A Spatial Decision Tree based on Topological Relationships for Classifying Hotspot Occurences in Bengkalis Riau Indonesia

Yaumil Miss Khoiriyah and Imas Sukaesih Sitanggang
Computer Science Department, Bogor Agricultural University, Bogor 16680, Indonesia
k.yaumil@gmail.com, imas.sitanggang@ipb.ac.id

***Abstract*—Forest fires in Riau province Indonesia, are frequently occurred every year especially in dry seasons. Hotspot is an indicator for forest fire events. Hotspots monitoring is an activity to prevent forest fires. Hotspot data are spatial data that are represented in points. In order to analyze the data, spatial algorithms are required. The extended spatial ID3 algorithm is a spatial classification algorithm for creating a spatial decision tree from spatial datasets. This research applied the extended spatial ID3 algorithm on the forest fires data in Bengkalis district, Riau province Indonesia. The data include hotspots and non-hotspots, weather data, socio-economic data, and geographical characteristics of the study area. The result of this research is a decision tree with the income source layer as the label of root node. As many 137 classification rules were generated from the tree. The accuracy of the tree is 75.66% on the forest fires dataset in Bengkalis district, Riau province.**

Keywords: *forest fires, hotspots, ID3, spatial decision tree*

I. Introduction

Riau province in Indonesia is known as the area with high case of forest fire events. In 1997, peat fires in Riau province caused smoke that spread to neighboring countries such as Singapura dan Malaysia [1]. Forest fires in Riau province are frequently occured every year until nowadays, especially in dry seasons.

Developing forest fires risk model is essential to minimize damages due forest fires. Several studies have been conducted in developing forest fire risk models by integrating Geographic Information Systems (GISs) and remote sensing. Some of the models are a forest fire risk model for the area of Sasamba in East Kalimantan Indonesia that was developed using the GIS and Complete Mapping Analysis (CMA) [2], a model of forest fire hazard in East Kalimantan, Indonesia built by applying remote sensing technology and GIS [3], and a forest fires risk model for West Kutai District in East Kalimantan Province, Indonesia that was developed using a GIS, remote sensing and Multi-criteria Analysis (MCA) [4].

In addition to dryland, fire risk models have been also developed for peatlands in order to minimize the incidence of forest fires in peatlands. Reference [5] created a model of peat fire risk in the District of Bengkalis, Riau Province Indonesia based on hotspot distribution, environmental and infrastructure aspects, using the method of Complete Mapping of Analysis (CMA).

Hotspot is an indicator for forest fires, detected by MODIS instrument on NASA's satellite. Each hotspot detection represents the center of a 1 km (approx.) pixel flagged as containing one or more fires, or other thermal anomalies (such as volcanoes) [6]. The hotspot data are spatial data. In order to analyze the data, spatial algorithms are required.

Reference [7] developed the ID3 spatial algorithm that can be applied to spatial data represented in polygons. Another research resulted the extended spatial ID3 algorithm that can work on spatial dataset including three types of spatial features namely point, line and polygon [8].

Applying classification algorithms in forest fires data is important to predict the possibility of hotspot occurences as indicators of forest fires events in a certain area. Reference [8] applied the extended spatial ID3 algorithm to the forest fires data in Rokan Hilir district in Riau Province, Indonesia. The result is a spatial decision tree for hotspot occurrences classification of the area. This research aims to apply the extended spatial ID3 algorithm [8] to forest fires data in a new area namely Bengkalis district, Riau province Indonesia in order to evaluate the performance of the algorithm in a new spatial dataset. The forest fire dataset consists of the target objects and the explanatory objects. The target objects are hotspots in 2008 and non hotspots whereas the

 978-1-4799-8075-8/14/$31.00

explanatory objects include physical objects, weather data, and socio-economics data. The decision tree can be used to predict hotspot occurences based on the characteristics of the area .

II. EXTENDED SPATIAL ID3 ALGORITHM

A. Spatial Relationship

Spatial data represents real objects based on geographical reference of the earth. The objects are represented by geometry such as point, line, polygon, and pixel [9]. Objects in spatial data have spatial relationships with its neighbors. Two types of the spatial relationships are topology and metric. Topology is a spatial relationship that concern about the geometry shape. It uses Boolean notation like OR for union and AND for subset of the shape [10]. Whereas metric is a spatial relationship that concern about distance relation [11].

In a spatial database, spatial objects are stored in layers in which objects have the same geometry type in a layer whether it is point, line or polygon. Spatial objects in a layer are characterized by spatial and non spatial attributes.

B. Extended Spatial ID3 Algorithm

To create decision trees using extended spatial ID3 algorithm [8], the algorithm is applied to the training dataset. The input dataset consists a set of explanatory layers which contain spatial features of explanatory objects and one target layer containing target objects.

The first step in the algorithm is determining spatial relations between two distinct layers. Let L_i dan L_j, $i \neq j$, for each featur r_i R=SpatRel(L_i, L_j), spatial measure for r_i is SpatMes(r_i). The spatial measure can be based on topology like area and count function or metric relation such distance. Spatial measure of R defined as in (1) [8]:

$$\text{SpatMes}(R) = f(\text{SpatMes}(r_1), \text{SpatMes}(r_2), .., \text{SpatMes}(r_n)) \quad (1)$$

Where r_i is a feature in R, i = 1, 2, …, n is number of feature in R, f is an aggregate function such as max, min, or sum.

A spatial relationship applied to L_i and L_j results a new layer, R. SJR (*Spatial Join Relation*) is defined for all features p in L_i and q in L_j as (2) [8]:

$$\text{SJR}=\{(p, \text{SpatMes}(r), q \mid r \text{ is feature in R associated with } p \text{ and } q\} \quad (2)$$

SpatMes is used to calculate spatial entropy. Let the target layer S has a target attribute C with l distinct classes (c_1, c_2, .., c_l). Spatial entropy for S represents the infomation needed to determine class for all dataset, as defined in (3) [8]:

$$H(S) = -\sum_{i=1}^{l} \frac{\text{SpatMes}(Sci)}{\text{SpatMes}(S)} \log_2 \frac{\text{SpatMes}(Sci)}{\text{SpatMes}(S)} \quad (3)$$

SpatMes(S) represents spatial measure of the layer S as defined in Equation 1.

Let V is the attribute of explanatory layer that has q distinct values (v_1, v_2, .. , v_q). For each value v_i associated with the target layer S, a new layer L(v_i, S) is produced. The expected entropy value for splitting is defined in (4) [8]:

$$H(S|L) = -\sum_{j=1}^{q} \frac{\text{SpatMes}(L(v_i,S))}{\text{SpatMes}(S)} H\big(L(v_j,S)\big) \quad (4)$$

H(S|L) represents amount of information needed to classified the objects based on explanatory layers.

The next step in the algorithm is to calculate spatial information gain as defined in (5) [8]:

$$\text{Gain}(L) = H(S) - H(S|L) \quad (5)$$

The layer with the highest value of spatial information gain will become the root node of the decision tree and it is selected to be a splitting layer .

III. MATERIALS AND METHODS

A. Study Area

The study area is Bengkalis district in Riau Province in Indonesia (**Fig. 1**). Bengkalis district has 7,793.93 km^2 [12].

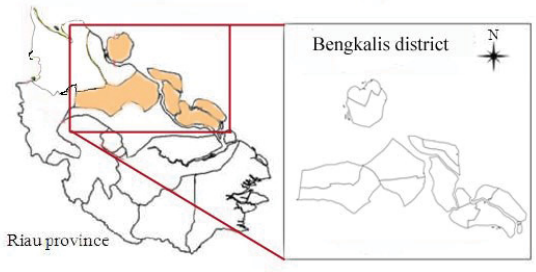

Fig. 1. Area of study

According to [12], Bengkalis district has the following borders:

1. The north side: Malaka strait;
2. The south side: Siak district and Meranti islands district;
3. The west side: Dumai city, Rokan Hilir and Rokan Hulu district, and;
4. The east side: Meranti islands district;

B. Dataset and Tools

The dataset contains spread and coordinates of hotspots in 2008 that was obtained from FIRMS MODIS Fire/Hotspot, NASA/University of Maryland, weather data of 2008 including precipitation (mm/day), screen temperature (K), and wind speed (m/s) are gathered from Meteorological Climatologically and Geophysical Agency (BMKG) Indonesia, socio-economic data of income source are collected from Statistics-Indonesia (BPS), and physical data (city centers, river, road, land cover) gathered from Geospatial Information Agency-Indonesia (BIG). Number of features in each spatial objects is given in **TABLE I**.

 978-1-4799-8075-8/14/$31.00

TABLE I
NUMBER OF FEATURES IN SPATIAL OBJECTS

Data	Spatial objects	Number of features
Physical	City center	17 points
	River	948 lines
	Road	44 lines
	Land cover	3026 polygons
Socio-economic	Income source	175 polygons
Weather	Precipitation	49 polygons
	Screen temperature	27 polygons
	Wind speed	41 polygons
Hotspot	Hotspot	685 points

Tools that used in this research are Windows 8 64-bit operating system, Quantum GIS 1.8.0-Lisboa for processing and visualization spatial data, PostgreSQL 1.16.0 as database management system, PostGIS 2.0 as the spatial extension for PostgreSQL, and the programming language Python 2.0 .

C. Data Praprocessing

1) Creating target layer

The target layer contains hotspots and non hotspot points as target objects for classifying a certain location into True class (hotspot occurrence) and False class (non hotspot occurrence). To prepare the target layer, as many 685 random points as non hotspots were generated near hotspots. The non hotspot points are located at least 0.907374 away from any hotspot data [13].

The input of the extended spatial ID3 algorithm is a dataset consisting of a target layer and explanatory layers. The explanatory layers are physical layers, socio-economic layers, and weather layers. Physical factors include in this work are city center, river, road, and land cover. The socio-economic factor is income source. Whereas weather aspects including precipitation, screen temperature, and wind speed data.

2) Distance calculating for explanatory layers of city center, river, and road objects

City center, river, and road objects are represented by points and lines. Thus, the metric operation can be applied to get the measure of objects in target layer to objects in explanatory layers. The aggregate function *min* is used to get the nearest distance of objects in target layer to each explanatory layers.

3) Validity testing of layers with polygon features

The explanatory layers land cover, income source, precipitation, temperature, and wind speed contains polygon features. Validity testing of the polygon features were applied in order to implement the aggregate function *sum* for all objects in the target layer that located inside the polygon features.

Data preprocessing results 8 explanatory layers and one target layer. Number of features in each layer is given in TABLE II.

TABLE II
LAYERS IN DATABASE

Layer name	Number of features	Number of distinct values
Distance to nearest city center in km	1370 points	3 (*low, medium, high*)
Distance to nearest river in km	1370 points	3 (*low, medium, high*)
Distance to nearest road in km	1370 points	3 (*low, medium, high*)
Income source	175 polygons	9 (*plantation*, *services*, dll)
Land cover	2937 polygons	12 (*plantation*, *swamp*, dll)
Precipitation in mm/day	49 polygons	5 (0, 1, 2, 3, 4)
Screen temperature in K	27 polygons	3 (297, 298, 299)
Wind speed in m/s	40 polygons	5 (0, 1, 2, 3, 4)
Target	1370 points	2 (*true, false*)

IV. Results and Discussion

The extended spatial ID3 algorithm has been implemented in Python [8]. The main Python moduleis to create decision tree. The input of this module are a list of explanatory layers and a target layer.

The *create decision tree* module work by choosing the best layer based on spatial gain information. The explanatory layer that became the best layer at the first iteration would be the root node of the decision tree. Each distinct value of the explanatory attribute in the best layer is assigned as a label of edge that connects the root node to an internal node The *create decision tree* module was run on the forest fire dataset and it resulted a spatial decision tree containing 137 leaf nodes with the income source layer as the first best splitting layer. Figure 2 shows a subtree of the resulted tree.

The decision tree was evaluated using two testing datasets: a forest fires dataset for Rokan Hilir district [8] and a forest fires dataset for Bengkalis district.

The accuracy of the decision tree on the Rokan Hilir dataset is 41.38% in which 192 objects of 464 objects in the dataset are correctly classified by the tree. The accuracy of the decision tree on the Bengkalis dataset is 75.66%, in which 404 objects of 534 objects in the dataset are correctly classified by the tree. Figure 3 and Figure 4 show the confusion matrix respectively for the Rokan Hilir and the Bengkalis testing set.

There are 137 rules generated from the decision tree for classifying hotspot occurences. The following 10 rules are considered as strong rules because these rules are supported by majority of spatial objects in the dataset.

Rule 1: IF *income_source* = *plantation* AND *land_cover* = *plantation* AND 1 mm/day ≤ *precipitation* < 2 mm/day AND 0 m/s ≤ *wind_speed* < 1 m/s AND 297 K ≤ *screen_temp* < 298 K AND

 978-1-4799-8075-8/14/$31.00

dist_road > 5 km AND *dist_river* ≤ 1.5 km, THEN *hotspot occurence* = **True**.
Total objects: 1315 points.

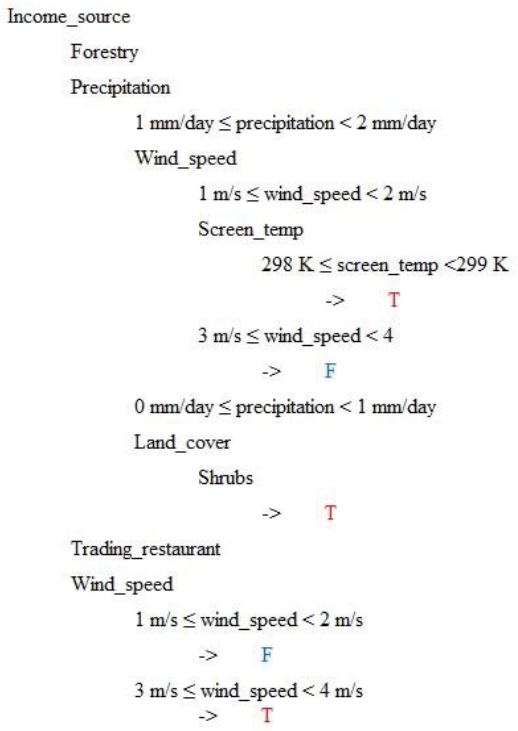

Fig. 2. Subtree

	True	False	Total
True	14	177	191
False	95	178	273
Total	109	355	464

Fig. 3. Confusion matrix for the testing data in Rokan Hilir

	True	False	Total
True	44	110	154
False	20	360	380
Total	64	470	534

Fig. 4. Confusion matrix for the testing data in Bengkalis

Rule 2: IF *income_source* = *plantation* AND *land_cover* = *plantation* AND 1 mm/day ≤ *precipitation* < 2 mm/day AND 1 m/s ≤ *wind_speed* < 2 m/s, THEN *hotspot occurence* = **True**.
Total objects: 954 points.

Rule 3: IF *income_source* = *plantation* AND *land_cover* = *plantation* AND 1 mm/day ≤ *precipitation* < 2 mm/day AND 1 m/s ≤ *wind_speed* < 2 m/s AND 297 K ≤ *screen_temp* < 298 K AND *dist_road* > 5 km AND *dist_river* > 3 km, THEN *hotspot occurence* = **True**.
Total objects: 557 points.

Rule 4: IF *income_source* = *plantation* AND *land_cover* = *plantation* AND 1 mm/day ≤ *precipitation* < 2 mm/day AND 3 m/s ≤ *wind_speed* < 4 m/s, THEN *hotspot occurence* = **False**.
Total objects: 304 points.

Rule 5: IF *income_source* = *plantation* AND *land_cover* = *plantation* AND 1 mm/day ≤ *precipitation* < 2 mm/day AND 3 m/s ≤ *wind_speed* < 4 m/s AND *dist_road* > 5 km AND *dist_river* > 3 km, THEN *hotspot occurence* = **True**.
Total objects: 140 points.

Rule 6: IF *income_source* = *plantation* AND *land_cover* = *plantation* AND *precipitation* ≥ 4 mm/day AND 1 m/s ≤ *wind_speed* < 2 m/s AND 297 K ≤ *screen_temp* < 298 K AND *dist_road* > 5 km AND *dist_river* ≤ 1.5 km, THEN *hotspot occurence* = **True**.
Total objects: 138 points.

Rule 7: IF *income_source* = *forestry* AND 3 mm/day ≤ *precipitation* ≤ 4 mm/day AND 1 m/s ≤ *wind_speed* < 2 m/s AND 297 K ≤ *screen_temp* < 298 K, THEN *hotspot occurence* = **False**.
Total objects: 69 points.

Rule 8: IF *income_source* = *plantation* AND *land_cover* = *plantation* AND 3 mm/day ≤ *precipitation* < 4 mm/day AND 1 m/s ≤ *wind_speed* < 2 m/s AND 297 K ≤ *screen_temp* < 298 K AND *dist_road* ≤ 2.5 km AND *dist_river* ≤ 1.5 km, THEN *hotspot occurence* = **True**.
Total objects: 52 points.

Rule 9: IF *income_source* = *plantation* AND *land_cover* = *plantation* AND *precipitation* ≥ 4 mm/day AND 1 m/s ≤ *wind_speed* < 2 m/s AND 297 K ≤ *screen_temp* < 298 K AND *dist_road* > 5 km AND 1.5 km < *dist_river* ≤ 3 km, THEN *hotspot occurence* = **True**.
Total objects: 52 points.

Rule 10: IF *income_source* = *forestry* AND 1 mm/day ≤ *precipitation* ≤ 2 mm/day AND 2 m/s ≤ *wind_speed* < 3 m/s AND 298 K ≤ *screen_temp* < 299 K, THEN *hotspot occurence* = **False**.
Total objects: 40 points.

Rule 1 shows that a hotspot is probably occurred in an area that has the following characteristics: 1) the income source of people living in the area is plantation, 2) the land cover type of the area is plantation, 3) precipitation in the area ranges from 1 to 2 mm/days, 4) temperature is in the interval of 297-298 °K, and 5) distance of the area to nearest road is around 1.5-5 km.

The decision tree shows that most potential areas with hotspot occurences in Bengkalis are the areas with plantation as the main income source and the land cover type. In addition, the such areas have the weather conditions as follows: precipitation of 1-2 mm/day, and wind speed of 0-1 m/s or 1-2 m/s . These conditions are supported by the rule number 1, 2, and

 978-1-4799-8075-8/14/$31.00

3 in which as many the 2286 objects in the dataset are associated with these rules. The rule 4 predicts that no hotspot in the area with 3-4 m/s of wind speed. This rule is supported by 304 objects with false class. If the wind speed in the area has value of 3-4 m/s, a further observation on distance of the area to the nearest river or road is required to classify hotspot occurrences in the area.

V.CONCLUSION

Developing a predictive model for hotspot occurences is essential to prevent the damage of forest fire events. This research has applied the extended spatial ID3 algorithm to the forest fires dataset in Bengkalis district, Riau province Indonesia. The results is a spatial decision tree model for classifying hotspot occurences in the study area. The spatial data include physical data (city center, river, road, land cover), weather data (wind speed, preciptation, temperature), and socio-economic data (income sorce). To evaluate the decision tree, the forest fires dataset in Rokan Hilir and Bengkalis were used. The accuracy of the tree on the Rokan Hilir dataset is 41.38% and for the Bengkalis dataset is 75.66%. As many 137 rules were generated from the tree in which the first factor to be tested for classification is income source. According to the tree, most of hotspots are probably occurred in plantation where the weather conditions as follows: precipitation of 1-2 mm/day, and wind speed of 0-1 m/s or 1-2 m/s.

REFERENCES

[1] L. Tacconi, *Kebakaran Hutan di Indonesia: Penyebab, Biaya, dan Implikasi Kebijakan.* Bogor: Center For International Forestry Research, 2003, ch 1.

[2] J. Boonyanuphap, F.G. Suratmo, I.N.S. Jaya, and F. Amhir, "GIS-based method in developing wildfire risk model (case study in Sasamba, East Kalimantan, Indonesia)," *Trop. For. Manage. J.*, vol. 7, no. 2, pp. 33—45, 2001.

[3] M. Darmawan, M. Aniya and S. Tsuyuki, "Forest fire hazard model using remote sensing and geographic information systems: Toward understanding of land and forest degradation in lowland areas of East Kalimantan, Indonesia," presented at the 2001 22nd Asian Conference on Remote Sensing, Singapore.

[4] P.H. Danan, "A RS/GIS–Based Multi-Criteria Approaches to Assess Forest Fire Hazard in Indonesia (Case Study: West Kutai District, East Kalimantan Province)," M.Sc. thesis, Info. Tech. for Nat. Res. Manage., Bogor Agricultural Univ., Bogor, Indonesia, 2008.

[5] M. Hadi, "Pemodelan spasial kerawanan kebakaran di lahan gambut: Studi kasus kabupaten Ben gkalis, provinsi Riau," M.Sc. thesis,, Bogor Agricultural Univ. Bogor, Indonesia, 2006.

[6] [NASA]. (2014, Jun 23). FIRMS FAQ. [Online]. Available: https://earthdata.nasa.gov/data/near-real-time-data/faq/firms#firms7.

[7] S. Rinzivillo, F. Turini, "Classification in geographical information systems," in *Proc. The 8th European Conf.of Principles and Practice of Knowledge Discovery in Databases,* New York, 2004, pp. 374—385.

[8] I. S. Sitanggang, R. Yaakob, N. Mustapha, A. N. Ainudin, "Classification model for hotspot occurences using spatial decision tree algorithm," Journal of Computer Science, vol. 2, 2013, pp. 244—251.

[9] J. Han, M. Kamber, J. Pei, *Data Mining Concepts and Techniques.* San Fransisco: Morgan Kauffman Publishers, 2011, 3rd ed.

[10] A. Brimicombe, C. Li, *Location-Based Services and Geo-Information Engineering.* Sussex: Wiley-Blackwell, 2009.

[11] M. Ester, H. P. Kriegel, J. Sander, "Spatial data mining: a database approach", in *Proc 5th Int. Symposium on Large Spatial Databases,* Berlin, 1997, pp. 47—68.

[12] Riau goverment. (2013, Dec 12). Kabupaten Bengkalis. [Online]. Available: http://www.riau.go.id/index.php?/detail/6.

[13] I. S. Sitanggang, R. Yaakob, N. Mustapha, A. N. Ainudin, "Predictive models for hotspot ocurence using decision tree algorithms and loistic regression," Journal of Applied Science, vol. 3, 2013, pp. 252—261.

 978-1-4799-8075-8/14/$31.00