Spatial temporal clustering for hotspot using kulldorff scan statistic method (KSS): A case in Riau Province

S A Hudjimartsu[1*], T Djatna[2], A Ambarwari[1] and Apriliantono[1]

[1] Department of Computer Science, Faculty of Math and Natural Sciences, Jl. Meranti Wing 20 Level 5 Kampus IPB, Babakan, Dramaga, Bogor, Jawa Barat
[2] Department Agroindustrial Technology, Faculty of Agricultural Technology, Gedung Fateta lt.2, Kampus IPB, Dramaga, Bogor, Jawa Barat

E-mail: shudjimartsu@gmail.com

Abstract. The forest fires in Indonesia occurs frequently in the dry season. Almost all the causes of forest fires are caused by the human activity itself. The impact of forest fires is the loss of biodiversity, pollution hazard and harm the economy of surrounding communities. To prevent fires required the method, one of them with spatial temporal clustering. Spatial temporal clustering formed grouping data so that the results of these groupings can be used as initial information on fire prevention. To analyze the fires, used hotspot data as early indicator of fire spot. Hotspot data consists of spatial and temporal dimensions can be processed using the Spatial Temporal Clustering with Kulldorff Scan Statistic (KSS). The result of this research is to the effectiveness of KSS method to cluster spatial hotspot in a case within Riau Province and produces two types of clusters, most cluster and secondary cluster. This cluster can be used as an early fire warning information.

1. Introduction

Recently the forest fires in Indonesia is an annual catastrophe that occurs when the dry season. About 61% of forest fires in Southeast Asia occurred in Indonesia in the period 1990-2010 [1]. Almost all of the causes of forest fires are caused by the human activity itself, among others forest fires and land clearing for plantations, sparks from plantation or forest and sabotage [2]. The impact of forest fires is the loss of biodiversity, pollution hazard and harm the economy of surrounding communities. Given the impact of the fires are very harmful and the factors causing the fire complex, it is important to develop early warning systems for the prevention of forest fires. Indication of land and forest fires can be known through the hotspots detected in a location at a certain time.

The hotspots area data derived from remote sensing satellite imagery. Hotspots data provide information such as location and time of occurrence, and as an early indicator of land and forest fires. To provide better information to assess that hotspots are a fire, there should be a method in processing. One of method is by clustering, grouping spatio temporal data to produce clustering that can be judged from grouping, the clustering can be regarded as a fire or not. Clustering method for spatio temporal data frequently used and adopted with KSS (Kulldorff Scan Statistic) [3]. KSS method was first introduce by Kulldorff et al. (1997) which is the result of the development of scan statistics method [4]. Other than that KSS method has been able to detect clustering of point in 3 dimensions, the dimension

spatial, the dimension of time (temporal) and as well as dimension of spatial and time (spatio temporal) [5].

2. Previous Work

Research on data processing hotspot has been done to prevent forest fires through the development of early warning systems. Sitanggang and Ismail (2010) observed hotspots data processing with classifying by decision tree method, using on algorithm j48 or C4.5 algorithms, the resulting in a ruleset that is used to predict the occurrence of hotspots based on parameters such as type of land, roads and river network [6]. In 2014 Sitanggang and Khoiriyah develop research in previously (2010), which connects between the spatial data in the data processing so as to generate spatial decision tree [7].

In addition to classification, data processing hotspot as an early warning system can also be done using clustering techniques. The study was conducted by Wulandari (2012), which applies the Dynamic Density Based Clustering Algorithm (DDBC) are known to be able to handle spatial and temporal data proposed [8]. Usman (2015) observed the clustering based on density for the distribution of hotspots as indicators of forest fires and peatland in Sumatra in 2002 and 2013. The method used for clustering is DBSCAN-in this study found 53 clusters in 2002 and 42 clusters in 2013 [9].

Annisa (2015) using Kuldorff Scan Statistics (KSS) the great method for detecting hotspots grouping based on spatial and temporal aspects. Clustering of hotspot using KSS method can detect where or when the occurrence grouping of hotspots and areas of distribution [10].

3. Methodology

This research was divide into three parts: area, tools and stages of research. Area of research in Riau Province with approximation of area 89,150.16 km^2 and has an area of peat land 40,087.76 km^2. The data that used are hotspots of the year 2010 – 2014, sourced from FIRM MODIS fire can downloaded at website http://earthdata.nasa.gov/data/nrt-data/firms. For peat land data sourced from ministry of environment and for administrative boundaries Riau Province sourced from geospatial information agency (BIG). Tools that used in this research are: R application, "SpatialEpi" and Quantum GIS. The stages are doing in this research consisted of six stages: collection and analysis data, preprocessing of data, implementation KSS method with Poisson model for determining likelihood, determine hotspots clustering with KSS method, cluster validation and visualization.

3.1. Collection Data

Hotspots data sourced from FIRM MODIS fire, which is data spatio temporal consist of location and time dimension. Dimensional location of the hotspot data is the latitude and longitude, and the time dimension is the date of occurrence of hotspots.

3.2. Preprocessing of Data

3.2.1. Preprocessing of hotspots data

Preprocessing on the hotspot data is divide into three phases:

a) Data Extraction
   At this phase, selected attribute from hotspot data which consist of latitude, longitude, and date only, the other attribute will be eliminated and divide data into two types (overlay/clip)
b) Selection of hotspots based on peatland
   This following result overlay (clip) with peatland.

Table 1. Number of hotspots in 2010-2014

Year	Before overlay (clip)	After overlay (clip)
2010	4,148	2,988
2011	6,864	4,827
2012	7,873	4,688
2013	15,103	10,948
2014	21,591	19,332

c) Selection distribution of hotspot locations by district
In the area of Riau Province grouping based on subdistrict boundary, so that the occurrence of hotspots can be found in each subdistrict.

3.3. Implementation of KSS method

The implementation of KSS method is based on the determination of the likelihood value on the scanning results, shape scanning of KSS method is a circular window. According to Wen (2009) that the radius of the circular window size varies from zero to a predetermined upper on any given centroid. Each circular window has a likelihood ratio λ (Z) which of each was calculate based on observed data and the probability model that has been chosen, in this research were taking Poisson models. Further, look for the maximum likelihood ratio and the candidate most likely cluster obtained, then the next stage is to test the significance of using a Monte Carlo approach.

3.4. Scanning Window

Clustering in KSS method based on scanning window with the phases as follows:

a) Random centroid/cell in research area.
b) Make a circle centered at centroid/cell of the closest neighborhood to the farthest distance and not exceed 50% of the population (Kulldorff et.al 1997).
c) Count the number of incident cases (the number of hotspot in an area) and population (the entire area of peatland) for each centroid neighborhood.
d) Calculating of the likelihood ratio of each pairwise.
e) Repeat steps 1 – 5 for all centroid in research area.

3.5. Determining of Hypothesis

Scanning window is used to detect hotspots, then grouped based on the average value of the hotspots in the scanning window. Let p be the average value of the occurrence of hotspots in the scanning window and q is the average value of the occurrence of hotspots outside the scanning window.

H0 : There is no clustering of hotspots in peatland (p = q).

H1 : There is clustering of hotspots in peatland (p > q) in circular scanning window (Z).

p = q can be interpreted by the number of case is same in every area/spread evenly. Meanwhile, p > q is defined as occurring grouping of hotspots in the scanning window.

3.6. Determining of Likelihood Ratio

Probability function, f(x) which represented the probability of the number of occurrences within a cell x described in equation 1 as follows (Kulldorff et.al 1997):

$$(x)=\begin{cases}\dfrac{e^{-p\mu(x)}(p\mu(x))^{n_x}}{n_{x!}}, x\in Z\\ \dfrac{e^{-p\mu(x)}(p\mu(x))^{n_x}}{n_{x!}}, x\notin Z\end{cases} \quad (1)$$

Where,
f(x) = Occurrence probability in cell x.
p = Average value of occurrence of hotspots in the scanning window.
q = Average value of occurrence of hotspots outside the scanning window.
μ(x) = Number of population in the cell x.
n_x = Number of cases in the cell x.
Z = circular scanning window.
e = 2.71828.

KSS method and Poisson models were used to compare the number of cases that are inside and outside the scanning window, which is used to find the hotspots. To calculate the likelihood ratio function $\lambda(Z)$ with Poisson models for each circular scanning window Z (Kulldorff 2014) can be seen in the following equation 2.

$$\lambda(Z)= \begin{cases} \left(\frac{n_z}{e_z}\right)^{n_z} \cdot \left(\frac{n_G-n_z}{n_G-e_z}\right)^{n_G-n_z}, \text{ if } n_z>e_z \\ 1, \text{ other} \end{cases} \quad (2)$$

Where,
n_z = Number of cases in the scanning window *Z*.
e_z = Expected of cases in the scanning window *Z*.
n_G = Number of cases in research area G.

In which expected value of cases e_z obtained from following equation 3.

$$e_z=\mu(Z)\left(\frac{n_G}{\mu_G}\right) \quad (3)$$

Where,
μ(Z) = Number of population in the scanning window *Z*.
n_G = Number of cases in the research area *G*.
μ(G) = Total of population in the research area *G*.

3.7. Cluster Validation

The results of the validation can be determined by calculating the statistical significance p-value or calculate using a Monte Carlo approach (Kulldorff et.al 1997). To calculate the p-value can use the following equation 4.

$$p=\frac{(T(x))\geq t_0}{m+1} \quad (4)$$

Where,
T(x) = The highest likelihood ratio value of data replication.
t_0 = The highest likelihood ratio value of real data.
m = Many simulations.

Monte Carlo calculations used to find the value of p that good, then the number of simulations is limited to values ending in 99, such as 1999, 2999, or 9999. To accept or reject H 0 is typically used alpha level value (α) by (α) = 0.1, (α) = 0.05, (α) = 0.01 and (α) = 0.001 (Kulldorff et.al 1997).

3.8. Cluster Analysis

At this stage, a cluster analysis is done based on physical characteristics of peat in Riau province, to get information on the peatland hotspot. The physical characteristics of peatland including the type and thickness of the peat land and land cover.

4. Results and Discussions

The results of the study the distribution of hotspots in the province of Riau, especially in peatland can be seen in the following figure.

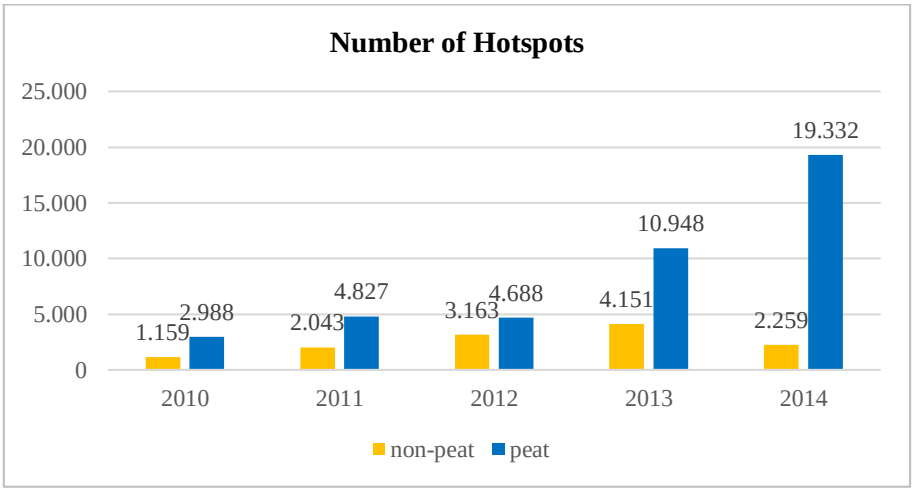

Figure 1. Number of hotspots in peat and non-peat Riau Province

Based on Figure 1, shows that the frequency of occurrence of hotspots in Riau particularly on peatland in the period 2010-2014 has increased.

Peatland is distinguished based on the type of maturity and thickness. Type maturity consists of Fibrists (immoldy), Hemists (half moldy), Saprists (molded) and the combination of these three types of maturity. Meanwhile, peatland based on the thickness divided into three: moderate depth, deep and very deep. The distribution of hot spots in the type of maturity peatlands can be seen in table 2.

Table 2. Distribution hotspots in type of maturity peatlands in 2010-2014

Type of maturity peatlands	2010	2011	2012	2013	2014	Average/year
Fibrists/Saprists (60/40),moderate depth	1	0	0	3	2	1.2
Hemists/min (30/70),shallow	11	22	52	95	33	42.6
Hemists/min (30/70), moderate depth	2	0	7	9	23	8.2
Hemists/min (90/10), moderate depth	0	1	2	0	1	0.8
Hemists/Saprists (60/40),deep	297	355	324	675	906	511.4
Hemists/Saprists (60/40),very deep	1,095	1,094	1,103	1,096	1,097	1,097
Hemists/Saprists (60/40), moderate depth	394	510	309	1,007	3,598	1,163.6
Saprists (100),deep	0	0	0	0	0	0
Saprists (100),moderate depth	68	178	121	131	303	160.2
Saprists/Hemists (60/40),deep	239	576	453	519	2894	936.2
Saprists/Hemists (60/40),very deep	620	900	1,011	1,546	3,315	1,478.4
Saprists/Hemists (60/40), moderate depth	121	125	75	153	413	177.4
Saprists/min (50/50),shallow	2	15	17	78	42	30.8
Saprists/min (50/50), moderate depth	30	77	104	470	143	164.8
Saprists/min (90/10), moderate depth	109	418	336	380	1106	469.8

From the type of peat maturity, which occurs on the type Hemists hotspot / Saprists (60/40) is very deep, Saprists / Hemists (60/40) is very deep, and Hemists / Saprists (60/40) deep enough. While the thickness of the peat is divided into five, which consists of:

1) Peaty soil, < 50cm (D0)
2) Shallow, 50-100cm (D1)
3) Moderate, 100-200cm (D2)
4) Deep, 200-400cm (D3)
5) Very deep, >400cm (D4)

Table 3. Distribution of hotspots in thickness of peat in 2010-2014

Thickness of peat	2010	2011	2012	2013	2014	Average/year
D0	2	22	24	9	9	13.2
D1	134	213	167	376	709	319.8
D2	796	1398	1087	2363	5933	2,315.4
D3	701	1440	1706	3345	4449	2,328.2
D4	1356	1767	1706	4863	8243	3,587

From Table 3, hotspot often occurs at the level of D4 (very deep, >400cm), D3 (in, 200-400cm) and D2 (Moderate, 100-200cm). This indicates that the thicker or more in peat, then the frequency of occurrence of a more frequent hotspot. Aside from factors peat, hotspots can occur because of land cover in Riau Province.

Table 4. Distribution of hotspots in cover in 2010-2014

Land Cover	2010	2011	2012	2013	2014	Average/year
Thicket swamp	296	396	327	932	2407	871.6
Swamp forest	2,091	3,444	3,653	8,530	13,718	6,287.2
Rubber plantation	22	24	20	66	74	41.2
Coconut on a former swamp forest> 5 years	199	334	260	416	1,553	552.4
Palm on former swamp forest <5 years	148	174	203	203	496	244.8
Palm on former swamp forest >5 years	77	158	69	323	204	166.2
Forest land concessions	36	164	59	199	191	129.8
Land cultivation of industrial plants	1	0	9	3	3	3.2
Plantation preparation	10	16	18	13	239	59.2
Rice fields and coconut	103	108	51	250	425	187.4
Intensive rice fields	0	0	0	0	0	0
Shrubs and marsh grasses and former fire	5	17	12	16	29	15.8
Shrubs, grass on a former rice field	1	5	9	5	4	4.8

Distribution of hot spots based on land cover with high frequency occurs in swamp forest in the period 2010 to 2014, with an annual average of about 6,287.2 hotspots.

4.1. Spatial Clustering with KSS Method

Results of clustering the hotspot data using KSS method within period from 2010 to 2014, obtained cluster consisting of two classes cluster composed of most cluster and the secondary cluster.

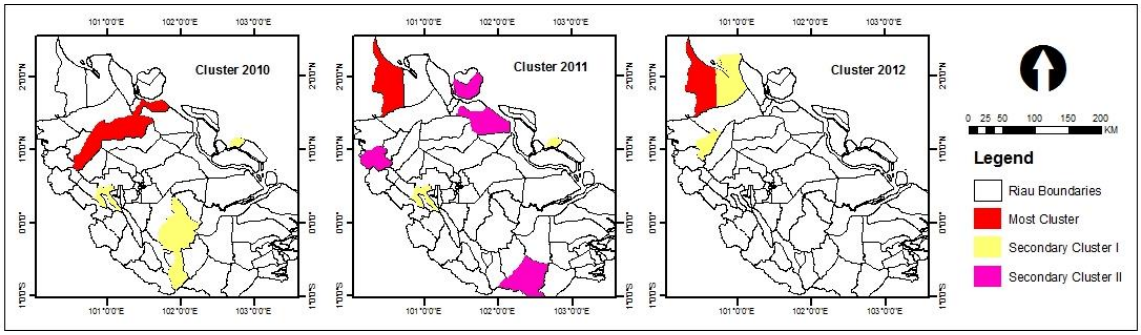

Figure 2. Hotspots clustering with KSS method in 2010 – 2012

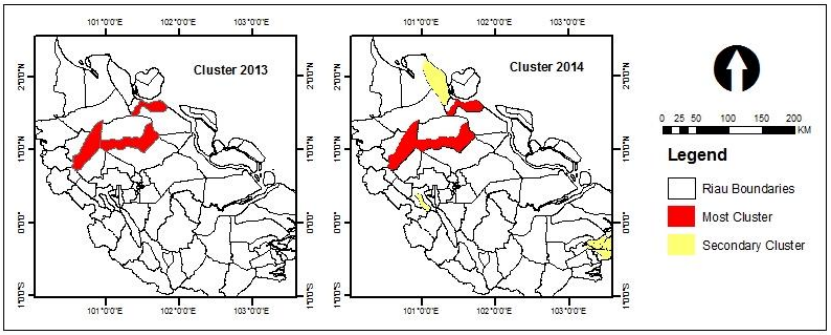

Figure 3. Hotspots clustering with KSS method in 2013 and 2014

From Figure 2 in 2010 occurred grouping with the largest density value is 0.498 hotspot/km^2 in the Bukit Kapu subdistrict. Whereas in 2010 and 2011, the grouping also occurs in the same district, started Sine districts with a density of 0,265 hotspot/km^2 and 0.185 hotspot/km^2. In 2013 the grouping occurred in the district of Pinggir with a density of 0.767 hotspot/km^2 and 2014 occurred in the district of Dumai west with a density value of 0.943 hotspot/km^2.

Referring to Usman (2014), that using DBSCAN in 2013 the density of hotspots highest on peatlands of Sumatra was found in the province of Riau, with a density value of 0.056 km^2, where most hotspots are in Rokan Hilir, Bengkalis, Dumai and Siak District. Compared with KSS method, the results are almost the same, because some districts which formed part of the subdistrict.

5. Conclusions

Based on the results obtained, it can be concluded:

1. Kulldorf Scan Statistic (KSS) is the good methods to analyze data with the spatial and temporal dimensions, such as the hotspot data.
2. In our case, the subdistrict that has the highest frequency of occurrence of hotspots is a subdistrict of Bukit Kapu, Bagan Sine and Dumai.
3. Based on the type of the maturity and thickness of the peat, the emergence of hotspots dominated by Hemists/Saprists (60/40), with the thickness of the peat is deep and very deep. While on land cover, hotspot dominated by peat swamp forests.

References

[1] Stibig H J, Achard F, Carboni S and Raši R and Miettinen R 2013 Change in tropical forest cover of southeast asia from 1990 to 2010 *Biogeosciences* **11** 247–258
[2] Syaufina L 2008 Kebakaran hutan dan lahan di Indonesia *Malang Bayumedia Publishing*
[3] Wen S and Kedem B 2009 A semiparametric cluster detection method a comprehensive power comparison with kulldorff âA˘Zs method *International Journal of Health Geographics BioMed Central USA*
[4] Naus J l 1965 Clustering of random points in two dimensions *Biometrika **52** 263-267*

[5] Kulldorff M 1997 A spatial scan statistic *Communications in Statistics: Theory and Methods* **26** 1481–1496
[6] Sitanggang L S and Ismail M H 2010 Hotspot occurrences classification using decision tree method: Case study in the rokan hilir, Riau province, Indonesia In *Proc. Eighth Int. Conf. ICT and Knowledge Engineering* 46–50
[7] Khoiriyah Y M and Sitanggang I S 2014 A spatial decision tree based on topological relationships for classifying hotspot occurences in bengkalis riau Indonesia In *Proc. Int Advanced Computer Science and Information Systems (ICACSIS) Conf* 268–272
[8] Wulandari F 2012 Penerapan dynamic density based clustering pada data kebakaran hutan [Undergraduate Thesis] Bogor (ID): Bogor Agricultural University
[9] Usman M, Sitanggang I S and Syaufina L 2015 Hotspot distribution analyses based on peat characteristics using density-based spatial clustering *The 1st International Symposium on LAPAN IPB Satellite for Food Security and Environmental Monitoring, Procedia Environmental Sciences*
[10] Kirana, Annisa P, Sitanggang I S and Syaufina L 2016 Hotspot pattern distribution in peat land area in Sumatera based on spatio temporal clustering *Procedia Environmental Sciences* **33** 635-645