VOL. 12, NO. 13, JULY 2017 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

DATA MINING APPROACH FOR OUTLIER DETECTION ON HOTSPOT DATA AS FOREST AND LAND FIRE INDICATOR: A CASE STUDY IN RIAU PROVINCE INDONESIA

Imas Sukaesih Sitanggang, Mohamad Bentar Cahyadahrena and Shofyan
Department of Computer Science, Bogor Agricultural University, Bogor, Indonesia
E-Mail: imas.sitanggang@ipb.ac.id

ABSTRACT

Forest fire is one of environmental problems which has continuously repeated and causes serious problems in health and environment especially in Sumatera and Kalimantan Island Indonesia. A hotspot is an indicator for forest fires which is daily collecting by some national and international institutions. This study aims to identify outliers from a hotspot dataset based on its occurrence using data mining approach. The dataset consists of 4383 daily hotspots and 144 monthly hotspots in Riau Province in Sumatera Island for the period of 2001-2012. A medoid-based clustering algorithm namely Partitioning Around Medoids (PAM) was applied on the hotspot frequency datasets and results 17 best clusters in which the average frequency of outliers is 351.11. These outliers are mostly occurred in February, March, June, July, and August. In addition, Local Outlier Factor (LOF) algorithm was applied to identify outliers based on the location of hotspots. This study detects high number of outliers in years 2005, 2006, 2009, and 2012 based on the LOF. Majority outliers spread on several districts including Dumai, Rokan Hilir, Bengkalis and Pelalawan in Riau Province especially in the period of dry seasons namely February to March and June to August. Outlier distribution on hotspot data is important for forest and land fires prevention.

Keywords: clustering, forest, land fire, hotspot, local outlier factor, outlier detection.

1. INTRODUCTION

Hotspots data for the whole area of Indonesia are daily recorded by satellites such as the National Oceanic and Atmospheric (NOAA) satellite and the AQUA-TERRA satellite through remote sensing technology that results a large dataset of hotspots. A hotspot is one of indicators for forest fire that indicates a location that has relatively higher temperature than surrounding areas. Several institutions in Indonesia including Ministry of Forestry, the Institute of Aeronautics and Space LAPAN and Ministry of Environment monitor, summary, and report the hotspots data to be used by related parties in preventing activities of forest and land fires. Further analysis of large number of hotspots data requires data mining techniques to discover interesting patterns and knowledge that can be potentially used in forest and land fires prevention.

Data mining techniques have been applied to analyse hotspots data in relation to several influencing factors for forest and land fire events. Classification algorithms were used to create prediction models for hotspots occurrences based on physical characteristics, socio-economy and weather conditions of the study area including Riau Province Indonesia [1, 2, 3, 4]. These works applied the decision tree based classification methods on spatial and nonspatial data. The decision tree algorithm was improved in the work of Sitanggang *et al.*, [1] by involving spatial measurement in the Information Gain formula. Spatial information gain was used to select best splitting layer in the spatial dataset in growing the tree. Furthermore, the spatial decision tree was applied to predict hotspot occurrences in peatland in the study area Rokan Hilir District, Indonesia. The work of Nurpratami and Sitanggang [3] classified the hotspots occurrences based on city centre, river, road, income source, land cover, population, precipitation, school, temperature, and wind speed using the spatial entropy-based decision tree algorithm in the study area of Bengkalis district, Riau province Indonesia. In addition, a spatial visualization of decision tree based classification model for hotspots was developed in the work of Amri and Sitanggang [4]. A geographical information system (GIS) was developed utilizing OpenGeo Suite 3.0 for classifying hotspot occurrences in Riau Province Indonesia [4].

Furthermore, association rule mining has been conducted to obtain characteristics of the area in Riau Province Indonesia where hotspots are probably occurred [5, 6]. Sitanggang [5] discovered the possible influence factors on the occurrence of fire events in the study area of Rokan Hilir Riau Province Indonesia. The Apriori algorithm was applied on a forest fire dataset which contains data on physical environment (land cover, river, road and city center), socio-economic (income source, population, and number of school), whether (precipitation, wind speed, and screen temperature), and peatlands. Arincy and Sitanggang [6] adopted the multidimensional association rule mining method using Frequent Pattern Growth algorithm (FP-Growth) algorithm and Equivalence Class Transformation (ECLAT) algorithm to determine association patterns between hotspot occurrences and its supporting factors.

In addition to classification and association rule mining tasks, hotspot data were also analysed using clustering algorithms [7, 8, 9]. The widely used clustering algorithm K-Means was applied by Fuad *et al.,* [7] to group hotspot datasets as the results of OLAP (On-line Analytical Processing) operations including roll up and drill down. A visualization module for hotspot clusters

was developed based on the combination of the dimension *time* and *location*. The density based clustering algorithm DBSCAN (Density-Based Spatial Clustering Algorithm with Noise) was used to cluster hotspots in the study area of Sumatra in the years of 2002 and 2013 [8]. The study found that there were changes in the distribution pattern of hotspots and land cover of peat land from 2002 to 2013. The distribution of hotspot clusters in 2002 is mostly found in moderate depth (100-200 cm) peat land, whereas the distribution of hotspot clusters in 2013 is commonly occurred in very deep (> 400 cm) peat land [8]. Kirana *et al*., [9] studied the distribution pattern of hotspot clusters in the peatland areas in Sumatera in the year 2014 using the Kulldorff's Scan Statistics (KSS) method with Poisson model. The assessment of cluster significance was done using Monte Carlo replication. The results showed that the KSS is reliable to detect the clusters of hotspots with the accuracy of 95% in which Riau and South Sumatera province have the highest density of cluster distributions of hotspots [9].

In addition to hotspot clusters, outliers in hotspot datasets are important to be discovered. Outliers in spatial datasets are considered as objects that are located far from the groups of objects. Fire brigades should pay attentions to outliers on hotspot data because the outliers are also potential to be real fires. This study aims to identify outliers in hotspot data using data mining techniques. An outlier is an object in a dataset that is much different from the rest of objects. Outlier detections can be performed using several methods including statistical approach, neural network and machine learning, as discussed in [10]. Clustering is one of methods in data mining that can be used to identify outliers and noises in a dataset. Several works in outlier detections which are based on density, distance, clustering and partition methods are reviewed by Singh and Aggarwal [11].

This study discovers outliers in hotspot in Riau province Indonesia data using the K-Medoids clustering algorithm. Anomalies in the hotspot data indicate high frequency of hotspot occurrences in a certain location and time period. In addition to clustering approach, outliers in hotspot data are identified using the Local Outlier Factor (LOF) algorithm. LOF assigns a hotspot as an outlier if these hotspots are located far from the hotspot groups. Outlier detection from hotspot data gives information about the distribution of hotspots that deviate from other hotspots in term of frequency and its locations. The results are important in preventing forest and land fires so that the effects of fire events can be minimized in the future time.

2. LITERATURE REVIEW

Outlier detection on hotspot data was conducted in our previous works by applying clustering algorithms in data mining including K-Means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and Ordering Point to Identify the Clustering Structure (OPTICS). The previous work by Thah and Sitanggang [12] detected contextual outliers on hotspot data in Riau province for the period 2001 to 2009 based on climate context, i.e. rainfall using K-Means algorithm. The result showed that there were 54 objects detected as contextual outliers, many of them occurred in February, March, June, July, and August [12]. Those objects detected as contextual outliers are hotspots which have high daily occurrences with high rainfall. The contextual outliers have the average of daily occurrences is 65.76 hotspots with the average of rainfall is 37.15 mm [12].

The work by Suci and Sitanggang [13] developed a web-based application to detect outliers on hotspot data using the framework Shiny with the R programming language. The application has several functions including summary and visualization of the selected data, clustering hotspot data using K-Means algorithm, visualization of the clustering results, and displaying global and collective outliers and visualization of outlier spread on the Riau Province Map.

Another work on clustering based outlier detection using the density-based algorithm has been conducted on hotspots data in Riau Province in between year 2001 to 2012 [14]. The algorithm used is DBSCAN [15]. The highest number of outliers is occurred in 2005 which reaches 1241 hotspots at the SSE of 0.084, Eps-neighborhood of data objects (Eps) of 0.01 and minpts of 2. MinPts is the minimum required number of neighbor objects. Those outliers scatter in 11 districts/cities and 136 sub-districts in Riau province in which as many 186 outliers were found Rokan Hulu [14].

Another density-based clustering algorithm namely OPTICS [16] was applied to determine outliers in hotspot data [17]. 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. The results shows that outliers are commonly occurred in 2007 at the parameter Eps of 0.01 and MinPts of 6. As many 906 outliers are identified in hotspot data in 2007 with SSE of clustering results of 0.0219. Outliers are mostly found in February spreading on several districts including Siak district in Riau Province.

3. MATERIALS AND METHODS

3.1 Data

Data used in this work are hotspots in Riau Province, Sumatera Island Indonesia for the period of 2001-2014. The data were obtained from Fire Information for Resource Management System (FIRMS) MODIS Fire/Hotspot, NASA. The hotspot data have 12 attributes describing hotspot specification including location of the point of hotspot, acquired date, confidence and satellite used to record the hotspots. Hotspot datasets as an input for clustering and LOF algorithms were prepared based on three attributes of hotspot namely longitude, latitude and acquired date.

3.2 Research steps

Partitioning-based clustering methods, such as K-means and K-medoids, detect outliers by determining distance between objects and clusters of objects. Using

these algorithms we can partition the dataset into several clusters and examine each object related to the clusters to determine whether an object is considered as an outlier or not. In clustering approach, an object is identified as an outlier if the object is far away from the center of nearest cluster. Furthermore the clustering algorithm may result small clusters. All objects in a small cluster are also considered as outliers.

In order to discover outliers using the clustering and LOF algorithms, several steps were performed. First is data pre-processing to prepare a task relevant dataset as input of the algorithm? The second step is clustering to group objects into clusters based on the similarity of objects. The best clustering result is determined based on the measurement of cluster quality such as sum of squared error (SSE). The third step is to identify outliers in term of hotspot frequency based on the best clustering results. The last step is to discover outliers in term of hotspot locations based on the LOF. Comparison of the results with the previous works is performed to evaluate performance of clustering and LOF algorithm in outlier detection on hotspot data.

3.3 Pre-processing data

Data cleaning and transformation were performed to obtain the task relevant data for clustering and LOF algorithms. Two datasets were prepared for outlier detection. The first dataset contains hotspot frequency in the study area in the period of 2001-2012. The dataset contains 4383 objects of daily hotspot frequency and 144 objects of monthly hotspot frequency. Figure-1 shows the time series decomposition for hotspot data in the period of 2001-2012. The label *data* indicates daily frequency of hotspots. As we can see on the *trend* graph in Figure-1, hotspot frequency increases at the beginning of 2005 and tends to decline in 2007 and continue increasing in 2009.

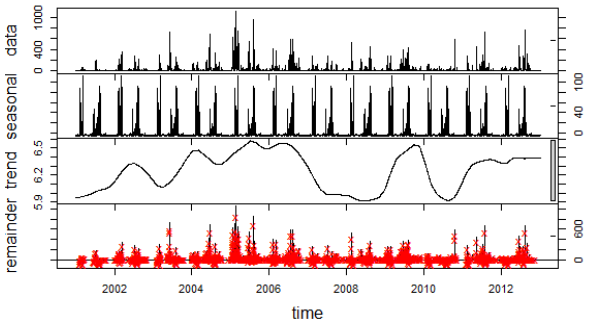

Figure-1. Time series decomposition for hotspot data in 2001- 2012 [18].

The second dataset contains hotspot locations in Riau Province in the period of 2001-2014. Figure-2 shows high number of hotspots with the confidence level greater than or equal to 70% in Riau Province is occurred in the years 2005 and 2014.

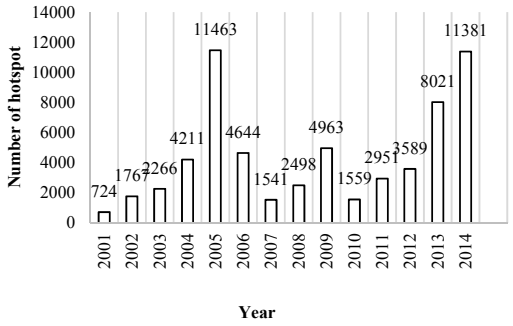

Figure-2. Number of hotspot in the period of 2001-2014 in Riau Province.

3.4 Data clustering

Clustering is a data mining technique to partition a dataset into several groups according to similarity of the objects in the dataset. There are several approaches in clustering including hierarchical, partitioning, and density-based clustering. This work applied a partitioning clustering algorithm namely a representative object-based algorithm K-Medoids [19]. The results are compared to those from a centroid-based algorithm K-Means [19].

In order to evaluate the quality of clustering, sum of squared error (SSE) is used [20]. The K-Means algorithm selects the initial centers of clusters arbitrarily and calculates the mean value of the objects in a cluster as a reference point [19]. A dataset may contain an object with extremely large value. This object may distort the distribution of the data and the value of centroid. Therefore a drawback of K-mean clustering is sensitive to outliers, objects that have much different characteristics to the rest objects.

Instead of determining mean value of objects in a cluster, the k-medoids algorithm takes medoid which is most centrally located object in a cluster. Medoid is considered as representative object of a cluster. Similarity between other objects and a medoid of a cluster is calculated in order to assign the objects to a cluster. Clustering of objects in the dataset is conducted based on the principle of minimizing the sum of the dissimilarities between each object and its corresponding representative object [19]. One of the k-medoids algorithms is the Partitioning Around Medoids (PAM) algorithm [19].

This work used the pam function that is available in the statistical tool R to cluster hotspots frequency in 2001 to 2012 with number of cluster (k) is 2 to 10. The best clustering was selected based on the value of sum of squared error (SSE).

3.5 Outlier detection based on clustering approach

In clustering approach, an object is identified as an outlier if the object is far away from the center of nearest cluster. In addition, the clustering algorithm may result small clusters. All objects in a small cluster are also considered as outliers. Furthermore, if an object in the dataset is not a member of any cluster then this object is considered as an outlier.

According to Han *et al.*, [19] outliers can be grouped into three types as follows:

a) **Global outliers.** A global outlier is an abject in the dataset that deviates significantly from the rest of the dataset. This type of outlier is sometimes called an anomaly.
b) **Contextual outliers.** A contextual outlier is an object in the dataset which deviates significantly with respect to a specific context of the object. To identify this type of outlier we need to define contextual attributes and behavioral attributes of the object. The contextual attributes of an object determine the object's context whereas behavioral attributes define the object's characteristics. The behavioral attributes are used to evaluate whether the object is an outlier in the context to which it belongs.
c) **Collective outliers.** A collective outlier is a group of objects in a dataset that deviate significantly from the whole data set.

The K-means and K-medoids clustering algorithm can be used to identify global outlier as an object that is far from the centroid and the medoid respectively. In order to determine the global outlier, outlier score is calculated as follows [19].

$$Outlier\ Score = \frac{dist(o,c_o)}{I_{c_o}} \quad (1)$$

where
o: object in the dataset
c_o: nearest centroid or center from the object o
$dist(o,c_o)$: distance function between the object to nearest centroid.
I_{c_o}: average value of $dist(o,c_o)$

An object is considered as a global outlier if it has high the value of outlier score. This study identifies global and collective outliers in the hotspot dataset based on clustering approach.

3.6 Density-based local outliers detection

Local outlier factor (LOF) identifies density-based local outliers. Figure-3 illustrates two outliers (O1, O2) and three clusters (C1, C2, C3) of objects. Objects in O1 are local outliers and O2 is a global outlier. Using the clustering approach, objects in O1 could not be detected as outliers. LOF calculates maximum distance of an object from its neighbors in which number of neighbors (k) is defined by users.

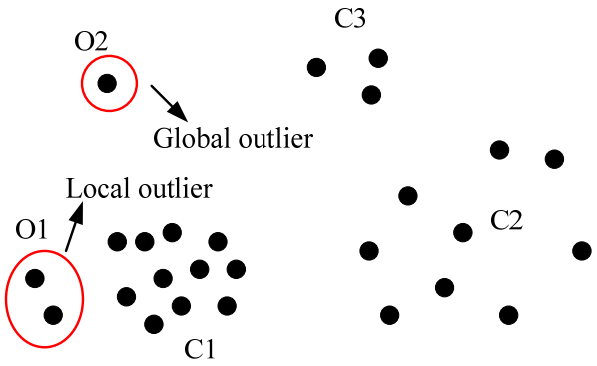

Figure-3. Illustration of global and local outliers.

In order to determine LOF, reachability distance of an object p w.r.t object o is calculated based on the following formula [21]:

$$reach\text{-}dist_k(p,o) = \max\{k\text{-}distance(o), d(p,o)\} \quad (2)$$

In this study, p is a hotspot, o is another hotspot near p. A hotspot o is a nearest neighbor of p with the given value of k. $d(p,o)$ denotes the distance between p and o. $k\text{-}distance(o)$ represents k-distance of a hotspot o. Furthermore Equation (1) is included in the formula *local reachability density* as follows [21]:

$$lrd_{MinPts}(p) = 1/\left(\frac{\sum_{o\in N_{MinPts}(p)} reach\text{-}dist_{MinPts}(p,o)}{|N_{MinPts}(P)|}\right) \quad (3)$$

Local outlier factor of an object p is calculated as follows [21]:

$$LOF_{MinPts}(p) = \frac{\sum_{o\in N_{MinPts}(p)} \frac{lrd_{MinPts}(o)}{lrd_{MinPts}(p)})}{|N_{MinPts}(P)|} \quad (4)$$

In this study, the outlier factor is calculated for hotspots with MinPts of 2 in order to determine which hotspots that are identified as outliers.

4. RESULT AND ANALYSIS

4.1 Outlier detection based on clustering approach

In our previous work, clustering hotspot data was performed using the K-means algorithm that is available in R. Summary of clustering results with number of cluster 10 (k=10) is given in Table-1. These results were used to identify collective outliers based on the cluster size. Objects in small clusters are considered as outliers in the dataset. These objects are members of cluster 5, cluster 7, and cluster 10 in which number of objects in these clusters is less that 1% of the dataset size. Frequency of daily hotspots detected as collective outliers ranges from 284 to 1118 (Figure-4).

Table-1. K-Means clustering results [18].

Cluster ID	Cluster size	Centroid	Min	Max	Percentage
1	248	47.6	35	64	5.66%
2	138	81.19	65	103	3.15%
3	90	125.7	104	157	2.05%
4	40	190.5	159	218	0.91%
5	19	447.26	400	557	0.43%
6	520	20.55	12	34	11.86%
7	17	713.47	589	1118	0.39%
8	38	250.05	222	287	0.87%
9	3250	1.83	0	11	74.15%
10	23	337.61	295	388	0.52%

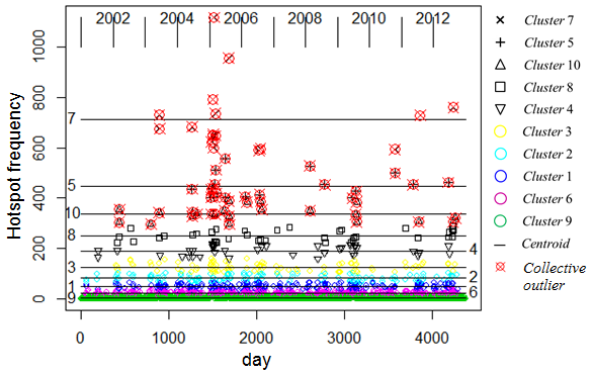

Figure-4. Collective outliers on hotspot data in 2001-2012 [18].

In addition to collective outliers, global outliers were determined based on the outlier score. These types of outliers are mostly found in the cluster 5, 7, and 10. Figure-5 shows the global outliers and its positions with respect to other objects in the dataset.

In addition to K-means clustering, this work was also applied the Partitioning around Medoids (PAM) algorithm to identify collective outliers. The result is provided in Table-2. Small clusters are in the cluster 13, 14, 15, 16 and 17 with number of cluster's members below 1%. Figure-6 shows the plot of objects in these small clusters that are considered as collective outliers and its positions with respect to other objects in the dataset.

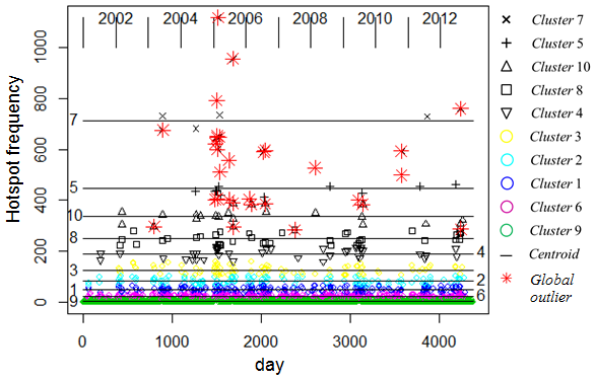

Figure-5. Global outliers on hotspot data in 2001-2012 [18].

Table-3 summaries all outliers obtained in this study and the date when these outliers were occurred. Based on the K-means clustering results, as many 61 outliers were found in the hotspot dataset. These outliers are frquently occurred in February, March, June, July and August in which the highest frequency of hotspots was obtained in 2005.

Table-2. PAM clustering results.

Cluster ID	Frequency of hotspot as medoid	Date of hotspot as medoid	Cluster size	Percentage
1	0	12/31/2012	2163	49.34%
2	8	2/19/2001	413	9.42%
3	3	2/14/2001	674	15.37%
4	15	6/22/2002	258	5.88%
5	49	7/31/2012	120	2.73%
6	66	3/4/2001	101	2.30%
7	24	9/1/2012	203	4.63%
8	36	7/11/2001	146	3.33%
9	89	8/15/2012	73	1.66%
10	141	6/10/2009	56	1.27%
11	209	6/18/2012	41	0.93%
12	110	4/8/2005	50	1.14%
13	335	3/4/2005	17	0.38%
14	266	1/23/2005	28	0.63%
15	648	3/7/2005	16	0.36%
16	428	8/2/2009	22	0.50%
17	956	8/7/2005	2	0.04%

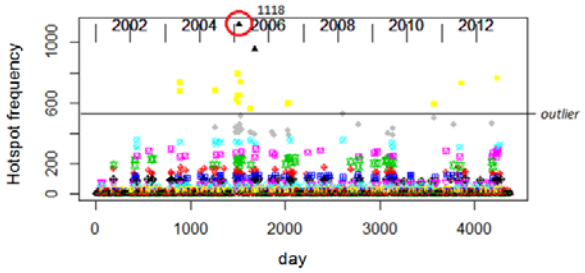

Figure-6. Collective outliers as results of PAM algorithm on the hotspot data in 2001- 2012.

As many 24 outliers were found in 2005. Average frequency of hotspots that are identified as outliers is 481.22. Moreover, the PAM clustering results 123 outliers in the hotspot dataset. Outliers are mostly accourred in February, March, June, July and August. The highest frequency of hotspots occurred in 2005 in which as many 36 outliers were found in this year. Average frequency of hotspots that are identified as outliers is 351.11. According to Table-3, there are no outliers identified in April in the period of 2001-2012. In addition based on K-means clustering results, outliers are not found in January, February, March, and May in 2001, 2004, 2007, 2009, 2010, and 2012. Table-3 shows that there are no outliers identified in November and December in the period of 2001-2012.

Figure-7 shows that outliers in 2005 spread in the whole of Riau province. In January 2005, outliers are detected at the north, northeast, and east parts of the province which include Dumai, Indragiri Hilir district, Siak district and Bengkalis district. In June and August 2005, outliers are occurred at the west, southwest, and south parts of the province. Figure-8 shows that based on the PAM algorithm, outliers are detected almost in the whole of study area namely Riau Province. In February 2005 outliers are found in Pekan Baru, Indragiri Hilir and Dumai. In March 2005, outliers are mostly occurred in Dumai. Moreover, in June, July and August 2005 outliers spread in Riau Province.

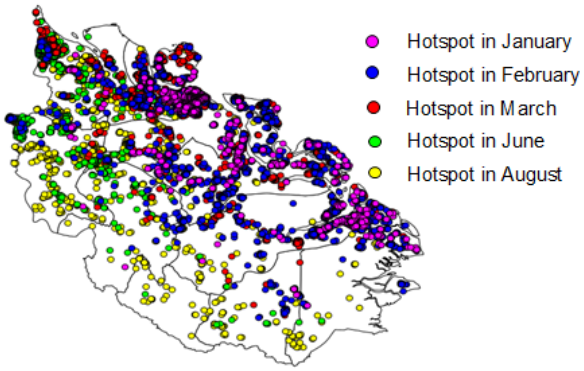

Figure-7. Outliers on hotspot in 2005 as results of K-means algorithm [18].

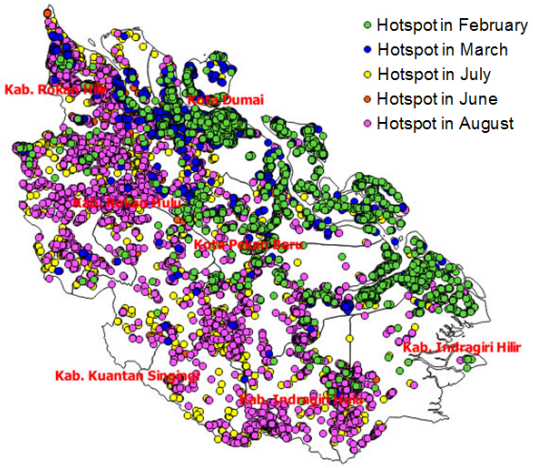

Figure-8. Outliers on hotspot in 2005 as results of PAM clustering algorithm.

4.2 Density-based local outliers detection

The experiments on hotspot dataset in the period of 2001-2014 were performed by determining the parameter k on the local outlier factor algorithm. The results shows that outlier score of hotspot tend to increase as the values of k decrease. Figure-9 provides outlier score of hotspots at the parameter k of 2, 3, 4, 5, and 6.

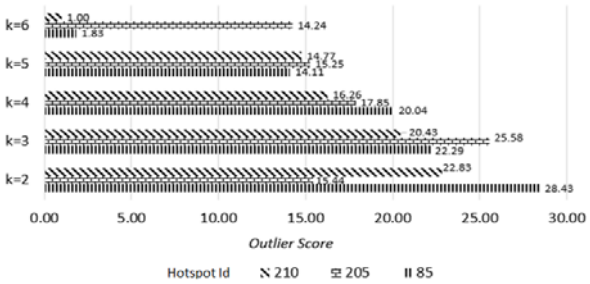

Figure-9. Outlier score at k= 2, 3, 4, 5, and 6.

Figure-10 provides number of outliers on hotspot datasets that are identified using the LOF algorithm at k of 2 in comparison to the results of the DBSCAN algorithm [14]. Both DBSCAN and LOF algorithms give the same trend of outliers frequency in the period of 2001-2014 in which the LOF algorithm detects more outliers that DBSCAN (Table-4). High number of outliers is found on the hotspot datasets in the years 2005, 2006, 2009, 2012.

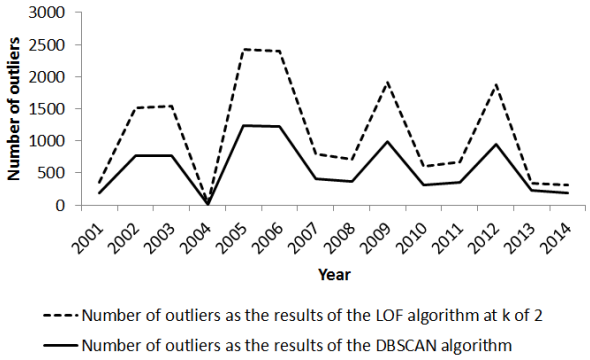

Figure-10. Number of outlier at k= 2, 3, 4, 5, and 6.

As many 1190 outliers spread in all districts in Riau province in 2005 (see Table-4) which are members of two clusters. Those outliers are mostly found in Dumai (228 outliers), Rokan Hilir (264 outliers), Bengkalis (265 outliers) and Pelalawan (163 outliers) as shown on Figure-11.

Table-3. Outliers in the period of 2001 - 2012, represented in Date (frequency).

Year	Jan	Feb	Mar	May	Jun	Jul	Aug	Sep	Oct
2001						**9(191)[1]**			
2002		15(224)[1] 17(190)[1]	8(304)[2] 10(247)[1] 12(354)[2]			14(279)[1]	15(227)[1] 20(228)[1]		
2003			4 (296)[2]		6 (341)[2] 8 (730)[2] 9(243)[1] 10 (676)[2]	31(250)[1]			
2004				20(191)[1]	15 (435)[2] 17 (683)[2] 19 (344)[2] 20 (328)[2] 22(254)[1] 23(191)[1]		6(259)[1] 11(337)[2]		
2005	20 (402)[2] 22 (434)[2] 24 (337)[2] 25 (623)[2]	1(217)[1] 5(205)[1] 7 (792)[2] 8(218)[1] 9 (438)[2] 10 (647)[2] 12 (652)[2] 14 (600)[2] 16 (406)[2] 19(216)[1] 21(1118)[2] 22(229)[1] 23(206)[1] 25(186)[1]	4 (335)[2] 5 (454)[2] 7 (648)[2] 9 (736)[2] 14(227)[1] 16 (511)[2] 18(275)1 20(420)[2] 21(222)[1]		20 (349)[1] 23(191)[2] 24 (557)[1] 25 (401)[1]		5(239)[1] 7 (956)[2] 9 (388)[2] 10 (328)[2] 16 (295)[2]		
2006		6(270)[1] 8(405)[2]	3 (383)[2] 5(254)[1]			4(196)[1] 16(589)[2] 25(414)[2]	5(596)[2] 6(386)[2] 8(235)[1] 17(354)[2] 21(205)[1]		4(232)[1] 6(210)[1]
2007		11(271)[1]				3(284)[2]			
2008		19 (527)[2] 21 (350)[2]		18(210)[1]			1(182)[1] 3(183)[1] 4(241)[1] 6(453)[2] 15(224)[1]		
2009	20(199)[1] 22(269)[1]	18(202)[1] 16(274)[1]			17 (400)[1] 19(180)[2] 20(230)[2]	4(241)[1] 13(214)[1] 17(205)[1] 24 (334)[2] 31(257)[1]	2 (428)[1] 4 (384)[1] 6 (305)[1] 7(202)[2]		
2010									15(593)[1] 17(498)[2]
2011		17(280)[1]		8(189)[1] 9 (453)[2]		1(729)[1] 5(242)[1] 12(306)[2] 21(183)[1]	1(729)[2]		
2012					14(462)[2] 16(244)[1] 18(209)[1] 23(266)[1] 25(181)[1]	10(761)[1] 27(275)[1] 30(245)[1]	8(306)[1] 10(761)[1] 12(287)[1] 14(270)[2]	4(321)[2]	

[1]PAM, [2] PAM and K-Means

As a comparison, Figure-12 shows the plot of outlier distribution in 2014. The LOF algorithm detects as many 119 outliers in the hotspot dataset in 2014 that is much lower than number of outliers in 2005. According to Figure-12, high number of outliers is also occurred in Dumai (21 outliers), Bengkalis (33 outliers), Pelalawan

(27 outliers), Rokan Hilir (28 outliers) and Siak (27 outliers).

Table-4. Number of outliers in the period of 2001-2014.

Year	Number of outliers as the results of DBSCAN [14]	Number of outliers as the results of LOF algorithm at k of 2
2001	188	164
2002	769	740
2003	769	768
2004	14	12
2005	1241	1190
2006	1229	1166
2007	409	380
2008	366	345
2009	990	917
2010	311	289
2011	349	323
2012	946	920
2013	228	117
2014	189	119

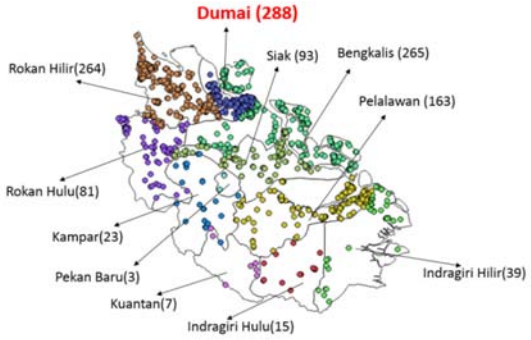

Figure-11. Outliers on hotspot data in Riau Province in 2005.

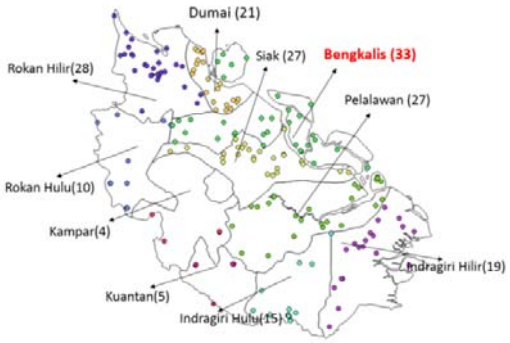

Figure-12. Outliers on hotspot data in Riau Province in 2014.

Outliers in 2005 and 2014 are mostly found in two periods namely February to March and June to August as shown in Figure-13. Riau province has two periods of dry seasons that are February to March and July to September. High number of hotspots and its outliers in those two periods of dry season indicates frequent forest and land fire events.

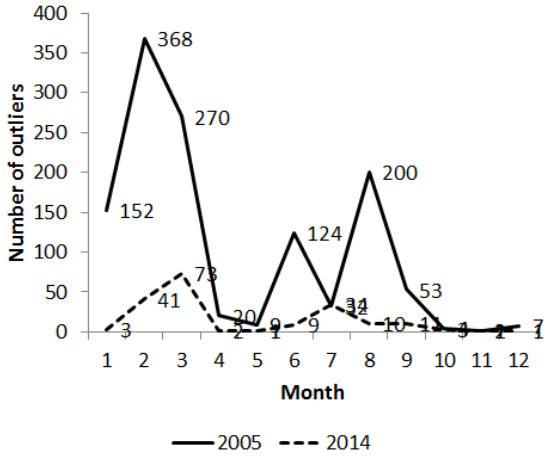

Figure-13. Number of monthly outliers in Riau Province in 2005 and 2014.

5. CONCLUSIONS

This work applied the PAM clustering and Local Outlier Factor (LOF) algorithm to identify outliers in the hotspot dataset in Riau Province Indonesia. Based on the PAM algorithm, this work identifies as many 123 outliers with the average of hotspots frequency that is considered as outliers is 351.11. The lowest frequency of hotspot occurrence as outlier is 180 and the highest frequency is 1118. Based on the clustering results of both K-means and PAM algorithm, outliers are mostly found in February, March, June, July, and August in the period of 2002-2012. The results shows that outliers are not occurred in April, November and December in the period of 2001-2012. The distribution of outliers which represent the high frequency of hotspot occurrences is essential for forest and land fires prevention.

The LOF algorithm identifies as high number of outliers in years 2005, 2006, 2009, and 2012 in which as many 1190 outliers are detected in 2005. Those outliers in 2015 cover almost all districts in Riau Province where the dense areas are Dumai, Rokan Hilir, Bengkalis and Pelalawan. Several outliers were again found in those districts in 2014. Outliers that are hotspots located far from the hotspot clusters may become the real forest and land fires. Therefore fire fighters and related parties should pay attentions to those areas where outliers are frequently occurred especially in the period of dry seasons namely February to March and June to August.

REFERENCES

[1] I.S. Sitanggang, R. Yaakob, N. Mustapha and A.N. Ainuddin. 2013. Classification model for hotspot occurrences using spatial decision tree algorithm. Journal of Computer Science. 9 (2): 244-251

[2] I.S. Sitanggang, R. Yaakob, N. Mustapha and A.N. Ainuddin. 2014. A decision tree based on spatial relationships for predicting hotspots in peatlands. TELKOMNIKA (Telecommunication Computing Electronics and Control). 12 (2): 511-518.

[3] I.D. Nurpratami and I.S Sitanggang. 2015. Classification rules for hotspot occurrences using spatial entropy based decision tree algorithm. Procedia Environmental Sciences. 24(2015: 120-126.

[4] A. Amri and I.S. Sitanggang. 2015. A geographic information system for hotspot occurrences classification in Riau Province Indonesia. Procedia Environmental Sciences. 24 (2015): 127-131.

[5] I S. Sitanggang. 2013. Spatial multidimensional association rules mining in forest fire data. Journal of Data Analysis and Information Processing. 1(4): 90-96.

[6] N. Arincy and I. S. Sitanggang. 2015. Association rules mining on forest fires data using FP-Growth and ECLAT algorithm. Proceedings of the3rd International Conference on Adaptive and Intelligent Agroindustry (ICAIA), Bogor, 2015: 274-277.

[7] T. Fuad, I.S. Sitanggang, and Annisa. 2010. K-Means clustering visualization of web-based OLAP operations for hotspot data. Proceedings of the International Symposium on Information Technology 2010 (ITSim 2010).

[8] M. Usman, I.S. Sitanggang, and L. Syaufina. 2015. Hotspot distribution analyses based on peat characteristics using density-based spatial clustering. Procedia Environmental Sciences. 24(2015): 132-140.

[9] A.P. Kirana, I.S. Sitanggang, and L. Syaufina L. 2015. Poisson clustering process on hotspot in peatland area using Kulldorff's Scan Statistics Method. TELKOMNIKA (Telecommunication Computing Electronics and Control). 13 (4): 1376-1383.

[10] V. J. Hodge and J. Austin. 2004. A survey of outlier detection methodologies. Artificial Intelligence Review. 22 (2): 85-126.

[11] J. Singh and S. Aggarwal. 2013. Survey on outlier detection in data mining. International Journal of Computer Applications. 67(19): 29-32.

[12] P.H. Thah and I.S. Sitanggang. 2015. Contextual outlier detection on hotspot data in Riau Province using K-Means algorithm. Procedia Environmental Sciences. 33(2016): 258-268.

[13] A.M.Y.A Suci, and I.S. Sitanggang. 2016. Web-Based application for outliers detection on hotspot data using K-Means algorithm and Shiny framework. IOP Conference Series: Earth and Environmental Science. 31(1): 1-8.

[14] P. Sukmasetya and I. S. Sitanggang. 2016. Outlier detection on hotspots data in Riau Province using DBSCAN algorithm. IOP Conference Series: Earth and Environmental Science. 31(1): 1-4.

[15] M. Ester, H. Kriegel, J. Sander, and X. Xu. 1996. A Density-based algorithm for discovering cluster in large spatial database with noise. Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining, AAAI Press. 1996: 226-231.

[16] M. Ankerst, M. M. Breunig, H. P. Kriegel and J. Sander. 1999. OPTICS: Ordering points to identify the clustering structure. ACM Press

[17] N. L. Febriana and I.S. Sitanggang. 2016. Outlier detection on hotspot data in Riau province using OPTICS algorithm. A paper presented in The 3rd International Seminar on Sciences, 3rd November 2016, IPB International Convention Center, Bogor, Indonesia

[18] I.S. Sitanggang and D.A.M. Baehaki, 2015. Global and collective outliers detection on hotspot data as forest fires indicator in Riau Province, Indonesia. Proceedings of the Second IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services. 2015: 66-70.

[19] J. Han, M. Kamber, and J Pei. 2012. Data Mining: Concepts and Techniques. Morgan Kaufmann, Massachusetts.

[20] P. Tan, V. Kumar, and M. Steinbach. 2006. Introduction to Data Mining. Pearson Education.

[21] M. Breunig, H. P. Kriegel, T. Ng. Raymond and J. Sander. 2000. LOF: Identifying density-based local outliers. Proceedings of the 2000 ACM Sigmod International Conference on Management of Data, ACM. 2000: 93-104.