Clustering Haze Trajectory of Peatland Fires in Riau Province using K-Means Algorithm

H Khairat[1], I Sukaesih Sitanggang[1]* and D E Nuryanto[2]

[1]Department of Computer Science, Bogor Agriculture University, Indonesia.
[2]Meteorology Climatology and Geophysics Board, Indonesia.

E-mail: hamidatul.khairat@gmail.com, imas.sitanggang@ipb.ac.id, danang.eko@bmkg.go.id

Abstract. Haze is one of the impacts caused by forest and land fires. In order to determine how far the movement of haze from forest and land fires including in peatland, this research aims to determine the haze trajectory patterns from peatland fires in Riau Province in 2015 using HYSPLIT model. Furthermore, the haze trajectory patterns were grouped using the K-Means algorithm. The results show that the trajectory consists of 4887 positions of haze that moves to the northeast and northwest. The clustering results using K-Means show that the largest cluster has 3702 positions of haze and these haze movements are located at the average of height of 23.554 m AGL. Those haze positions were located in Riau, North Sumatra, Malaysia, and Malacca Strait.

1. Introduction

Peatland area has significant roles for biodiversity conservation and climate regulation, and provides important support for human welfare. The Province of Riau (Indonesia) has high deforestation rate in recent years and the majority of the deforestation has occurred on the peat soil. Deforestation and fire on the peatland area resulted in the increasing of CO and CO_2 emissions. The existence of haze is one of the impacts caused by forest and land fires, especially when it happens in peatland area. Haze caused by peatland fires will be blown away by the wind and can lead to a wide variety of health problems for the surroundinghuman. Therefore, peatland fire prevention is an important activity to reduce the damage and losses. In order to minimize the negative impacts of peatland fires, trajectory or movement pattern of haze as the results of peatland fires need to be identified and analyzed.

This work aims to analyze the haze trajectory of the peatland fires in the Province of Riau by implementing Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT) model and K-Means algorithm. HYSPLIT model is designed to support a variety of simulations related to atmospheric transport and dispersion of pollutants [1]. K-Means clustering is one of methods that splits the data into several clusters [2]. In this work, HYSPLIT is used to model the meteorological data from peatland fires in Riau Province which aims to get the haze trajectory pattern. Furthermore, the haze trajectory patterns are grouped based on the movement direction of haze using K-Means algorithm. Based on the results, it is expected to be known how far the movement of haze from peatland fires.

2. Data and Method

2.1. Data

The data used in this research are the meteorological data in 2015. Data were obtained from the National Oceanic and Atmospheric Administration (NOAA), which can be accessed at this URL http://www.ready.noaa.gov/ready2-bin /extract/extracta.pl/. We focus the study to the area of Sumatra Island and its surrounding. The meteorological data are weekly data consisting of complex meteorological parameters which are stored in a bin-format file. The dataset includes the variables such as temperature (°C), relative humidity (%), wind direction (degree) and wind speed (m/s).

In addition to meteorological data, this work also utilized the sequence data of hotspots in Riau Province during 2015 of which occurrence's duration of 2 days. These sequences are obtained from the previous research [3], and consists of 86 hotspots. Attributes that are available in the hotspot sequences are the coordinates of the hotspots which are the longitude and the latitude.

Meteorological data in Sumatra were selected from global meteorological data. The hotspots in sequences are used as the initial location of haze trajectory. Meteorological data are selected based on locations of hotspots that occurred in the sequence. The selection was done by matching the date of hotspot occurrences with weekly meteorological data.

2.2. Determination of haze trajectory using HYSPLIT

Trajectory patterns contain information about spatial and temporal movements of objects. The spatial information indicates locations of points passed-through during the movement, while the temporal information indicates the time required for the movement from one point to another. The trajectory shows displacement or transition from one point to another in each transition time [4].

Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT) is a model used to calculate the simple trajectory of air parcels to complex transport, dispersion, chemical transformation and deposition simulation. The initial development of HYSPLIT is a joint effort between NOAA and the Australian Bureau of Meteorology. HYSPLIT model is designed to support a variety of simulations related to atmospheric transport and dispersion of pollutants. Calculation method applies Lagrangian and Euler approaches. Lagrangian approach uses a moving object as a reference for calculating of air diffusion moved from their initial locations. Euler approach uses a 3-dimensional grid as a frame of reference for calculating the concentration of air pollutants [1].

2.3. Clustering haze trajectory using K-Means algorithm

Clustering is a process of grouping a set of objects into several clusters or groups so that objects in the same cluster have a high degree of similarity, however the objects are very different from objects in other clusters. Dissimilarities and similarities of the objects are calculated based on the attribute values that describe the objects. There are various methods that can be used to measure similarity value between object, one of them is the Euclidean distance [2].

K-Means clustering is a method that partitions data into one or more cluster(s). This method is efficient in large data processing with fast computation time [2]. The steps in K-Means algorithm are as follows:

1. Determine the value of k as the number of clusters to be formed along with the centre point of the cluster (centroid).
2. Calculate the distance of each object to each centroid using Euclidean distance.
3. Assign each object by the shortest distance between objects and the centroid.
4. Determine the new centroid by calculating the average value of the objects that exist in the same cluster.
5. Perform steps 2 through 4 until the centroid does not change.

In this work, clustering of haze trajectory patterns was done using K-Means algorithm. Attributes used in clustering are longitude and latitude coordinates of haze trajectory, as well as the height of

haze positions. K-Means algorithm is available on 'fpc' package in R. Clustering results are analysed to determine how far the movement of haze in each cluster and regions affected by haze.

3. Results and Discussions

3.1. Data Selection
The hotspot sequence data consist of 6 sequence periods from July to October 2015. There are 86 hotspots in the sequences. Table 1 shows the selected hotspot coordinates of each sequence period of hotspots that are used as initial points of trajectory. Meteorological data were selected based on hotspot sequence data. Table 2 shows periods of meteorological data that are used in this study.

Table 1. Selected hotspot coordinates in sequences in 2015

Sequence period of hotspots	Longitude	Latitude	Duration (day)	District
July 9th – July 11th	100.445	2.080	3	Rokan Hilir
July 22nd – July 23rd	100.550	1.845	2	Rokan Hilir
July 26th – July 29th	102.662	-0.583	4	Indragiri Hulu
August 30th – September 1st	102.103	0.660	3	Siak
September 1st – September 2nd	101.281	0.432	2	Kampar
October 21st – October 22nd	102.493	-0.279	2	Indragiri Hulu

Table 2. Period of meteorological data based on the hotspot sequences

Sequence period of hotspots		Meteorological data
First day	Last day	
July 9th 2015	July 11th 2015	gdas1.jul15.w2
July 22nd 2015	July 23rd 2015	gdas1.jul15.w4
July 26th 2015	July 29th 2015	gdas1.jul15.w4 and gdas1.jul15.w5
August 30th 2015	September 1st 2015	gdas1.aug15.w5 and gdas1.sep15.w1
September 1st 2015	September 2nd 2015	gdas1.sep15.w1
October 21st 2015	October 22nd 2015	gdas1.oct15.w3 and gdas1.oct15.w4

3.2. Determination of Haze Trajectory using HYSPLIT
Haze trajectory patterns were generated using HYSPLIT with the input including meteorological data and hotspot sequences. Attributes of the hotspot sequence are the coordinates of hotspots that are used as starting points of haze trajectory. In this work, the hotspot sequence consists of 6 day periods of hotspot occurrences. Table 3 shows values of parameters for HYSPLIT in determining haze trajectory.

Table 3. Parameters for HYSPLIT to determine haze trajectory

Parameter	Value
Starting time	Initial date the sequence period of hotspots
Starting locations	Longitude and latitude coordinates of hotspots in the sequence
Height	10 m AGL
Total run time	Duration of hotspot occurrences (hour)
Direction	Forward trajectory
Top of model	10 km AGL
GIS-output	GIS-point
Vertical coordinate	m AGL (Above Ground Level)

The first step to get the haze trajectory in HYSPLIT is setting trajectory parameters. Trajectory parameters include time and location of the initial movement of haze, duration of simulation, and height of the haze source. Time, duration, and location of the initial movement of haze are customized based on the sequence data of hotspots. Height of the haze source depends on the vegetation in the location. This work assumed a tree with a height of 10 meters as height of the haze source. In addition, it is assumed that maximum height of haze (top of model) is the troposphere with the height of 10 km.

The results of haze trajectory in the period from July 9th until July 11th 2015 are shown in Figure 1. The haze trajectory contains time information with haze trajectory duration in hour, locations of haze trajectory, and height of haze in unit of m AGL (Above Ground Level). Figure 1 shows that in the period from July 9th until July 11th 2015 starting at 00:00 UTC or 07:00 am there were 18 trajectories or haze movements. The results show that the haze trajectory moves to the northeast toward Malaysia. The height of haze is between 0 m AGL and 1500 m AGL.
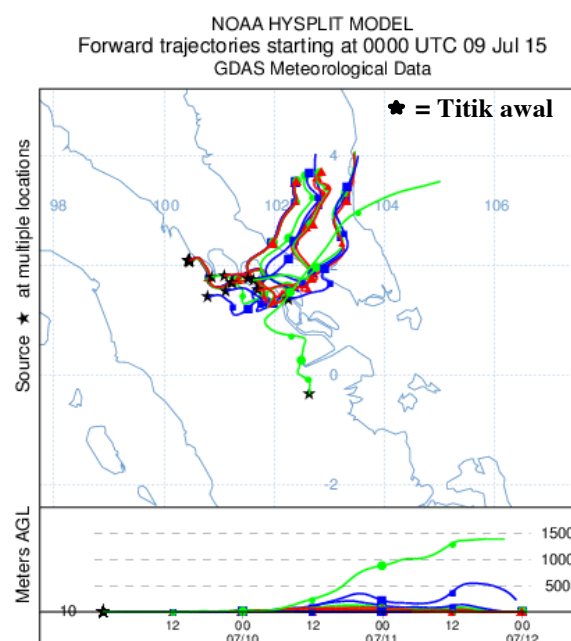

Figure 1. Haze trajectory in the period of July 9th to July 11th 2015

The haze trajectory in the period from July 9th until July 11th 2015 has 1241 positions of haze movements in which a maximum height of haze is 1395 m AGL. The entire period of the hotspot sequence in 2015 consists of 12 to 18 hotspots as initial points of haze trajectory. In this period, as many 4887 positions were obtained with the movement directions of haze to the northeast and northwest. Table 4 shows a summary of haze trajectory in 2015.

Table 4. Summary of haze trajectory in 2015

Date of haze trajectory	Movement direction of haze trajectory	Height level of haze (m AGL)	Height of haze in average (m AGL)
July 9th – July 11th	Johor and Pahang (Malaysia)	1395	54.817
July 22nd – July 23rd	Riau and North Sumatra	615	73.125
July 26th – July 29th	Nanggroe Aceh Darussalam	1300	387.189
August 30th – September 1st	North Sumatra and Malacca Strait	1070	103.221
September 1st – September 2nd	Riau and North Sumatra	665	57.286
October 21st – October 22nd	Riau	282	34.912

3.3. Clustering Haze Trajectory using K-Means Algorithm

Clustering haze trajectory was performed on the datasets consisting of latitude and longitude of haze trajectory and its height. Clustering was done using K-Means function that is available on 'fpc' package in R. In this research, number of clusters (k) was selected from 2 to 10 and the clustering results were evaluated by calculating the Sum Square Error (SSE). Figure 2 shows the SSE value and accuracy for each cluster.
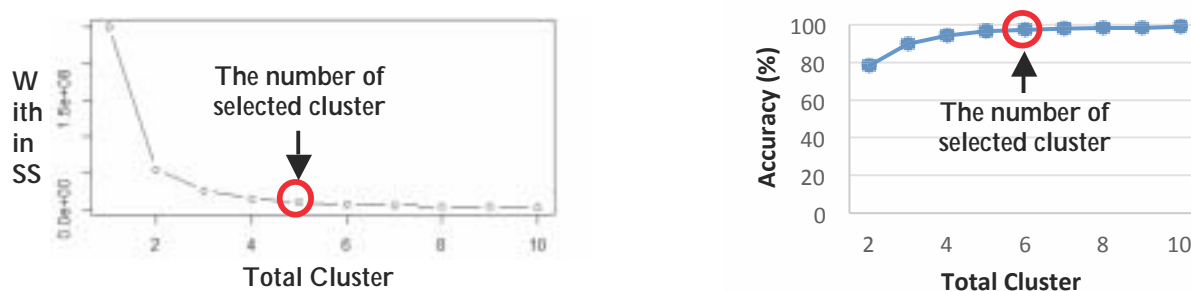

Figure 2. SSE value and clustering accuracy

The value of number of clusters (k) was selected such that percentage of clustering quality is high enough to represent the clustering results with the smallest value of SSE. Based on the experimental results, k of 5 was selected to cluster haze trajectory. The accuracy of clustering reaches 96.4%.

Table 5 shows the clustering results of haze trajectory at k of 5. Cluster 1 has the largest number of members which are positions of haze movement. As many 3702 positions (75.75% of overall positions) of haze trajectory are members of cluster 1. Cluster 2 has the smallest number of members which contains 89 positions of haze trajectory or about 1.82% of all data.

Table 5. Results of haze trajectory clustering with k of 5

Cluster	Number of members	Average (Center)			Percentage (%)
		Longitude	Latitude	Height (m AGL)	
1	3702	101.699	1.317	23.554	75.75
2	89	100.444	2.707	1138.831	1.82
3	169	99.210	2.821	797.639	3.46
4	287	99.886	2.342	467.498	5.87
5	640	100.744	1.598	180.673	13.10

Figure 3 shows a plot of five haze trajectory clusters. Locations of haze spread in Sumatra and a point of haze is located outside the territory of Sumatra that is around the Strait of Malacca and Malaysia.
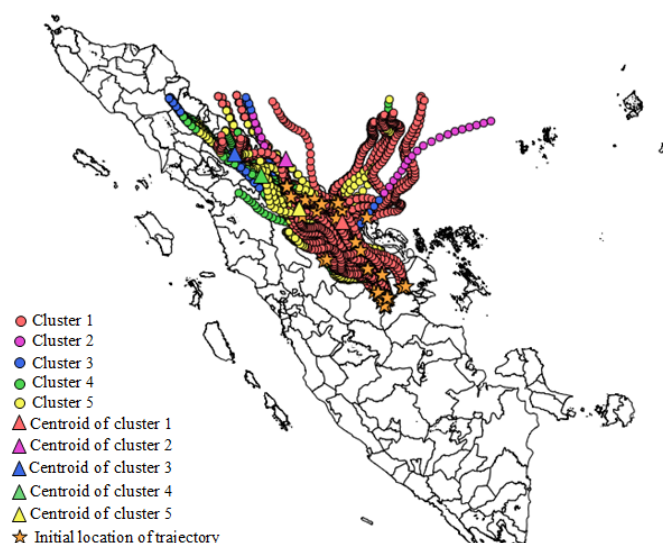
Figure 3. Haze trajectory clusters in Riau Province in 2015

Haze trajectory clusters were analysed based on height of haze in average on each cluster. According to Table 5, cluster 2 has the smallest number of haze trajectory positions with the average of haze height is 1138.831 m AGL. Haze positions move from the source of fire at height of 10 m AGL to 1138.831 m AGL. Cluster 1 has 3702 positions of haze trajectory in which the lowest height value in average is 23.554 m AGL. In this situation, the haze is dangerous for human health. The haze positions are located in 10 districts and cities in Riau province including Bengkalis, Indragiri Hilir, Indragiri Hulu, Kampar, Pelalawan, Rokan Hilir, Rokan Hulu, Siak, Pekan Baru City, and Dumai City, 4 districts in North Sumatra namely Asahan, Labuhan Batu, Serdang Bedagai, and Simalungun, 3 states in Malaysia namely Johor, Negeri Sembilan and Pahang. In addition, haze spread around the Strait of Malacca.

4. Summary
This study generated haze trajectory from peatland fires. The trajectory pattern mining results as many 4887 movement positions of haze that move towards the northeast and northwest of Riau Province. The results of haze trajectory clustering show that the largest cluster contains as many 3702 locations of haze movements in hotspot sequence period in which the lowest height of haze in average is 23.554 m AGL. The haze movements are located in 10 districts and cities in Riau Province, 4 districts in North Sumatra, 3 states in Malaysia, as well as around the Strait of Malacca.

References
[1]   Stein A F, Draxler R R, Rolph G D, Stunder B J B, Cohen M D and Ngan F 2015 *Bull Amer Meteor Soc*. **96** 2059-2077
[2]   Han J, Kamber M and Pei J 2011 *Data Mining: Concepts and Techniques* Ed 3 (US: Morgan Kaufmann)
[3]   Abriantini G, Sitanggang I S, and Trisminingsih R 2016 Hotspot Sequential Pattern Visualization in Peatland of Sumatra and Kalimantan using Shiny Framework. *The 3rd International Symposium on LAPAN – IPB SATELLITE* (LISAT), October 25-26, 2016, Bogor, Indonesia.
[4]   Giannotti F, Nanni M, Pedreschi D and Pinelli F 2007 *Trajectory Pattern Mining* (Italy: University of Pisa)