Outlier Detection on Hotspots Data in Riau Province using DBSCAN Algorithm

Pristi Sukmasetya[1]*, Imas S. Sitanggang[2]
Computer Science Department Bogor Agricultural University, Jalan Meranti Wing 20 level 5 Kampus IPB, Bogor, West Java, Indonesia

E-mail: pristisukmasetya92@gmail.com

Abstract. Indonesia has serious problems in forest fires. One of the potential factors which indicates forest fires is hotspot. Hotspot is a forest fires indicator that detects a location with relatively higher temperature in comparison with nearby positions. One possible prevention efforts for forest fires is by detecting outliers on hotspots data. This study detects outlier on hotspots data in Riau Province in between year 2001 to 2012 using the DBSCAN algorithm and determines the distribution of outlier hotspots by region and time. The experiment results show that the highest occurrence of outliers is in 2005. The number of outliers on hotspots data reaches 1241 hotspots with the sum of square error (SSE) is 0.084. Outlier hotspots in Riau Province in 2005 spread across 11 districts/cities and 136 districts. In 2005 the highest outlier are found in Rokan Hulu with the number of outliers is 186 points. The highest frequency of hotspot that is considered as outliers is found in August 2005, with a total of 355 outliers in which as many 97 of these outliers are occurred in Rokan Hulu District.

1. Introduction

Indonesia is the largest archipelago country which has serious forest fire problems. Historical data between year 2001-2012 shows about average of 20.000 hotspots warning in Sumatra region every year, with confidence level of detection is more than 30 percent [1]. This indicates the changeover of hotspots occurs dynamically in a region over time.

Ref. [2] shows one of indicator of forest fires that will likely occurs is the hotspot. Monitoring of hotspots is done using remote sensing by satellites. The hotspot data takes form time-series , which the hotspot data are periodically observed every day by satellites such as NOAA series with AVHRR sensor onboard or Terra/Aqua satellite with MODIS sensor onboard. It is possible that the distribution of hotspots gather in a space naturally, so that data can be analyzed using clustering techniques. However it is also possible that there are hotspots which scattered far from the hotspot distribution, so the probability of hotspot as outliers also can be analyzed. Some clustering techniques are K-means, hierarchical cluster, DBSCAN and ST-DBSCAN. Among all the clustering techniques mentioned before, DBSCAN algorithm can be used to find cluster from large spatial database [3].

*corresponding author

Outlier is a set of data which is considered to have different properties or deviate when compare with other majority data. Outlier analysis is also known as anomaly analysis, anomaly detection or detection of deviations [4]. Outliers often contain useful information about the characteristics of the abnormal data. An object is said to be an outlier if (1) the object is not belongs in any cluster, (2) there is a great distance between the object to the closest formed cluster, (3) the object is included in a small cluster or separate clusters. The object is not included in any cluster can be detected using the algorithm DBSCAN [4].

Therefore, outlier detection from hotspot data can give information about the distribution of hotspots. If it deviates than it can be concluded there are anomalous dispersion of hotspots occurrence. The frequency of occurrence of the hotspot outliers is analyzed as an indicator of forest fires.

In this present paper, the authors implemented outlier detection on hotspot data in Riau Province in between year 2001 to 2012 using the DBSCAN algorithm to locate outliers based on its location and time of occurrence.

2. Data and Method

2.1. Data

The attributes of the data consist of latitude, longitude and acq_date. Latitude and longitude attributes describe the location of hotspots in Riau Province, while the acq_date attribute describes the date of hotspot occurrence in a certain time. Fig. 1. represents the number of hotspots in Riau province.

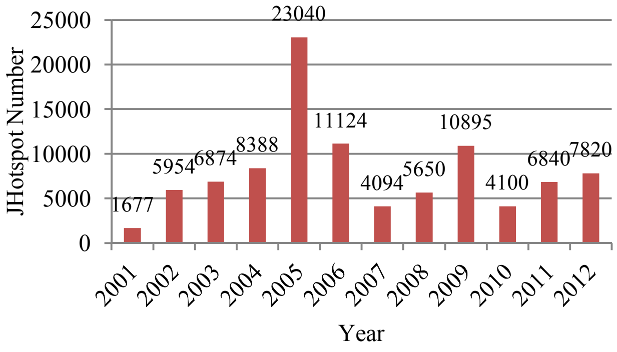

Figure 1. Hotspot data in Riau Province.

2.2. Spatial hotspot clustering using DBSCAN algorithm

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a clustering algorithm which is developed based on the density level of data. DBSCAN find a central object which has a dense neighborhood. DBSCAN connects the center of the object with its neighborhood to create an area as a cluster [4].

DBSCAN algorithm was first introduced which is designed to find clusters and its outliers on spatial data. This algorithm has two main parameters, namely Eps and MinPts. Eps is the distance between center object of cluster and it's neighborhood, while MinPts is the minimum required number of neighbor objects from center object of cluster [6]. DBSCAN concept is as follows:

- The neighborhood which located within a radius Eps is called *Eps-neighborhood* of data objects,
- If the *Eps-neighborhood* of an object containing at least a minimum number *MinPts*, then the object is called a core object,

- A point *p* is said *directly density-reachable* from a point *q*, if *p* is adjacent with point *q* with a certain distance (ε) and point *q* is the center point,
- An object *p* is *density reachable* from object *q* with Eps and MinPts in a set of objects *D* if there is a chain of objects (p_1, p_2, ..., p_n), where p_1 = q and p_n = p, which p_i + 1 *density reachable* from p_i with Eps and MinPts, for $1 \leq i \leq n$, p_i member of D,
- An object p is *density-connected* to object q with Eps and MinPts in a set of objects D if there is an object *o* which member of D where p and q are density reachable from o with Eps and MinPts.

2.3. Outlier detection based on clustering

Outliers are objects or data that deviate significantly from the majority objects or data. In periodic data or time-series data, an outlier is defined as a data point that differs significantly from all the data [4].

One way to detect outliers is clustering. Clustering-based method assumes that the normal objects data belong to large and dense groups, while outliers belong to small and sparse group, or not belonging to any group. DBSCAN define outliers as data that do not include in any cluster, or noise [4].The result of clustering had been evaluated between Sum of Square Error (SSE) and silhouette index. This work uses SSE to evaluate clustering results. SSE formula is used to obtain the total number of clusters that minimizes the square error, $p \in C_i$ is each data point in cluster i, m_i is the centroid of the cluster i, and d is the distance to each cluster i [4]. The SSE formula is given as in (1).

$$\mathrm{SSE}= \sum_{i=1}^{K} \sum_{x \in c_i} d\left(p,m_i\right)^2 \quad (1)$$

SSE formula on DBSCAN is obtained as follows, firstly calculate the center object (centroid) of each cluster which has been formed then calculate the distance to each cluster using the Euclidean distance formula and lastly find the total sum.

Calculation of SSE is done using RapidMiner. on each year dataset. The smallest SSE is used as a reference to determine the value of Eps and MinPts which will be used in DBSCAN algorithm to detect outliers.

2.4. Outliers Analysis

At this stage, an analysis of outliers was carried out from hotspots data in Riau Province based on location and time of occurrence of hotspots using the DBSCAN algorithm. The terms outliers in this paper is defined as a hotspot that stray away from the other hotspots location in general.

3. Results and Discussions

3.1. Spatial Hotspot Clustering

Various formation of experiments was carried out to determine parameter values for the DBSCAN algorithm. The values of Eps are 0.1, 0.06, 0.05, 0.04, 0.03, 0.02, and 0.01. The values of MinPts which were used in experiment are 1, 2, 3, 4, 5, 6.

The combination of Eps and MinPts used to cluster data hotspots in Riau Province from 2001 to 2012. Comparing of the these various Eps and MinPts produce the best Eps and MinPts value, and both will be used to form clusters of hotspot distribution data from 2001 to 2012 using the DBSCAN algorithm. For example, to find a cluster with variations in the value of Eps of 0.05 and MinPts of 2 we execute the R code as follows:

```
library(fpc)
tabel<-read.table ("HOTSPOT_2005.csv",header= TRUE,sep=",")
```

```
tabel$acq_date <- NULL
sapply(tabel, class)

#DBSCAN
x <- as.matrix(tabel)
ds <- dbscan(x, eps=0.05, MinPts=2)
plot(ds, x)
ds
```

Function *library* is used to call available packages in R software. In this case, function *library* is used to call the fpc packages that contains a module of DBSCAN algorithm and its toolbox. DBSCAN is executed using two parameters namely Eps and MinPts in accordance with the variation mentioned before to cluster hotspots data each year. After the clustering results are obtained, sum of square error (SSE) is calculated.

3.2. Outlier detection based on DBSCAN algorithm

Outlier detection was done after clustering using the DBSCAN algorithm. Outliers in clustering approach are defined as data that do not belong to any cluster. Table 1 shows the clustering result using the DBSCAN algorithm and outliers in the hot spot data from 2001 to 2012.

Table 1. Clustering result using DBSCAN algorithm.

Year	Eps	MinPts	Number of Cluster	Number of outlier	Percentage	SSE
2001	0.02	2	175	188	11.20%	0.048
2002	0.01	2	457	769	12.90%	0.036
2003	0.02	2	497	375	5.45%	0.022
2004	0.10	2	25	14	0.16%	0.026
2005	**0.01**	**2**	**864**	**1241**	**5.38%**	**0.084**
2006	0.02	2	862	1229	11.00%	0.032
2007	0.02	2	446	409	9.99%	0.021
2008	0.02	2	489	366	6.48%	0.018
2009	0.02	2	768	990	9.09%	0.022
2010	0.02	2	389	311	7.58%	0.021
2011	0.02	2	500	349	5.10%	0.021
2012	0.02	2	742	946	12.09%	0.024

3.3. Outlier Analysis

The highest frequency outlier in hotspots data is in the year 2005. In that year large and severe fires were occurred, so year 2005 was practically suitable and then chosen for analysis. Fig. 2. shows a map of the pattern of hotspots spread in 2005 in Riau Province.

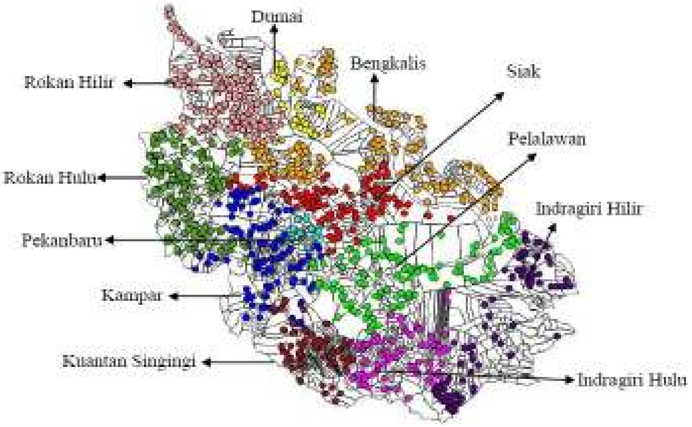

Figure 2. Hotspot spread in Riau province in 2005

Outlier detection on hotspots data could be analyzed by location or time of hotspots occurrence. Clustering using the DBSCAN algorithm results the highest occurrence of outliers in year 2005 with 1241 outlier with 864 clusters and SSE about 0.084. Outlier on hotspots data in Riau Province in year 2005 spread across 136 sub-districts and 11 districts/cities. The number of outliers based on districts/cities in Riau Province are as follows, Bengkalis-165 hotspots , Indragiri Hilir-97 hotspots, Indragiri Hulu-93 hotspots, Kampar-134 hotspots, Dumai City-37 hotspots, The capital city of Pekanbaru-20 hotspots, Kuantan Singingi-71 hotspots, Pelalawan-132 hotspots, Rokan Hilir-175 hotspots, Rokan Hulu-186 hotspots, and Siak-131 hotspots.

Fig. 3. shows a Riau map with outlier patterns in the year 2005 based on location. Each district / city is distinguished by distinct color.

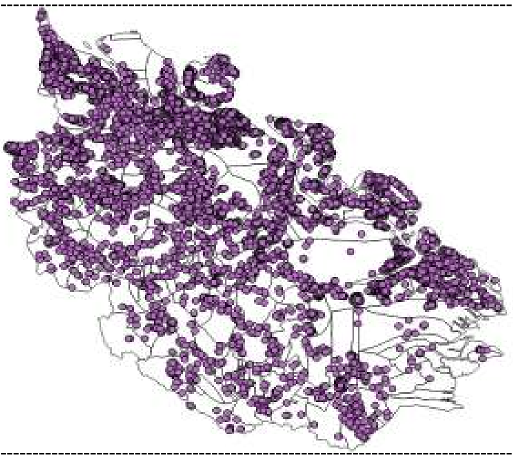

Figure 3. Riau map with outlier pattern in year 2005

In year 2005, the highest outlier of hotspots is occurred in Rokan Hulu district which has 186 outliers. The hotspots are scattered in 13 sub-districts. From all Rokan Hulu's sub-district, the as many 50 hotspot outliers are found in Tambusai sub-district. The outliers spread over 9 villages in which in West Tambusai village and East Tambusai village 8 outliers are detected. Fig. 4. shows the spread of outliers on hotspots data in Rokan Hulu.

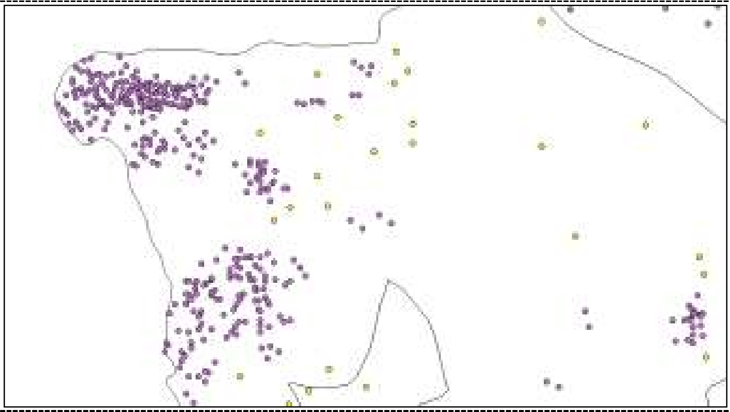

Figure 4. Rokan hulu outlier

In term of time of hotspot occurrence, as many 355 objects that are considered as outliers are occurred in August 2005. In this period Rokan Hulu district has the highest number of outliers i.e. 97 outliers compared to other districts.

Outliers in hotspots data, can be used as an alternative for consideration in keeping attention to the location of the hotspot which stray away that may become trigger to forest fires, judging from the number occurrence of hotspots outliers.

4. Summary

This work is able to obtain information of outliers occurrence on hotspots data in Riau Province from the year 2001 to 2012. The information gained including information about the name of villages, sub-districts, districts/cities and time of outliers occurrences. The highest number of outliers is occurred in 2005 with the number outliers of 1241, SSE of 0.084, Eps of 0,01 and minpts of 2. Outlier points spread in 11 districts/cities and 136 sub-districts in Riau province, with the highest is located in Rokan Hulu for 186 outliers. Outliers on hotspots data are scattered in 13 districts and are mostly found in Tambusai sub-district, as 50 points, with the largest is in West Tambusai and East Tambusai as 8 points respectively. Throughout year 2005, the occurrence of outliers reached 355 hotspots and most prevalent in Rokan Hulu district as many 97 points. This outlier refers to hotspot which its location faraway from majority hotspot. The information obtained from this research can be used as an alternative for consideration in keeping attention to the location of the hotspot outliers that may become trigger to forest fires.

References

[1] K. Austine, A. Alisjahbana, N. Sizer. (2013). *Data terbaru menunjukkan kebakaran hutan di indonesia adalah krisis yang telah berlangsung sejak lama*. [Online]. Available: http://insights.wri.org/news/2013/06/ data-terbaru-menunjukkan-kebakaran-hutan-di-indonesia-adalah-krisis-yang-telahberlangs#

[2] W.C. Adinugroho, *et al*. *Panduan Pengendalian Kebakaran Hutan dan Lahan Gambut*. Bogor, Indonesia: Wetlands International, 2005.

[3] M. N. Gaonkar and K. Sawant, "AutoEpsDBSCAN: DBSCAN with Eps automatic for large dataset," in *International Journal on Advanced Computer Theory and Engineering*. India, 2013, pp. 2319-2526.

[4] M. Ester, et al, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining.* Oregon, United States, 1996, pp. 226-231