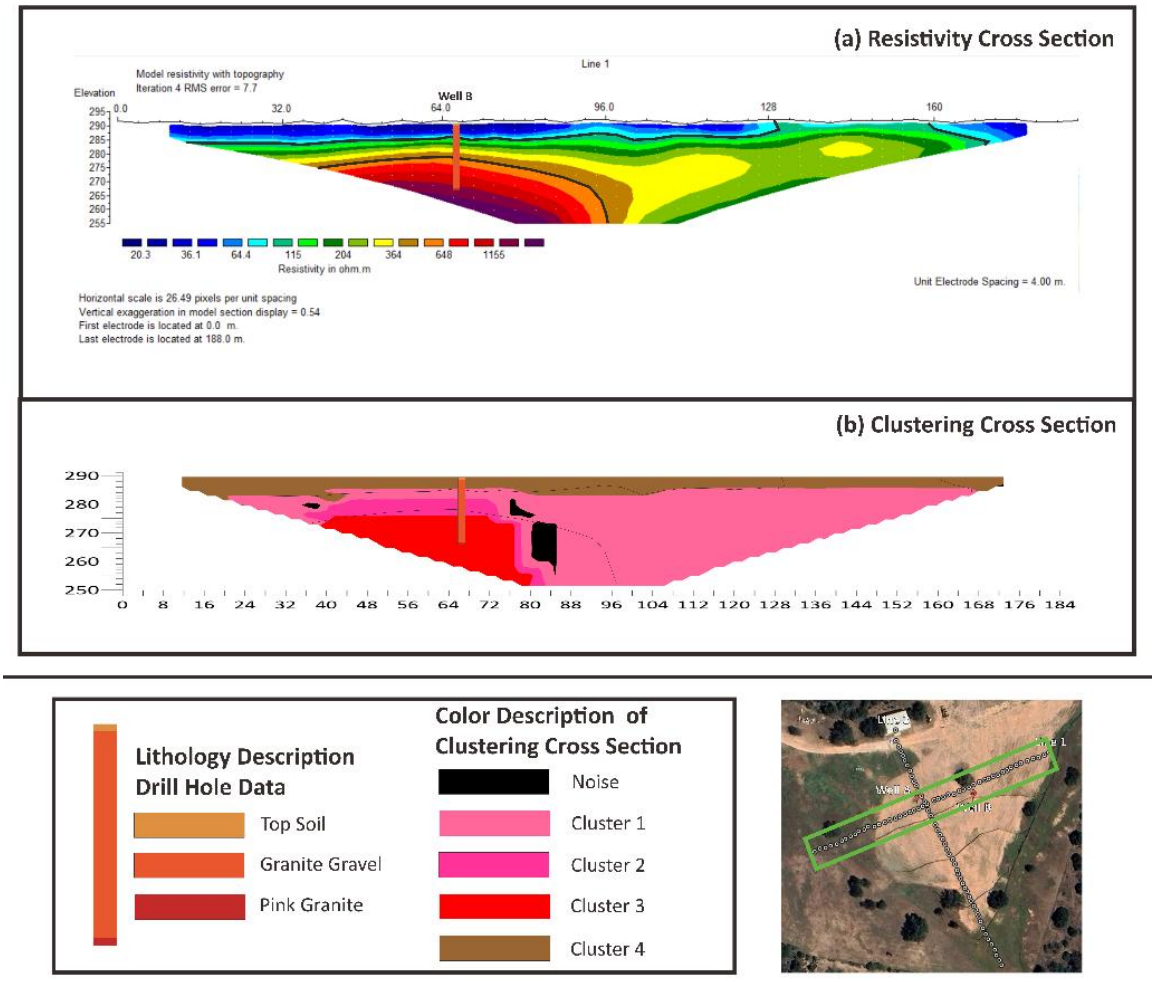


**Figure 6.** Cross Section Comparison (a) Inversion Result Resistivity Cross Section, (b) DBSCAN Clustering Cross Section

## IV. CONCLUSIONS

The optimal parameters are obtained by combining the input parameters when running the DBSCAN algorithm by validating using the silhouette coefficient. The cross-section of the corresponding cluster results shows that the use of the DBSCAN algorithm has been optimized. The application of the DBSCAN algorithm can provide more optimal results in determining the lithological boundaries in the subsurface based on the distribution of resistivity data by considering the optimal parameters used.

The inversion results produce an RMS error of 7.7%. Interpreting the inversion results on the resistivity cross-section, low resistivity values (<100) ohm.m as soil. Medium resistivity value (100-500) ohm.m is suspected as granite gravel aquifer, and high resistivity (> 500) ohm.m is indicated as pink granite bedrock. The clustering cross section shows cluster 1 as soil, cluster 2 as unsaturated granite gravel aquifer, cluster 3 as saturated granite gravel, cluster 4 as pink granite bedrock, and noise of 5.54%.

In showing the distribution of granite rocks in the subsurface, the clustering cross-section can show a better data distribution pattern than the resistivity cross-section. Data with relevant values are included in a cluster in the clustering cross-section. Relevant values are included in the same cluster so that one cluster and another cluster have a visible contrast in value difference.

## V. ACKNOWLEDGMENT

The authors are grateful to the Geophysical Engineering Department of UPN 'Veteran' Yogyakarta for providing opportunities and facilities for the research. The author would also like to thank the United States Geological Survey (USGS) for providing access to the open-source earth data provider website.

## VI. REFERENCES

[1] Reynolds, M. (2011). An Introduction to Applied and Environmental Geophysics. The University of Michigan.

[2] Sabor, K., Jougnot, D., Guerin, R., Steck, B. 2021. A data mining approach for improved interpretation of ERT inverted sections using the DBSCAN clustering algorithm. Geophys. J. Int. (2021) 225, 1304–1318.

[3] Suyanto. 2019. Data Mining Untuk Klasifikasi dan Klasterisasi Data Edisi Revisi. Bandung: Penerbit Informatika.

[4] Ward, W.O.C., Wilkinson, P.B., Chambers, J.E., Oxby, L.S. & Bai, L., 2014. Distribution-based fuzzy clustering of electrical resistivity tomography images for interface detection. Geophys. J. Int., 197(1), 310–321.

[5] Nofriansyah, D., & Nurcahyo, G. W. 2015. Algoritma Data Mining dan Pengujian. Sleman: Penerbit Deepublish.

[6] Kriegel, H.-P., Kroger, P., Sander, J., & Zimek, A. 2011. Density-Based Clustering. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discover 1(3): p.231-240.

[7] Pietrzykowski, Marcin. 2020. Comparison of mini-models Based on Various Clustering Algorithms. Procedia Computer Science 176 (2020) 3563–3570.

[8] Ester, M., Kriegel, H., Xu, X. & Sander, J., 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Assoc. Adv. Artif. Intell., 96(34), 226–231.

[9] Dewi, D. A. I. C., & Pramita, D. A. K. (2019). Analisis Perbandingan Metode Elbow dan Silhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali. Matrix: Jurnal Manajemen Teknologi Dan Informatika, 9(3), 102–109.

[10] Telford, e. a. (1990). Applied Geophysics Second Edition. New York: Cambridge University Press.

[11] Loke, M. (1999). Electrical Imaging Surveys for Environtmental and Engineering Studies.

[12] Adha, Rimelda, Nurhaliza, N., Soleha, U., Mustakin. 2021. Perbandingan Algoritma DBSCAN dan K-Means Clustering untuk Pengelompokan Kasus Covid-19 di Dunia. Jurnal Sains, Teknologi dan Industri, Vol. 18, No. 2, Juni 2021, pp.206 – 211.

[13] Devi, F., Gunawan, W., Kurniaputra. R. A., 2021. Implementasi Algoritma DBSCAN Dalam Pengambilan Data Menggunakan Scatterplot. Jurnal Ilmu Komputer dan Teknologi Informasi. Vol. 6 No: 2.

[14] Grandis, Hendra. (2009). Pengantar Inversi Geofisika. Himpunan Ahli Geofisika Indonesia (HAGI)

[15] Lowrie, William. (2007). Fundamental of Geophysics. New York: Cambridge University.

[16] Menke, William. (1989). Geophysical Data Analysis: Discrete Inverse Theory. Academic Press.