Comparison of DBSCAN and PCA-DBSCAN Algorithm for Grouping Earthquake Area

1st Mustakim
Departement of Information System
Puzzle Research Data Technology (Predatech)
Universitas Islam Negeri Sultan Syarif Kasim Riau
Pekanbaru – Riau, Indonesia
mustakim@uin-suska.ac.id

2nd Emi Rahmi
Departement of Information System
Puzzle Research Data Technology (Predatech)
Universitas Islam Negeri Sultan Syarif Kasim Riau
Pekanbaru – Riau, Indonesia
emi.rahmi@students.uin-suska.ac.id

3rd Mediantiwi Rahmawita Mundzir
Departement of Information System
Puzzle Research Data Technology (Predatech)
Universitas Islam Negeri Sultan Syarif Kasim Riau
Pekanbaru – Riau, Indonesia
mediantiwi.rahmawita@uin-suska.ac.id

4th Said Thaufik Rizaldi
Departement of Information System
Puzzle Research Data Technology (Predatech)
Universitas Islam Negeri Sultan Syarif Kasim Riau
Pekanbaru – Riau, Indonesia
11753101376@students.uin-suska.ac.id

5th Okfalisa
Departement of Informatic Engineering
Universitas Islam Negeri Sultan Syarif Kasim Riau
Pekanbaru – Riau, Indonesia
okfalisa@uin-suska.ac.id

6th Idria Maita
Departement of Information System
Universitas Islam Negeri Sultan Syarif Kasim Riau
Pekanbaru – Riau, Indonesia
idria@uin-suska.ac.id

2021 International Congress of Advanced Technology and Engineering (ICOTEN) | 978-1-6654-1224-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICOTEN52080.2021.9493497

***Abstract*—Geologically, the territory of Indonesia is where the three active tectonic plates meet which are always moving and colliding with each other, resulting in earthquakes, volcanic pathways, and faults. Earthquake is a natural disaster that cannot be avoided or prevented, but the consequences of earthquakes can be minimized. Based on data obtained from Meteorology, Climatology and Geophysics Agency (MCGA), earthquakes often occur in Indonesia. Data obtained from earthquakes can be grouped to map the area of earthquake occurrence and an analysis will be carried out to determine the characteristics of earthquake clustering areas. The clustering in this is study conducted with two experiments, first experiment is Density-Based Spatial Clustering of Applications with Noise (DBSCAN) without dimensional reduction and second experiment is DBSCAN clustering with dimensional reduction using Principal Component Analysis (PCA). The best cluster results can be found by calculating the value of Silhouette Index (SI) of each cluster. From the two experiments, the highest SI value was obtained in experiment using PCA, which was 0.4137. Then the second experiment was used as the best cluster results with the highest Dept and Magnitude features in clusters 19 and 17 which showed the 5 main regions where earthquakes often occur are Sumatra, Banda Sea, Moluccan Sea, Irian Jaya and Sulawesi**

Keywords*— Climatology and Geophysics Agency, DBSCAN, DBSCAN-PCA, Earthquake Area, PCA

I. Introduction

Geologically, the territory of Indonesia is at the meeting point of three active tectonic plates, namely the Indo-Australian Plate in the south, the Eurasian Plate in the north and the Pacific Plate in the East. The three plates move and collide with each other so that the Indo-Australian Plate dips beneath the Eurasian plate and causes earthquakes, volcanic pathways, and faults [1]. The Meteorology, Climatology and Geophysics Agency (BMKG) is carrying out government duties in the fields of Meteorology, Climatology, Air Quality and Geophysics according to the law.

Based on data obtained from BMKG, the activity of earthquakes in Indonesia is very high, on average, 400 times earthquakes every month were recorded. In 1991 to 2007, there were 24 major earthquakes, including in Aceh on December 26th, 2004 earthquake with 9.3 RS power. This earthquake was followed by a large tsunami which caused the loss of hundred thousand lives and caused the loss of trillions rupiah in assets, as well as the Yogyakarta earthquake on May 26th, 2006 which caused severe infrastructure damage. The Padang earthquake on September 30 2009 with 7.9 on the Richter Scale (RS) reached 4.8 trillion rupiahs loss, with 1,195 people killed, 271,540 units damaged. An earthquake with a tsunami in Aceh 2004 claimed nearly 300,000 lives in Indonesia, Thailand, India, Sri Lanka, Maldives and Africa [2].

Earthquake is natural disaster that cannot be avoided or prevented, but the consequences of earthquakes can be minimized. The data obtained from the earthquake event can be grouped to find out the spread area of the earthquake and to map the area. By knowing the spread area of the earthquake, people who lives in the area can build earthquake resistant buildings to minimize the losses that can be caused by earthquakes. This grouping was done using clustering techniques in Data Science field.

Clustering technique is the process of grouping a set of data objects into several groups or clusters so that objects in a cluster have a high similarity, but will be different from objects in other clusters [3]. The clustering technique used is Density-Based Spatial Clustering of Applications with Noise (DBSCAN). DBSCAN is a partition-based cluster type where denser regions are considered clusters and areas with low density are called noise [4]. C. Kondal Raj conducted a study which compared K-Means algorithm, K-Medoids, and

978-1-6654-1224-7/21/$31.00 ©2021 IEEE

DBSCAN with the title Comparison of K-Means, K-Medoids, DBSCAN Algorithms Using DNA Microarray Datasets, stated that the efficiency of DBSCAN algorithm is better than K-Means and K-Medoids, and it is one of the technique to find clusters in arbitrary form in large spatial databases [5]. Some advantages of DBSCAN algorithm are able to recognize non-convex clusters, the partitions with the most appropriate clusters are obtained automatically, and do not need to use an index to determine the appropriate number of clusters in a partition [6]. The advantage of DBSCAN is that it can identify clusters arbitrarily, able to overcome noise and outliers, and do not require the specification of the expected number of clusters in the data [7][8][9].

The disadvantage of data mining clustering techniques is with the data that has a very large feature [10], it raises some noises and problem in algorithmic processing time [11]. It was revealed by several previous researchers that the most optimal technique in performing data reduction is by applying Principal Component Analysis (PCA). PCA is a statistical procedure used to simplify data [12], thus forming a new coordinate system with maximum variance and used for grouping data based on data similarity [13]. Aside from being a dimensional reduction algorithm, PCA is able to overcome the initial centroid problems of K-Means clustering by applying eigen vector as the initial centroid determinant in the cluster [14]. So it is expected that the presence of PCA will be able to provide strong support to improve cluster validity in clustering algorithms such as DBSCAN.

In this study, two experiments will be carried out, the first cluster was carried out without doing dimension reduction and the second cluster was done after making dimensional reduction using PCA. Both experiments were conducted to find the best cluster results by comparing the value of the Silhoutte Index (SI) of each experiment. The best experiments based on SI will be applied to an earthquake grouping information system in Indonesia by BMKG.

II. Material and Method

Research that carried out by conducting field observations, observations and interviews with certain parties as supporters, collecting data as the main ingredient of the research process. Data that has passed preprocessing will be conducted with two models. The first is to do data clustering using PCA and secondly to cluster without PCA reduction. It was done to prove whether the role of PCA can work optimally as a dimensional reduction method and produce good cluster validity or vice versa. This experiment will be used as a standard reference in building information systems for BMKG. Figure 1 below is a methodology carried out in research in general:

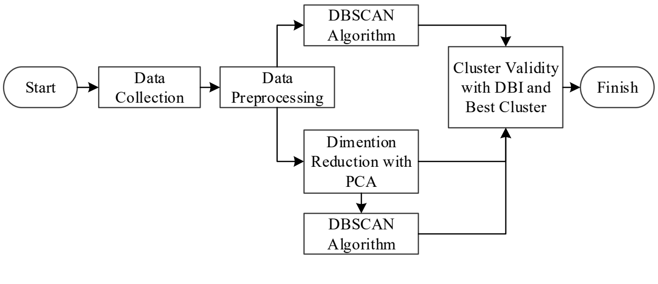

Fig. 1. Research Methodology

A. Data Mining and PCA

1) Data mining is the process of analyzing data in various variables and the end result is useful information [15]. Data sources can include databases, data warehouses, webs, other information repositories, or data that is dynamically transmitted to the system [3]. Preprocessing is known in data mining. This data is generally in the form of noise, large size, and from various sources [16]. The purpose of preprocessing is to guarantee the quality of data used [17] as well as some standard rules such as cleaning, transforming and normalization [18]. Principal Component Analysis (PCA) is used to reduce the dimensions of large size data [12]. PCA was developed by Pearson in 1901 and *developed* independently by Hotelling in 1933 and Jolliffe in 1986 [19]. The purpose of PCA is to reduce variables without removing information [13]. In general, the steps of the PCA algorithm are as follows: (1) Normalization of data by subtracting each data with Mean, (2) Calculate the covariance matrix, (3) Calculate eigen vector and eigen value, (4) Select the component and feature vector and (5) New data set [19].

B. Dencity-Based Spatial Clustering of Applications with Noise (DBSCAN)

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was introduced by Esther, this algorithm is a non-parametric clustering technique [20]. Data can be grouped in many ways depending on the grouping algorithm used [8]. DBSCAN algorithm is a method of large-scale data grouping [21]. DBSCAN gets clusters by finding the number of points in a specified distance from a particular point in the dataset and it can find data in the form of random and eliminate data noise [22][23][24].

III. Result and Analysis

A. Data Collection and Preprocessing

Data collection is a stage to get the data used in conducting research. Data collection was done by accessing the official website of Meteorology, Climatology and Geophysics Agency (BMKG) online. Data collected is earthquake data from 2012 to 2016 with 26,120 data. The data obtained was cleaned by cleaning process. Cleaning in this study is deleting unused data. The next step is to transform 8 attributes consisting of Day, Month, Year, Lat, Long, Dept, Mag and Region. The date attribute on the dataset used is separated into three attributes, day, month, and year. The next step is normalization process, which was carried out to equalize the range of values from all features used so that features with a large range of values do not dominate the results of calculating the distance between data points. The results of data normalization can be seen in Table 1 below:

TABLE I. Data Normalization

No	Region	Day	Month	Year	Lat	Long	Depth	Mag
1	Sumbawa Region	0	0,00	0,00	0,17	0,52	0,01	0,48
2	Northern Sumatra	0	0,00	0,00	0,83	0,10	0,01	0,25
3	Seram	0	0,00	0,00	0,41	0,74	0,07	0,41
4	Java	0	0,00	0,00	0,23	0,28	0,05	0,28

No	Region	Day	Month	Year	Lat	Long	Depth	Mag
5	Banda Sea	0	0,00	0,00	0,38	0,74	0,26	0,34
...	...	...	...	...	...	...	...	...
26.120	Northern Sumatra	1	1,00	1,00	0,72	0,09	0,09	0,39

B. *DBSCAN without PCA Dimension Reduction*

DBSCAN is a clustering algorithm based on density or data density. DBSCAN has the best ability to detect clusters of various shapes and sizes [16]. Earthquake data is clustered using DBSCAN algorithm with several experiments which combined Eps and MinPts parameters to find the best cluster results. To determine the quality of cluster results in this study by using silhouette index. In this study two experiments will be conducted, the first cluster is carried out without reducing the dimensions and the second cluster is carried out after applying dimensional reduction.

Data that has been normalized using min-max normalization will be grouped using DBSCAN algorithm. In this study, the euclidean distance was used to calculate the distance of each data. The application of DBSCAN in this study used RapidMiner tool. The results of DBSCAN cluster without dimension reduction with several combinations of Eps and MinPts can be seen in Table 2.

TABLE II. RESULT OF DBSCAN CLUSTER WITHOUT PCA DIMENSION REDUCTION

No	Eps	MinPts	Silhouette Index	Number of Cluster	Noise
1	1,7	5	-0,0871	227	20382
2	1,7	6	-0,0186	141	21353
3	1,7	7	0,0423	99	22001
4	1,7	8	0,0826	71	22497
5	1,7	9	0,1225	120	22803
6	1,7	10	0,1507	54	23039
7	1,8	5	0,0818	5	964
8	1,8	6	0,1165	4	1271
9	1,8	7	0,1075	3	1576
10	1,8	8	0,0409	4	1916
11	1,8	9	0,0937	6	2258
12	1,8	10	0,0535	4	2629

From Table 2 above, it can be seen that the 12 experiments were conducted with epsilon 1.7 and 1.8 and MinPts values between 5 and 10. The results obtained from the above experiments, the 10th experiment with the number of clusters 54 was the best experiment. The cluster results that have the highest silhouette index value was obtained by a combination of Eps = 1.7 and minPts = 10 which results in a silhouette index value of 0.1507.

C. *DBSCAN with PCA Dimension Reduction*

The dimension reduction method used in this study is Principal Component Analysis (PCA). PCA dimension reduction in this study used RapidMiner tools. The results of dimensional reductionsing PCA changed the 11 main attributes to 6 attributes according to PCA principle which can reduce data. This process will not reduce the quality of data that will be used in data mining process. The dataset that has been reduced by PCA can be seen in Table 3 below.

TABLE III. RESULT OF DBSCAN CLUSTER WITHOUT PCA DIMENSION REDUCTION

No	*Region*	Pc_1	Pc_2	Pc_3	Pc_4	Pc_5	Pc_6
1	Sumbawa Region	-0,55	0,43	0,52	0,19	0,09	0,09
2	Northern Sumatra	-0,54	0,40	0,51	-0,39	-0,44	-0,08
3	Seram, Indonesia	-0,52	0,41	0,50	-0,08	0,27	0,02
4	Java, Indonesia	-0,56	0,44	0,53	0,17	-0,17	-0,04
5	Banda Sea	-0,52	0,41	0,50	-0,04	0,28	0,06
6	Minahassa Peninsula, Sulawesi	-0,52	0,40	0,50	-0,27	0,06	-0,08
7	Minahassa Peninsula, Sulawesi	-0,52	0,40	0,50	-0,25	0,06	-0,19
8	Sumbawa Region	-0,55	0,44	0,53	0,25	0,04	-0,13
9	North of Halmahera	-0,50	0,37	0,49	-0,53	0,17	0,06
10	Minahassa Peninsula, Sulawesi	-0,52	0,40	0,50	-0,28	0,09	-0,12
...	...	...	...	...	...	...	...
26.120	Northern Sumatra	0,54	-0,42	-0,53	-0,12	-0,52	-0,07

The experiments carried out on this data applied Eps 0.15, 0.16 and 0.17 and MinPts from 5 to 10 with number of experiment is 18. It can be seen that the 18 experiments have various SI values and were not too different between experiments. The cluster result with the highest silhouette index was obtained by a combination of Eps = 0.16 and minPts = 5 and with silhouette index value of 0.4137. Based on the two experiments, it can be seen that the best cluster is obtained by applying PCA reduction dimensions on the data, because the value of silhouette index is higher. Comparison of silhouette index and noise values in both experiments can be seen in Figure 2 and Figure 3.

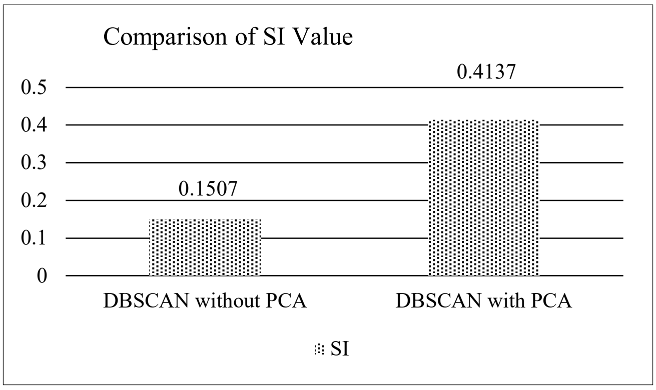

Fig. 2. Comparison of Silhouette Index

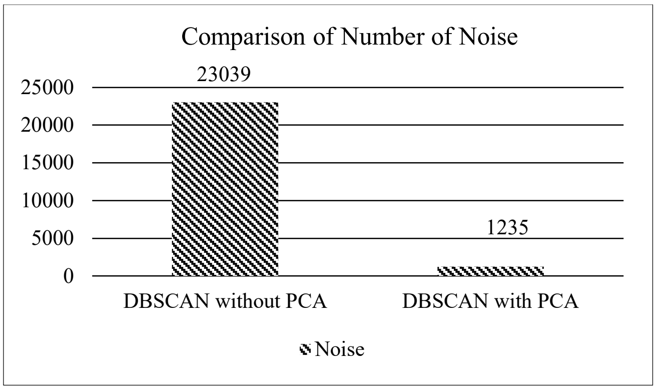

Fig. 3. Comparison of Noise Values

D. Analysis of Cluster Result

Based on cluster results, the areas that most frequent earthquakes are Minahassa Peninsula and Northern Sumatra. This is caused by the western part of Indonesia precisely in Sumatra region located in the meeting points of Indo-Australian Plate and Eurasia. The second meeting point of the plate was said to be the path of Mediterranean/Trans-Asiatic. Whereas in Minahassa Peninsula there are several active volcanoes and faults that become earthquake-prone areas. Based on result, it is known that the highest average depth is in cluster 19, which is 572.75 Km and the lowest average value of depth is in cluster 14, which is 10.4 Km. While comparison of the average magnitude of each cluster is known that the highest average magnitude value is in cluster 17, which is 4.89 SR and the lowest average magnitude value is in cluster 9, which is 3.01 SR.

E. Mapping of Cluster Result

Mapping the area of earthquake in Indonesia is mapped according to the best cluster results, the map can be seen in Figure 4.

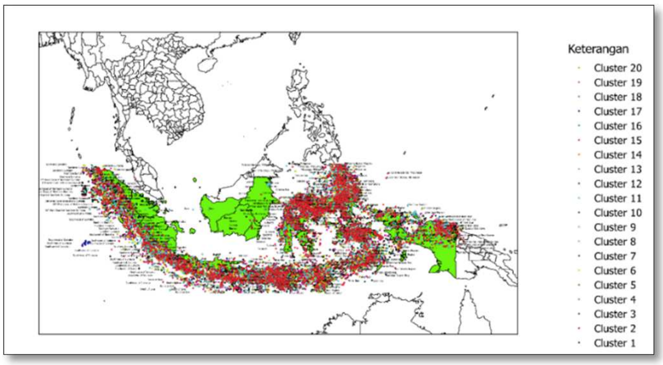

Fig. 4. Maping Eartquake Area of Indonesia

From the mapping result, it can be seen that the occurrence of earthquakes in Indonesia is more dominant in the ocean area. On the map it can also be seen that the earthquake occurence on Borneo Islan is not as many as an earthquake point on another island. It was influenced by the activity of Indo-Australian plate which moves beneath the Eurasian plate, as well as the Pacific plate which moves westward.

IV. Conclusion

The best cluster results were obtained after dimension reduction using PCA on the data before grouping. The highest value of silhouette index obtained was 0.4137 with the number of clusters formed were 20 clusters. So that dimensional reduction data was used as a reference to be applied to earthquake grouping information system. Earthquakes most often occur in Sumatra region, Banda Sea, Moluccan Sea, Irian Jaya and Sulawesi because the area is the meeting point of world's active tectonic plates. Based on the results of clustering, it was known and validated that the highest Dept and Magnitude were in clusters 19 and 17 with 572.75 km and 4.89 SR respectively. The cluster was occupied by 5 regions above with the intensity of earthquake strength above average, the results of analysis from this research come from the dataset used. The weakness of this research is that the clustering process will be repeated when there are several simultan changes in the dataset, which affect the cluster groups formed.

References

[1] M. Fuady, R. Munadi, and M. A. K. Fuady, ‘Disaster mitigation in Indonesia: between plans and reality’, *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1087, no. 1, p. 12011, 2021.

[2] E. Kertapati, ‘Aktivitas Gempa Bumi di Indonesia Perspektif Regional Pada Karakteristik Gempa Bumi Merusak’, *Pus. Survei Geol. Bandung*, vol. 109, 2006.

[3] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Elsevier, 2011.

[4] P. Batra Nagpal and P. Ahlawat Mann, ‘Comparative Study of Density based Clustering Algorithms’, *Int. J. Comput. Appl.*, vol. 27, no. 11, pp. 44–47, Aug. 2011.

[5] C. K. Raj, ‘Comparison of K Means, K Medoids, DBSCAN Algorithms Using DNA Microarray Dataset’, *Int. J. Comput. Appl. Math.*, vol. 12, no. 1, pp. 344–355, 2017.

[6] S. Scitovski, ‘A density-based clustering algorithm for earthquake zoning’, *Comput. Geosci.*, vol. 110, pp. 90–95, 2018.

[7] S. K. Bandyopadhyay and T. U. Paul, ‘Segmentation of brain tumour from mri image analysis of k-means and dbscan clustering’, *Int. J. Res. Eng. Sci.*, vol. 1, no. 1, pp. 48–57, 2013.

[8] K. Mumtaz and K. Duraiswamy, ‘An Analysis on Density Based Clustering of Multi Dimensional Spatial Data’, *Indian J. Comput. Sci. Eng.*, vol. 1, pp. 8–12, Jun. 2010.

[9] E. Baralis, T. Cerquitelli, S. Chiusano, L. Grimaudo, and X. Xiao, ‘Analysis of twitter data using a multiple-level clustering strategy’, in *International Conference on Model and Data Engineering*, 2013, pp. 13–24.

[10] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, ‘An efficient k-means clustering algorithm: analysis and implementation’, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, 2002.

[11] F. Cao, M. Estert, W. Qian, and A. Zhou, ‘Density-Based Clustering over an Evolving Data Stream with Noise’, in *Proceedings of the 2006 SIAM International Conference on Data Mining*, Apr. 2006, pp. 328–339.

[12] S. Wold, K. Esbensen, and P. Geladi, ‘Principal component analysis’, *Chemom. Intell. Lab. Syst.*, vol. 2, no. 1–3, pp. 37–52, Aug. 1987.

[13] K. Vijay and K. Selvakumar, ‘Brain FMRI clustering using interaction K-means algorithm with PCA’, in *2015 International Conference on Communications and Signal Processing (ICCSP)*, 2015, pp. 909–913.

[14] Mustakim, ‘Centroid k-means clustering optimization using eigenvector principal component analysis’, *J. Theor. Appl. Inf. Technol.*, vol. 95, no. 15, pp. 3534–3542, 2017.

[15] K. Sumathi, S. Kannan, and K. Nagarajan, ‘Data mining: analysis of student database using classification techniques’, *Int. J. Comput. Appl.*, vol. 141, no. 8, pp. 22–27, 2016.

[16] W. I. D. Mining, ‘Data mining: Concepts and techniques’, *Morgan Kaufinann*, vol. 10, pp. 559–569, 2006.

[17] S. García, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining*, vol. 72. Cham: Springer International Publishing, 2015.

[18] J. Quackenbush, ‘Microarray data normalization and transformation’, *Nat. Genet.*, vol. 32, no. 4, pp. 496–501, 2002.

[19] C. Marsboom, D. Vrebos, J. Staes, and P. Meire, ‘Using dimension reduction PCA to identify ecosystem service bundles’, *Ecol. Indic.*, vol. 87, pp. 209–260, 2018.

[20] T. N. Tran, K. Drab, and M. Daszykowski, ‘Revised DBSCAN algorithm to cluster data with dense adjacent clusters’, *Chemom. Intell. Lab. Syst.*, vol. 120, pp. 92–96, 2013.

[21] Y. Chen, S. Tang, N. Bouguila, C. Wang, J. Du, and H. Li, ‘A fast clustering algorithm based on pruning unnecessary distance computations in DBSCAN for high-dimensional data’, *Pattern Recognit.*, vol. 83, pp. 375–387, 2018.

[22] 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, 2016.

[23] M. M. A. Patwary, D. Palsetia, A. Agrawal, W. Liao, F. Manne, and A. Choudhary, ‘A new scalable parallel DBSCAN algorithm using the disjoint-set data structure’, in *SC ’12: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, 2012, pp. 1–11.

[24] B. Borah and D. K. Bhattacharyya, ‘An improved sampling-based DBSCAN for large spatial databases’, in *International Conference on Intelligent Sensing and Information Processing, 2004. Proceedings of*, 2004, pp. 92–96.