

## Determination of Optimal Epsilon (Eps) Value on DBSCAN Algorithm to Clustering Data on Peatland Hotspots in Sumatra

Nadia Rahmah<sup>1\*</sup>, Imas Sukaesih Sitanggang<sup>1</sup>

<sup>1</sup>Computer Science Departement, FMIPA, Bogor Agricultural University, Bogor 16680

E-mail: [nadiarahmahh@gmail.com](mailto:nadiarahmahh@gmail.com)

**Abstract.** In this work we determine the optimal epsilon value on peatland on DBSCAN Algorithm to clustering data on peatland hotspots in sumatera. DBSCAN is a base algorithm for density based data clustering which contain noise and outliers. We found using this method that the area which has the highest density of hotspots in Sumatra in 2013 peatland is contained in cluster 1 of Riau Province that is equal to 2112 hotspots.

### 1. Introduction

Monitoring of hotspots area could be performed by knowing if the hotspots are randomly distributed or clustered. DBSCAN is a base algorithm for density based clustering containing large amount of data which has noise and outliers. DBSCAN has 2 parameters namely Eps and MinPts. However, conventional DBSCAN cannot produce optimal Eps value. DBSCAN modifications is required to determine the optimal Eps value automatically. Determination of the optimal Eps value is aquired using DMDBSCAN algorithm on a single density level corresponding to the k-dist plot. DMDBSCAN is one method of DBSCAN algorithm modification on Eps value to obtain optimal value automatically with different densities.

The parameter of values are used in the clustering data on hotspots peatlands in Sumatra. The data set of peatlands hotspots in Sumatera in the period 2013. This research modifies DBSCAN algorithm using the R programming language to get the optimal Eps value automatically to clustering data on peatland hotspots in Sumatera. Implementation of DBSCAN modifications in determining the value of parameter Eps is expected to be a reference to the DBSCAN algorithm, when searching the Eps value and determining the spread of hotspots that could be known clusters of the region that have the potential occurrence of peatlands fire.



## 2. Results and Discussions

### Determination of Optimal Eps Value Using R Programming Language

This phase is done by modifying the algorithm DBSCAN in the programming language R. Modifications done by inserting a search algorithm to the algorithm DBSCAN Eps value on R found in accordance with the package fpc. DMBSCAN pseudocode is presented in Figure 1.

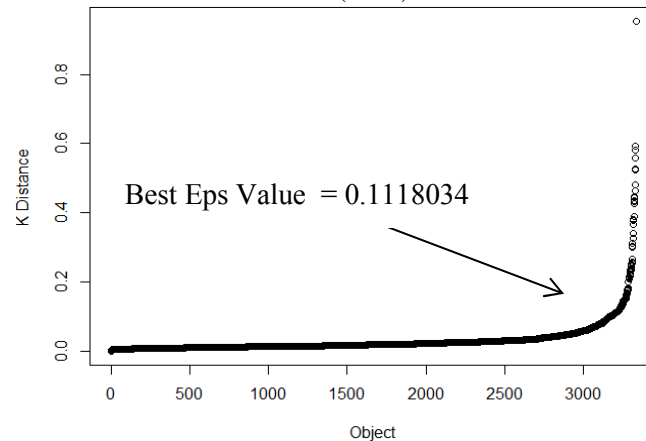
<i>Algorithm 1 The pseudo code of the proposed technique DMBSCAN to find suitable Epsi for each level of density in data set</i>	
Purpose	<i>To find suitable values of Eps</i>
Input	<i>Data set of size n</i>
Output	<i>Eps for each varied density</i>
Procedure	<pre> 1  for i 2  for j = 1 to n 3    d(i,j) ← find distance (x<sub>i</sub>, x<sub>j</sub>) 4  find minimum values of distances to nearest 3 5  end for 6  end for 7  sort distances ascending and plot to find each value 8  Eps corresponds to critical change in curves </pre>

**Figure 1** Pseudocode DMBSCAN Algorithm (Elbatta 2012)

The algorithm begins with the calculation euclidean distance on each pair of data latitude and longitude by dist function. Furthermore dist transform data into a matrix using as.matrix function in R. A function of matrix is used because the results dist upper triangular form that needs to be normalized into the matrix intact. Normalization is used to facilitate the search for the  $k$  nearest neighbors on each line of the distance calculation.

After calculating the distance and the matrix, the initialization is done to count the number of rows in the data by using the function nrow. Further searches carried out three nearest neighbors of each matrix line spacing for sorting is done in ascending for each result the closest distance to the neighbors. Sorting the results of each line of the nearest neighbor made a plot with the x-axis and y-axis is the object and the distance  $k$  nearest neighbors. Plots that have formed in ascending calculated the difference in slope of the line to get the value of Eps (Gaonkar and Sawant 2013). A point which has a slope changes or changes in the slope of the plot will be a significant Eps optimal value (Elbatta 2012).

In this study, an algorithm DMBSCAN was adopted is to use single-level density to determine the optimal value Eps so that only produce one value Eps alone. Eps rate determination for single-level density is obtained by calculating the slope between points with equation  $(y_2 - y_1) / (x_2 - x_1)$  using a threshold value of 1% so that the slope has a slope of 1% difference is an optimal Eps value. Figure 2 depicts the result of a plot that has been sorted in ascending Using data hotspot 2013. Searches Eps value was done by calculating the slope of the line from any point and sought-after pair of points that have the greatest slope to locate the point. The slope of the line is located at the point of 0.1118034, a point which is the optimal value Eps (Elbatta 2012).

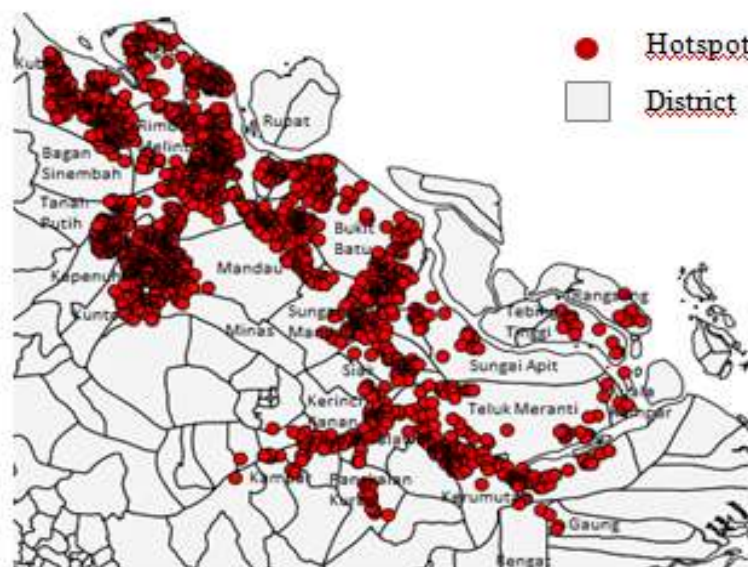


**Figure 2** Points sorted by distance to the 3rd nearest neighbor

### *Clustering of Data Hotspots Using Optimal Eps Value on DBSCAN Algorithm*

Clustering process is successfully executed from the search results DBSCAN modification Eps optimal value automatically with the idea of k-dist plot to generate the parameter Eps = 0.1118034 and MinPts = 3. This parameter is used for clustering processes that generate 43 clusters with outliers at 54 points. Then the area has the highest density of hotspots in the area of peatland in Sumatra in 2013 contained in cluster 1 (Riau Province) that is equal to 2112 hotspots.

Figure 3 shows the area in Riau province were included in cluster 1. Districts included in cluster 1 is Bengkalis, Pelalawan, Rokan Hilir, Siak, Rokan Hulu and Indragiri Hulu. Subdistrict contained in the districts included in cluster 1 can be seen in Figure 3.



**Figure 3** Area/Subdistrict was included in *cluster 1*

### *Cluster Analyst*

The analysis process performed analysis of running time to the result of determining the optimal Eps automatically on DBSCAN algorithms for clustering data using R. This analysis includes a comparison generated in this studied with research conducted Usman (2014).

This research resulted in a runtime of 1.72 seconds to 32.29 seconds for the system and user time use 2 GB of RAM. The system time is the speed of the system running the job before the user enter data. User time is the time used after users enter data into the system. Therefore, this research is

more efficient in classifying the hotspots data because it can determine the optimal Eps value automatically.

### 3. Conclusion

In this study, an algorithm DMDBSCAN was adopted is to use single-level density to determine the optimal value Eps so that only produce one value Eps alone. The slope of the line is located at the point of 0.1118034, a point which is the optimal value Eps. Clustering process is successfully executed from the search results of DBSCAN modification. We obtain the an Eps optimal value automatically with the idea of k-dist plot resulting in the generatiopn of parameters Eps = 0.1118034 and MinPts = 3. This parameter is used for clustering processes that generate 43 clusters with outliers at 54 points. It follows that the area which has the highest density of hotspots in Sumatra in 2013 peatland is contained in cluster 1 (Riau Province) that is equal to 2112 hotspots. The advantage of this method is that the optimal Eps value can be determined automatically.

### References

- [1] Elbatta MNT. 2012. An improvement for DBSCAN algorithm for best results in varied densities [disertasi]. Gaza (PS): Islamic University of Gaza.
- [2] Ester M, Kriegel HP, Sander J, Xu X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. Di dalam: Simoudis E, editor. *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*; 1996 Agu 4-6. hlm 226-231.
- [3] Gaonkar MN, Sawant K. 2013. AutoEps DBSCAN: DBSCAN with Eps automatic for large dataset. *International Journal on Advanced Computer Theory and Engineering*. 2(2): 11-16.
- [4] Han J, Kamber M, Pei J. 2012. *Data Mining: Concepts and Techniques*. 3rd ed. San Francisco (US): Morgan-Kaufman.
- [5] Usman M. 2014. *Spatial clustering* berbasis densitas untuk penyebaran titik panas sebagai indikator kebakaran hutan dan lahan gambut di Sumatera [tesis]. Bogor (ID): Institut Pertanian Bogor.