Outlier Detection on Hotspot Data in Riau Province using OPTICS Algorithm

N L Febriana¹ and I S Sitanggang¹

¹ Department of Computer Science, Jalan Meranti Wing 20 Level 5 Kampus IPB, Bogor, West Java, Indonesia.

Email: nurlaila76@gmail.com, imas.sitanggang@ipb.ac.id

Abstract. Hotspot is a point that shows the coordinate of an area with relatively higher temperature than another surrounding area. Hotspot is one of indicators of forest and land fires. An effort in forest and land fires prevention is detecting hotspot occurrences and its outlier. The purpose of this study is to detect outlier occurrences on hotspot data in Riau Province using the density based algorithm namely Ordering Point to Identify the Clustering Structure (OPTICS). The data used are hotspots in Riau Province for the period of 2001 to 2012. In order to find the best clustering results, OPTICS was executed on the parameter Eps of 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.1 and MinPts of 1 to 6. This study found that outliers are commonly occurred in hotspot data in 2007 at the parameter Eps of 0.01 and MinPts of 6. This study identifies 906 outliers in hotspot data in 2007 with SSE of clustering result of 0.0219. Outliers are mostly found in February spreading on several districts including Siak district in Riau Province.

1. Introduction
Detecting hotspot occurrence is an important activity in forest and land fire prevention. However, forest and land fire can be located far from the position of hotspot. In addition, some hotspot locations are difficult to be reached and it may also too far from the water resources. This condition causes the fire supression activities to be delayed, forest and land fires occur much longer, and fires propagate to the other regions. Therefore losses values due to fires are increased.

The study by [1] reports that in the period July to October 2015, trajectory CO and CO₂ pollutants from peat fires in Riau Province spread to the west, northwest, east, northeast, and north of the Riau Province to the province of West Sumatra, North Sumatra, Aceh and Riau Islands. On July 9, July 22, July 26, August 30 and October 21, 2015, high pollutants concentration spread in some districts in Riau province including Rokan Hilir, Bengkalis, Meranti island, Indragiri Hulu, and Siak. In the study by [1], haze trajectory patterns are obtained using the HYSPLIT model and the CO and CO₂ pollutants are analyzed using K-Means algorithm.

Because forest and land fires have negative impacts for the environment, monitoring hotspot distribution is required to minimize damages due to fires. In addition, identifying hotspot distribution especially those are located apart from other hotspots is important in forest and land fires prevention. Those hotspots which are occurred at isolated locations are considered as outliers. In addition to determine hotspots in dense clusters, outliers on hotspot data should be detected because those outliers are potentially to be fires. Several clustering algorithms including K-Means, DBSCAN and Kulldorff’s
Scan Statistics Method have been applied to analyze hotspot distribution in the previous works [2], [3], [4], [5], [6]. Studies on outliers detection on hotspot datasets have been done by applied clustering algorithms. The work by [7] applied the DBSCAN algorithm to detect outliers on hotspot data in 2005 in Riau Province. This work find as many 1241 outliers appeared in the study area. In addition, the work of [8] applied k-means clustering algorithm to discover outliers on hotspot datasets. This study identifies 59 collective outliers and 30 global outliers on the hotspot dataset. These outliers on the hotspot data are mostly occurred in February, March, June, and August. Another study by [9] detected contextual outliers on hotspot data in Riau province for the period 2001 to 2009 based on climate context, i.e. rainfall. The result of this study showed that there were 54 objects detected as contextual outliers, many of them occurred in February, March, June, July, and August. The contextual outliers have an average of daily occurrences of 65.76 hotspots with rainfall in average of 37.15 mm. This research aims to detect outlier occurrences on hotspot dataset in Riau Province on the period of 2001 to 2012 using the OPTICS algorithm. The results are compared to the output of DBSCAN algorithm [7].

2. Data and Methods

2.1. Hotspot Data
Hotspot data used in this study were collected from the National Aeronautics and Space Administration (NASA) Fire Information for Resource Management (FIRMS) http://earthdata.nasa.gov/. Data preprocessing including data cleaning and data transformation was done in the previous work by [8]. The datasets for clustering has three attributes namely longitude, latitude, and date of hotspot occurrence (acq_date).

2.2. Outlier Detection using OPTICS Algorithm
Ordering Points to Identify the Clustering Structure (OPTICS) is one of the density-based cluster algorithm that was introduced by [10]. In density-based cluster algorithms, objects are in dense regions are assigned as members of clusters. Noises are objects in less dense regions which are separately located from members of clusters. OPTICS is an extension of Density-Based Spatial Clustering Algorithm with Noise (DBSCAN) algorithm. The key idea in DBSCAN is also applied in OPTICS in which for each object of a cluster the neighbourhood of a given radius (Eps) has to contain at least a minimum number of objects (MinPts) [10]. The algorithm has two parameters namely Epsilon (Eps) and Minimum Points (MinPts). Epsilon (Eps) is the distance of cluster centroid to its neighbourhoods and MinPts is minimum number of object in one cluster [11]. OPTICS does not create cluster explicitly. It creates ordering of all objects in the dataset based on the density-based clustering structure. There are two values are assigned to each object namely a core-distance and a reachability-distance which are defined as follows [10]:

- Given $p$ as an object from database $D$; $\varepsilon$ is a distance value; $N_\varepsilon(p)$ is $\varepsilon$-neighbourhood of $p$; MinPts is integer; MinPts_distance ($p$) is a distance between $p$ to its MinPts’ Neighbourhood. Core distance of the object $p$ is a minimum distance ($\varepsilon'$) between $p$ and another object in its $\varepsilon$-neighbourhood. $p$ is considered as a core object if this neighbour consists of $N_\varepsilon(p)$, and otherwise the core distance of the object $p$ is considered as UNDEFINED.

$$\text{core} \_ \text{distance}_{\varepsilon,\text{MinPts}}(p) = \begin{cases} \text{UNDEFINED, if Card}(N_\varepsilon(p)) < \text{MinPts} \\ \text{MinPts} \_ \text{distance}(p), \text{otherwise} \end{cases}$$  \quad (1)

- Given $p$ and $o$ are objects from the database $D$; $N_\varepsilon(o)$ is number of $o$’s neighbours in $\varepsilon$-neighbourhood; MinPts is integer. The reachability-distance of $p$ with respect to $o$ is defined as follows
reachability_distance_{\varepsilon, \text{MinPts}}(p, o) = \begin{cases} 
\text{UNDEFINED, if } |N_\varepsilon(p)| < \text{MinPts} \\
\max(\text{core_distance}(o), \text{distance}(o, p)), \text{otherwise}
\end{cases}

(2)

Core distance of the object \(p\) is a minimum distance (\(\varepsilon'\)) between \(p\) and another object in its \(\varepsilon\)-neighbourhood. \(p\) is considered as a core object if its neighbour consists of \(N_\varepsilon(p)\), and otherwise the core distance of the object \(p\) is considered as \textit{UNDEFINED}.

In the OPTICS algorithm, reachability-distance does not allowed to be smaller than the core distance. If this situation is happened then there will be no objects in \(o\) neighbourhood [10]. Euclidean distance id used to calculate the core-distance and reachability-distance.

Sum Square Error (SSE) was calculate to evaluate clustering results in which the best clustering was selected based on minimum SSE. The formula of SSE is as follows [12]:

\[\text{SSE} = \sum_{i=1}^{K} \sum_{x \in C_i} d(p, m_i)^2\]

(3)

where \(d\): distance between the objects; \(m_i\): a centroid of cluster \(i\); \(p \in C\); \(p\) is an object in cluster \(i\).

3. Result and Discussion

OPTICS algorithm that is available in Weka was used to generate clusters on a hotspot dataset as well as to identify outliers on the dataset. For all objects in the dataset, OPTICS assigns the values of Core-Distance and Reachability-Distance. Table 1 provides some examples of objects as the results of OPTICS algorithm.

| Id | Objects (longitude, latitude) | Core-Distance | Reachability-Distance |
|----|-------------------------------|---------------|----------------------|
| 0  | 101.476, 1.673                | UNDEFINED     | UNDEFINED            |
| 1  | 102.807, 1.134                | UNDEFINED     | UNDEFINED            |
| 10 | 102.005, 0.796                | 0.00064       | UNDEFINED            |
| 127| 102.005, 0.795                | 0.00084       | 0.00064              |
| 173| 102.004, 0.797                | 0.0007        | 0.00064              |

Core-distance in table 1 indicates the distance of an object within a certain radius that meets MinPts. Reachability-distance denotes distance between an object with another object. If number of objects within a certain radius does not meet MinPts, then the distance is considered as UNDEFINED. In this study, if core distance and reachability distance of an object is UNDEFINED then the object is considered as an outlier. An object is classified as an outlier or abnormal if it meets the following criteria (1) The object is not included in any cluster, (2) the object is far away from the nearest cluster, (3) the object is a member in a small cluster [12].

Several values of parameters Eps and MinPts are used to determine the best clustering results in which outliers are identified in the best cluster. The value of Eps are 0.01, 0.02, 0.03, 0.04, 0.05, 0.06 (in degree) whereas the MintPts starts from 1 to 6 as shown by the following table.
Table 2 represents clustering results on hotspot dataset in 2001 to 2012 at the Eps of 0.01 and MinPts of 6. Based on table 2 the outliers mostly appeared in 2007. Figure 1 shows number of cluster’s member and outliers as the results of OPTICS algorithm in a histogram.

Table 2. Clustering results of OPTICS algorithm at the Eps of 0.01 and MinPts of 6.

| Year | Number of object in cluster | Outlier | SSE  |
|------|----------------------------|---------|------|
| 2001 | 1,264                      | 410     | 0.0060 |
| 2002 | 5,181                      | 773     | 0.0728 |
| 2003 | 6,041                      | 833     | 0.0337 |
| 2004 | 7,498                      | 890     | 0.0395 |
| 2005 | 22,369                     | 671     | 0.0772 |
| 2006 | 10,245                     | 879     | 0.0545 |
| 2007 | **3,188**                  | **906** | **0.0219** |
| 2008 | 4,843                      | 807     | 0.0278 |
| 2009 | 10,167                     | 728     | 0.0457 |
| 2010 | 3,414                      | 686     | 0.0194 |
| 2011 | 5,997                      | 843     | 0.0320 |
| 2012 | 7,034                      | 786     | 0.0361 |

Figure 1. Number of objects in cluster and outliers.

From the figure 1 in average as many 767 outliers are detected on hotspot datasets each year in the period of 2001 to 2012. These outliers are hotspots that are located not in clusters meaning that distance between one a hotspot to another is greater to 0.01 or about 1.1132 km. The OPTICS algorithm results the highest number of outliers on hotspots in 2007, the outlier’s distribution as shown in Figure 2.
Figure 2. Outlier occurrences in district level, Riau Province in 2007.

Figure 2 shows a plot of 906 outliers on hotspot data in 2007 in which those outliers are mostly occurred in February. The outlier distribution in district level in 2007 is as follows: 100 outliers in Bengkalis District, 99 outliers in Indragiri Hilir District, 97 outliers in Indragiri Hulu District, 111 outliers in Kampar District, 26 outliers in Dumai City, 9 outliers in Pekanbaru City, 71 outliers in Kuantan Sengigi, 99 outliers in Pelalawan, 90 outliers in Rokan Hilir, 90 outliers in Rokan Hulu, and 112 outliers in Siak. Based on the distribution, in district level the outliers mostly appeared in Siak and 71 of them appeared on February. The outliers in Siak plotted in sub districts level, as follows.

Figure 3. Outlier occurrences in sub districts level in Siak District, Riau Province in 2007.
Figure 3 shows 112 outliers distributed 6 in Bunga Raya, 1 in Dayun, 10 in Kandis, 1 in Kerinci Kanan, 5 in Koto Gasib, 15 in Minas, 2 in Sabak Auh, 1 in Siak, 31 in Sungai Apit, 38 in Sungai Mandau, and 2 in Tualang.

The previous research by [7] identified outliers on hotspot datasets in 2001 to 2012 using the DBSCAN algorithm. Table 3 provide the comparison of clustering results using DBSCAN and OPTICS algorithms at the Minpts of 2. The output of the DBSCAN algorithm obtained from the previous work [7].

| Year | Eps | Number of object in cluster | Outlier % | SSE | Number of cluster | Outlier % | SSE |
|------|-----|-----------------------------|-----------|-----|-------------------|-----------|-----|
| 2001 | 0.02| 1515                        | 159       | 9.48%| 0.024             | 175       | 188 | 11.21%| 0.048|
| 2002 | 0.02| 5904                        | 50        | 0.84%| 0.068             | 457       | 769 | 12.91%| 0.036|
| 2003 | 0.02| 6808                        | 74        | 1.08%| 0.1               | 497       | 375 | 5.45% | 0.022|
| 2004 | 0.1 | 8236                        | 152       | 1.81%| 0.027             | 25        | 14  | 0.16% | 0.026|
| 2005 | 0.01| 22902                       | 138       | 0.59%| 0.041             | 864       | 1241| 5.38% | 0.084|
| 2006 | 0.01| 11054                       | 70        | 0.63%| 0.141             | 862       | 1229| 11.04%| 0.032|
| 2007 | 0.02| 4048                        | 46        | 1.12%| 0.070             | 446       | 409 | 9.99% | 0.021|
| 2008 | 0.02| 5620                        | 30        | 0.53%| 0.069             | 489       | 366 | 6.48% | 0.018|
| 2009 | 0.02| 10866                       | 29        | 0.27%| 0.083             | 768       | 990 | 9.09% | 0.022|
| 2010 | 0.02| 4057                        | 43        | 1.05%| 0.055             | 389       | 311 | 7.58% | 0.021|
| 2011 | 0.02| 6804                        | 36        | 0.53%| 0.071             | 500       | 349 | 5.10% | 0.021|
| 2012 | 0.02| 7786                        | 34        | 0.43%| 0.074             | 742       | 946 | 12.09%| 0.024|

Table 3 shows the differences of clustering result using OPTICS and DBSCAN with the same parameters. OPTICS algorithm provide a number of object in cluster, while DBSCAN algorithm provide number of cluster. Both of the algorithm have the same parameters and similar process with a very difference result. OPTICS algorithm do not provide cluster explicitly.

4. Conclusion
The OPTICS algorithm has successfully applied on the hotspot datasets in the period 2001 to 2012 to identify outliers on the datasets. An outlier on hotspot datasets is a hotspot that is located far from the clusters of hotspots. The distance of those outliers to other nearest hotspots is about 1.11 km. Experimental results at the parameter Eps of 0.01 and MinPts of 6 shows that about 767 outliers are found on hotspot dataset per year in the period 2001 to 2012 in which the highest number of outliers are found in 2007. As many 906 outliers found in 2007 that mostly spread in the districts of Bengkalis, Indragiri Hilir, Indragiri Hulu, Kampar, Pelalawan, Rokan Hilir, Rokan Hulu, and Siak.

References
[1] Ni'am M, Sitanggang I S, Nuryanto D E 2016 3rd Int. Symp. Of LISAT on Food Security and Environmental Monitoring (October 25th-26th Bogor)
[2] Kirana A P, Sitanggang I S, Syaufina L 2015. J. TELKOMNIKA 13(4) 1376-83
[3] Hermawati R, Sitanggang I S 2015 Proc.2nd Int. Symp. On LAPAN-IPB Satellite for Food Security and Enviromental Monitoring (Bogor) 317-23
[4] Usman M, Sitanggang I S, Syaufina L 2015 Proc. 1st Int. Symp. On LAPAN-IPB Satellite for Food Security and Enviromental Monitoring (Bogor) 132-40
[5] Fuad T, Sitanggang I S, Annisa 2010 IEEE Int. Symp. On Information Technology (June 15th-17th Kuala Lumpur) CFP1033E-CDR ISBN: 978-1-4244-6717-4
[6] Hayardisi G, Sitanggang I S, Syaufina L 2009 IEEE 2nd Conf. On Data Mining & Optimization (October 27th-28th) CFP0982H-PRT

[7] Sukmasetya P and Sitanggang I S 2016 IOP. Conf. On Earth and Environmental Science 31(1) 1-4

[8] Sitanggang I S and Baehaki D A M 2015 IEEE 2nd Proc. Int. Conf. On Spatial Data Mining and Geographical Knowledge Services (July 8th-10th Fuzhou) p 66 CFP1584-ART ISBN 978-1-4799-7749-9

[9] Thah P H, Sitanggang I S 2015 Proc.2nd Int. Symp. On LAPAN-IPB Satellite for Food Security and Environmental Monitoring (Bogor) 238-63

[10] Ankerst M, Breunig M M, Kriegel H P, Sander J 1999 OPTICS: Ordering Points to Identify the Clustering Structure (Philadelphia-Munich: Institute for Computer Science University of Munich) 3-4

[11] Ester M, Kriegel H P, Sander J, and Xu X A Density-Based Algorithm for Discovering Clusters in Large Spatial Database with Noise (Munich: Institute for Computer Science University of Munich) 2-3

[12] Han J, Kamber M, Pei J 2012 Data Mining Concepts and Technique (San Diego: Morgan Kaufman Publisher) 3rd