Wildfire Risk Map Based on DBSCAN Clustering and Cluster Density Evaluation

Muchamad Taufiq Anwar, Wiwien Hadikurniawati, Edy Winarno, Aji Supriyanto

Faculty of Information Technology, Universitas Stikubank Semarang, Jl. Tri Lomba Juang No 1 Semarang 50241, Central Java, Indonesia

taufiq@edu.unisbank.ac.id

Abstract. Wildfire risk analysis can be based on historical data of fire hotspot occurrence. Traditional wildfire risk analyses often rely on the use of administrative or grid polygons which has their own limitations. This research aims to develop a wildfire risk map by implementing DBSCAN clustering method to identify areas with wildfire risk based on historical data of wildfire hotspot occurrence points. The risk ranks for each area/cluster were then ranked/calculated based on the cluster density. The result showed that this method is capable of detecting major clusters/areas with their respective wildfire risk and that the majority of consequent fire occurrences were repeated inside the identified clusters/areas.

Keywords: wildfire risk map; clustering; DBSCAN; cluster density;

1. Introduction

Wildfire is one of the most notable disasters occurred around the world. Wildfires had caused a large amount of economic losses and environmental damaged. Preventive actions/managements are required in order to minimize the negative effect caused by wildfire. One of the main tasks in managing wildfire is the detection of the areas which have high wildfire risk. The determination of area with a high risk of wildfire or any geographical phenomena (*i.e.,* the spatial 'hotspot') often relied on the use of administrative or grid polygons which has their own limitations, as mentioned by Han and Shu [1]. Therefore, the detection/determination of high fire risk areas should implement 'unsupervised' methods that are independent of administrative or grid polygons. One of the available options is by implementing a clustering algorithm to detect clusters of points and to set boundaries for each cluster. This research aims to develop a wildfire risk map by implementing DBSCAN clustering method to identify areas with wildfire risk based on historical data of wildfire hotspot occurrence points. Risk ranks for each identified area/cluster were then ranked/calculated based on their density (number of hotspot points per km square). The resulted wildfire risk map is useful in wildfire management so that preventive/mitigating action can be done to minimize losses or other negative effects.

2. Methods

2.1. Wildfire Risk Assessment

Research often used historical data of wildfire to determine wildfire risk in certain areas. Recent research had explored the risk of fire spreading to the urban area in Australia [2]. Other research used

generalized additive model to estimate wildfire risk in Mediterranean area [3], and used artificial neural networks to predict wildfire risk in South Africa [4]. The historical data can be obtained from satellite images such as from NASA [5]. The risk model might be based on several variables related to wildfire, such as the meteorological variables [6-8], vegetation indices [6,9], and other variables [6,7,9]; or based only on historical wildfire data [10]. When relying only on historical wildfire data, assessments can be based on the number of or the location of wildfire hotspot occurrences. A recent reseach also had explored a rule-based approach to wildfire model based on historical wildfire data[11].

The identification of area with high fire occurrence/risk (the spatial 'hotspot') can also be done using spatial analysis such as the Anselin's Moran's I or LISA [12] and Getis-Ord Gi [13]. This approach is sensible since most geographical phenomena often show a spatial pattern (i.e. clustered) [14], including wildfire itself [15,16]. Even the evidence of spatial autocorrelation itself can be included in the model to make it better [17]. However, these analyses often use administrative polygons or grid polygons and require a high number of polygons to be used in a study area in order for the analysis to give a good result. Furthermore, the use of polygons itself also has its own drawbacks, namely the scale mismatch, shape mismatch, and location mismatch as mentioned in [1].

2.2. DBSCAN Clustering

Density-Based Clustering of Applications with Noise (DBSCAN) is a modern clustering algorithm which has the capability to detect cluster in spatial points data (coordinates) and not relying on polygons. DBSCAN was first introduced by Ester, et.al in 1996 [18]. DBSCAN has two parameters namely eps which represent maximum distance/radius from a point where the membership of other points enclosed within it is evaluated; and minpts which tell the minimum number of points to be considered as a member of a cluster within the radius of eps. DBSCAN works by giving each point a circle with the radius of determined eps, followed by the membership evaluation for each point enclosed in that circle. A point would fall into one of three categories, namely a core point, border point, or noise points. A point is assigned as core points if it has at least a number of member (enclosed) point equal to minpts within radius eps. A point is assigned as border points when a point is within eps but has a number of member points of less than minpts. Finally, if the point doesn't belong to the core or border points, it is assigned as noise points. Noise points are not a member of any clusters. A cluster is then defined as a set membership containing a combination of core points surrounded by border points. The pseudocode for the DBSCAN algorithm is presented in Algorithm 1 [19].

Algorithm 1: *DBSCAN Clustering*

```
1:  DBSCAN(D, eps, MinPts)
2:      C = 0
3:      For each unvisited point P in dataset D
4:          mark P as visited
5:          N = getNeighbors (P, eps)
6:          If sizeof(N) < MinPts then
7:              mark P as NOISE
8:          Else
9:              C = next cluster
10:             expandCluster(P, N, C, eps, MinPts)
11:         End If
12:     End For
13:
14: expandCluster(P, N, C, eps, MinPts)
15:         add P to cluster C
16:     For each point P' in N
17:         If P' is not visited
18:             mark P' as visited
19:                 N' = getNeighbors(P', eps)
20:             If sizeof(N') >= MinPts
21:                 N = N joined with N'
```

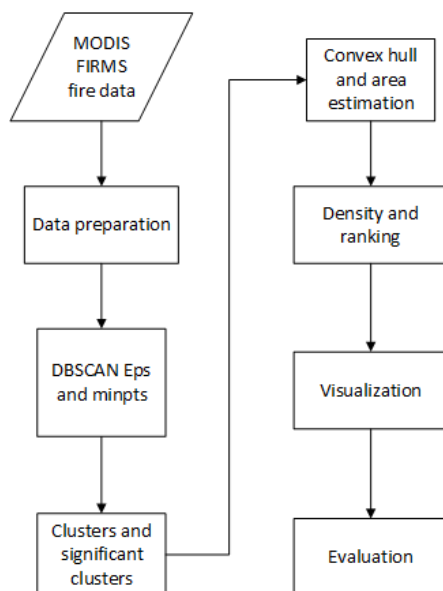|	
22:	**End If**
23:	**End If**
24:	**If** P' is not yet member of any cluster
25:	add P' to cluster C
26:	**End If**
27:	**End For**

The research framework for this paper is shown in Figure 1. For the purpose of this research, DBSCAN is used since it has advantages over other clustering algorithms such as the K-Means clustering and Scan statistics. First, unlike the K-Means, DBSCAN does not require us to set a determined number of cluster k. This determination of k, of course, is not suitable for this research, since the number of identified clusters is merely a result of density evaluation of points in the data, *i.e.,* an "unsupervised" algorithm. Furthermore, DBSCAN can detect clusters with arbitrary shapes, while K-Means and Scan statistics only evaluate and produce clusters of circular shape. This circular shape, of course, is not suitable to be implemented on geographical data since geographical phenomena come in many shapes. Recent use of DBSCAN in geographical clustering includes retail agglomerations [20], and location recommendation in location-based social networks [21]. Recent research also applied the DBSCAN to detect clusters in raster images [22]. Despite years of proposed improvements, the original DBSCAN could still perform well as long as we pick a reasonable parameter [23].

Figure 1. Research framework.

In order to calculate the density of a cluster, we need to estimate the area of each cluster, which first we need to determine the boundary of the clusters. One of the available options is the convex hull algorithm. In mathematics, a convex hull is the smallest set (area) which allows any two points in that set to be connected without leaving the set. It is well understood using the analogy of a stretched rubber band which encloses a set of points (*i.e.*, a cluster) [24]. Wildfire data were collected from the Moderate Resolution Imaging Spectroradiometer (MODIS) provided by the Fire Information for Resource Management System (FIRMS). This data is a result of automatic detection by NASA MODIS satellite for the anomaly of temperature/fire presence. For the purpose of this research, only data with 100 confidence were used. The data of the national wildfire archive were then filtered/clipped only for the study area.

DBSCAN clustering algorithm was done in R using 'dbscan' function in the 'dbscan' package. The determination of eps was first based on visual examination of 'knee' in the K-NN distance plot provided by 'kNNdistplot' function, provided in the same package. The experimentations of eps and minpts were done in multiple datasets to test their performance. The datasets were varied in the time and spatial scope. The best and final eps and minpts value to be used in this research were then picked based on the experimentation result. The DBSCAN algorithm resulted in points membership assigned as a member of a cluster or as a noise. Noise points were removed before the creation of convex hull polygons. Each polygon has the attribute of its area (in degree square). This attribute is automatically generated by the Convex Hull function in QGIS. These areas in degree square were then converted into the areas in kilometer square (at the equator, one degree of longitude/latitude is about 110.57 km). Our study area is located at the equator, spanning from latitude 02° 25' N to 01° 15' S. Clusters density were then calculated by dividing the number of points in each area/cluster by the cluster's area (in km square).

The risk ranks for each area/cluster were then ranked/calculated based on the density of the clusters. The rank is relative, where clusters with a density higher than the mean density are classified as a high

fire risk area, while clusters with a density lower than the mean density are classified as a medium fire risk area. The low fire risk area then is the area outside the identified clusters. The resulted clusters with their corresponding fire risk ranks were then visualized using QGIS. The resulted clusters were tested using the test dataset from the past two years. The tools used in this research are RStudio and QGIS 2.18.

3. Results and Discussion

The study area for this researh is Riau Province in Sumatra Island, Indonesia. Wildfire data from 2001 data 2017 were filtered only those with 100 confidence and were split into two datasets, 21252 training dataset has which ranged from 2001 to 2015, and 326 test dataset which ranged from 2016 to 2017. The wildfire locations of training dataset is shown in Figure 2a. The training dataset was run in the DBSCAN algorithm with eps of 0.02 degrees (~2.21 km) and minpts = 5. These values were based on experimentation which gave good clustering results. The DBSCAN algorithm in R package 'dbscan' resulted in 211 clusters with a total membership of 16142 points (76% of fire case data). Minor clusters that have historical fire counts less than 120 were then excluded. The 120 limits were estimated by the judgment of one fire case in 8 fire-months for 15 years. Convex hull polygons were created using QGIS function which also calculated the area of each cluster. The areas of each cluster (in kilometer square) were then calculated by multiplying the area (in degree square) with 110.57 km square (1 degree in equator ~ 110.57 km). Fire density for each cluster was then calculated by dividing the number of fire points in a cluster with the area (in kilometer square) of the corresponding cluster. The previous minor clusters exclusion had also eliminated low-density cluster with a density lower than 0,5 (which translated into less than one fire in an area of 2 km square). The result is 22 "significant" / major clusters which have high historical fire count and usually have a large area. These clusters were then assigned fire risk rank/label based on their density. The risk ranks were assigned based on their density compared to their mean. A density lower than the mean would be assigned as an area of medium fire risk, while a density higher than the mean is assigned as an area of high fire risk. The low fire risk areas are the areas which are not included in the final (major) clusters. This resulted in l4 medium fire risk clusters and 8 high fire risk clusters. One of the clusters in the high fire risk category has very high density and deemed as an outlier from other clusters. This cluster was then assigned as a very high fire density. The resulted clusters and their respective fire ranks are shown in Figure 2b. The area with a very high fire risk is located at the near top right of the study area.

Figure 2. (a) Locations of fire hotspot within the study area from 2001 to 2015. (b) Resulted clusters with their color-coded risk ranks. (c) Recent wildfire cases overlayed with the resulted clusters.

The resulted clusters were then tested against fire data from the last two years (2016-2017) which have 326 fire points as shown in Figure 2c. Layer overlay and intersection we used to calculate the number of points within the resulted clusters and resulted in 229 of 326 or 70% fire point fall within final significant clusters. The detailed information for each cluster is shown in Table 1. The resulted clusters might reveal similarity in wildfire variables such as meteorological factors, vegetation, or land

type (peat or non-peat). From the spatial hotspot analysis view, each cluster resulted from the DBSCAN algorithm may also represent a hotspot (High-High) surrounded by 'warm' spots (High-Low). The hotspots are analogous to the core points, while the warm spots are analogous to border points.

Table 1. The resulted clusters and their attributes.

Cluster ID	Number of points (Training dataset)	Area (km square)	Density (Points per km square)	Risk Rank	Number of points (Test dataset)
1	4846	1043,89	4,64	Very High	45
2	1536	533,96	2,88	High	1
3	206	97,33	2,12	High	1
4	1249	603,16	2,07	High	10
5	194	110,46	1,76	High	15
6	839	513,08	1,64	High	0
7	162	109,70	1,48	High	0
8	164	111,10	1,48	High	10
9	371	262,43	1,41	Medium	4
10	279	200,49	1,39	Medium	0
11	214	169,80	1,26	Medium	4
12	366	301,52	1,21	Medium	0
13	1683	1388,49	1,21	Medium	24
14	834	691,11	1,21	Medium	19
15	173	152,54	1,13	Medium	0
16	256	270,24	0,95	Medium	0
17	168	219,24	0,77	Medium	37
18	583	761,50	0,77	Medium	15
19	689	966,70	0,71	Medium	26
20	164	243,95	0,67	Medium	11
21	244	369,38	0,66	Medium	0
22	475	930,61	0,51	Medium	6
	15695				**228**

4. Conclusion and Future Research

Wildfire risk assessment can be based on historical data of fire hotspot occurrence. This research built a model for fire-risk prone areas using DBSCAN clustering applied on historical wildfire data from 2001 to 2015 and followed by density evaluation to determine risk rank. The resulted clusters were then tested against recent wildfire data from the last two years, and it showed that 70% of the recent wildfire occurrence falls within the resulted cluster, which indicates good performance. The identified cluster calls the authorities to perform wildfire ignition prevention and other mitigation actions [25].

However, the convex hull algorithm used in this research has its own limitation in estimating area since it cannot create proper cluster boundary in complex cluster shapes. In that case, other cluster boundary estimation such as concave hull or buffer distance might work better, but some parameter tuning might be needed. Area estimation (in kilometer) also need to be reconsidered in a study area far

from the equator. Future research might also try to explore other or richer risk-ranking method. The determination of eps and minpts in DBSCAN, and also the risk ranking method could also be based on experts' judgment or against a certain standard from the authority.

References
[1] Han, H., & Shu, X. (2017). A Self-Adjusting Approach to Identify Hotspots. International Review for Spatial Planning and Sustainable Development, 5(2), 104-112.
[2] Price, O., Borah, R., Bradstock, R., & Penman, T. (2015). An empirical wildfire risk analysis: the probability of a fire spreading to the urban interface in Sydney, Australia. International Journal of Wildland Fire, 24(5), 597-606.
[3] Ager, A. A., Preisler, H. K., Arca, B., Spano, D., & Salis, M. (2014). Wildfire risk estimation in the Mediterranean area. Environmetrics, 25(6), 384-396.
[4] Lall, S., & Mathibela, B. (2016, December). The application of artificial neural networks for wildfire risk prediction. In 2016 International Conference on Robotics and Automation for Humanitarian Applications (RAHA) (pp. 1-6). IEEE.
[5] Farahmand, A., Reager, J. T., Behrangi, A., Stavros, E. N., & Randerson, J. T. (2017, December). Using NASA Satellite Observations to Map Wildfire Risk in the United States for Allocation of Fire Management Resources. In AGU Fall Meeting Abstracts.
[6] Nami, M. H., Jaafari, A., Fallah, M., & Nabiuni, S. (2018). Spatial prediction of wildfire probability in the Hyrcanian ecoregion using evidential belief function model and GIS. International journal of environmental science and technology, 15(2), 373-384.
[7] Guo, F., Su, Z., Wang, G., Sun, L., Lin, F., & Liu, A. (2016). Wildfire ignition in the forests of southeast China: Identifying drivers and spatial distribution to predict wildfire likelihood. Applied Geography, 66, 12-21.
[8] Mitri, G., Jazi, M., & McWethy, D. (2015). Assessment of wildfire risk in Lebanon using geographic object-based image analysis. Photogrammetric Engineering & Remote Sensing, 81(6), 499-506.
[9] Amalina, P., Prasetyo, L. B., & Rushayati, S. B. (2016). Forest Fire Vulnerability Mapping in Way Kambas National Park. Procedia Environmental Sciences, 33, 239-252.
[10] Nisa, K. K., Andrianto, H. A., & Mardhiyyah, R. (2014, October). Hotspot clustering using DBSCAN algorithm and shiny web framework. In 2014 International Conference on Advanced Computer Science and Information System (pp. 129-132). IEEE.
[11] M. T. Anwar, H. D. Pumomo, S. Y. J. Prasetyo, and K. D. Hartomo, "Decision Tree Learning Approach To Wildfire Modeling on Peat and Non-Peat Land in Riau Province," in *2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2018, pp. 409–415.
[12] Anselin, L. (1995). Local indicators of spatial association—LISA. Geographical analysis, 27(2), 93-115.
[13] Getis, A., & ORD, J. (1992). The analysis of spatial association by use of distance statistics Geographical Analysis 24 (3): 189–206.
[14] Griffith, D., & Chun, Y. (2016). Spatial autocorrelation and uncertainty associated with remotely-sensed data. Remote Sensing, 8(7), 535.
[15] Chou, Y. H., Minnich, R. A., Salazar, L. A., Power, J. D., & Dezzani, R. J. (1990). Spatial autocorrelation of wildfire distribution in the Idyllwild quadrangle, San Jacinto Mountain, California. Photogrammetric Engineering and Remote Sensing, 56(11), 1507-1513.
[16] Koutsias, N., Allgöwer, B., & Conedera, M. (2002, November). What is common in wildland fire occurrence in Greece and Switzerland?–Statistics to study fire occurrence pattern. In Proceedings of the 4th International Conference on Forest Fire Research (pp. 18-23).
[17] Portier, J., Gauthier, S., Robitaille, A., & Bergeron, Y. (2018). Accounting for spatial autocorrelation improves the estimation of climate, physical environment and vegetation's effects on boreal forest's burn rates. Landscape ecology, 33(1), 19-34.
[18] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for

discovering clusters in large spatial databases with noise. In Kdd (Vol. 96, No. 34, pp. 226-231).

[19] Vijayalaksmi, S., & Punithavalli, M. (2012). A fast approach to clustering datasets using dbscan and pruning algorithms. International Journal of Computer Applications, 60(14).

[20] Pavlis, M., Dolega, L., & Singleton, A. (2018). A Modified DBSCAN Clustering Method to Estimate Retail Center Extent. Geographical Analysis, 50(2), 141-161.

[21] Yuan, Z., & Li, H. (2016, June). Location recommendation algorithm based on temporal and geographical similarity in location-based social networks. In 2016 12th World Congress on Intelligent Control and Automation (WCICA) (pp. 1697-1702). IEEE.

[22] Trifonov, G. M., Zhizhin, M. N., Melnikov, D. V., & Poyda, A. A. (2017). VIIRS Nightfire Remote Sensing Volcanoes. Procedia computer science, 119, 307-314.

[23] Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Transactions on Database Systems (TODS), 42(3), 19.

[24] Berg, M. D., Cheong, O., Kreveld, M. V., & Overmars, M. (2008). Computational geometry: algorithms and applications. Springer-Verlag TELOS.

[25] Calkin, D. E., Cohen, J. D., Finney, M. A., & Thompson, M. P. (2014). How risk management can prevent future wildfire disasters in the wildland-urban interface. Proceedings of the National Academy of Sciences, 111(2), 746-751.