Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 13 (2): 252-261, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.252.261

Predictive Models for Hotspots Occurrence using Decision Tree Algorithms and Logistic Regression

[1,2]I.S. Sitanggang, [2]R. Yaakob, [2]N. Mustapha and [3]A.N. Ainuddin
[1]Department of Computer Science, Bogor Agricultural University, Darmaga Campus,
Jl. Meranti Wing 20 Level V, Bogor, 16680, Indonesia
[2]Faculty of Computer Science and Information Technology, Universiti Putra Malaysia,
43400 Serdang Selangor, Malaysia
[3]Institute of Tropical Forestry and Forest Products (INTROP),
Universiti Putra Malaysia, 43400 Serdang Selangor, Malaysia

Abstract: Predictive models for hotspots (active fires) occurrence are essential to develop as one of activities in forest fires prevention in order to minimize damages because of forest fires. This work applied the decision tree algorithms i.e., ID3 and C4.5, as well as logistic regression on spatial data of forest fires for Rokan Hilir District in Riau Province in Indonesia. The data consist of ten explanatory layers (physical, weather and socio-economic data) and a target layer. Target objects in the target layer are hotspots 2008 and non-hotspot points which were randomly generated near hotspots. As many 561 target objects were prepared through several data preprocessing tasks. The results show that the C4.5 algorithm has better performance than the ID3 algorithm in terms of accuracy and the number of generated rules. The C4.5 decision tree has the accuracy of 65.24% with number of generated rules is 35 and the first test attribute of the tree is peatland type. Furthermore, the logistic regression model outperforms the decision tree algorithms with the accuracy of 68.63%.

Key words: Decision tree, logistic regression, hotspots occurrence, forest fires

INTRODUCTION

Forest fires cause many negative effects in various aspects of life such as natural environment, economic and health (Herawati *et al*., 2006). In order to minimize damages due to forest fires in peatland and dry land, forest fires risk models are essential to develop. Several studies have been conducted in developing forest fire risk models by integrating Geographic Information Systems (GISs) and remote sensing. Boonyanuphap (2001) constructed a forest fire risk model for the area of Sasamba in East Kalimantan Indonesia by applying the GIS and Complete Mapping Analysis (CMA). Darmawan *et al.* (2001) integrated the remote sensing technique with the GIS to create a model of forest fire hazard in East Kalimantan, Indonesia. Moreover, a forest fires risk model for West Kutai District in East Kalimantan Province, Indonesia was developed by Danan (2008) using a GIS, remote sensing and Multi-criteria Analysis (MCA).

In addition to dryland, fire risk models have been also developed for peatlands in order to minimize the incidence of forest fires in peatlands. A-GIS-based peat swamp forest fire hazard model that integrated Analytical Hierarchy Process (AHP) and a GIS was built for the region Pekan in Pahang Malaysia (Iwan *et al*., 2004). Hadi (2006) 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). Furthermore, Razali *et al.* (2010) developed a fire hazard model in peat swamp forest of Penor/Kuantan District of Pahang, Malaysia using the GIS involving the fuel types, roads and canal.

Currently, data mining techniques have been also applied in modeling forest fire risks. Stojanova *et al.* (2006) built predictive models based on geographic data, meteorological ALADIN data and MODIS satellite data. This work utilized logistic regression and decision trees (J48), as well as random forests, bagging and boosting of decision trees, to obtain predictive models of fires occurrence. Furthermore, the association mining method was employed to find association between forest fire factors from the

Corresponding Author: Imas Sukaesih Sitanggang, Department of Computer Science, Bogor Agricultural University, Darmaga Campus, Jl. Meranti Wing 20 Level V, Bogor, 16680, Indonesia Tel: +62 251 8625584

historical data of every fire in order to estimate the fire grade (Yu and Bian, 2007). Prasad and Ramakrishna (2008) proposed a novel system for identifying forest fires from digital satellite images using K-means clustering algorithm and fuzzy logic. The association rule algorithm namely the Apriori algorithm was utilized by Hu *et al*. (2009) to study the probability and intensity of forest fires based on terrain data and weather data including time, temperature, moisture, wind speed and rainfall.

Hotspots (active fires) indicate spatial distribution of fires. Predicting hotspots occurrence may be considered as one of the activities in fires prevention. Therefore, developing predictive models for hotspots occurrence is important as an early warning system in order to minimize damages because of forest fires. In this study, we developed predictive models for hotspots occurrence for the study area Rokan Hilir in Riau Province in Indonesia using the logistic regression and the decision tree algorithms namely ID3 and C4.5. The data used in this work include spread and coordinates of hotspots in 2008, physical, socio-economic, as well as weather data.

MATERIALS AND METHODS

Study area: The study area is Rokan Hilir district in Riau Province in Indonesia (Fig. 1) that covers the area of 8,881.59 km^2 or about 10 percent of Riau's total land area (Pemerintah Kabupaten Rokan Hilir, 2010). It is positioned in the western part of the north Sumatera, the southern part of Bengkalis district and Rokan Hulu district, the eastern of Dumai as well as the northern part of the north Sumatera and Melaka strait.

Riau is one of provinces in Sumatra that has high deforestation because of forest fires especially in dry season. A study by Uryu (2008) shows that Riau has lost more than 65% of forest (about 4 million hectares) in the last 25 years and the majority of the deforestation has occurred on peat soil. A clear relation between fires and deforestation was found in Riau in which more than 72,000 active fires (hotspots) were recorded in this province by NOAA AVHRR and MODIS satellite sensors in the period 1997-2007 (Uryu, 2008). Rokan Hilir is one of districts in Riau that had 454,000 ha of peatlands in 2002 or about 11.2% of the whole peatlands in Riau (Wahyunto *et al*., 2005). As many 517 hotspots were found in Rokan Hilir in 2008.

Data and tools: This study utilized several factors influencing fire events which are suggested to identify hotspots occurrence. The factors include spread and coordinates of hotspots (Thoha, 2006; Danan, 2008); weather data of 2008 including maximum daily temperature (Boonyanuphap, 2001), daily rainfall (Boonyanuphap, 2001) and speed of wind; human activity factors i.e., roads (Darmawan *et al*., 2001; Boonyanuphap, 2001), rivers, city centers, land cover (Darmawan *et al*., 2001); and peatland (Iwan *et al*., 2004; Hadi, 2006; Razali *et al*., 2010). The data were collected from several institutions. The data used in this work and its source are summarized in Table 1.

In order to conduct experiments, this work utilized some software as follows:

- Quantum GIS 1.7.2 (http://www.qgis.org) for spatial data analysis and visualization
- PostgreSQL 9.1 (http://www.postgresql.org) as the database management system
- PostGIS 1.5 (http://www.postgis.org) for spatial data analysis

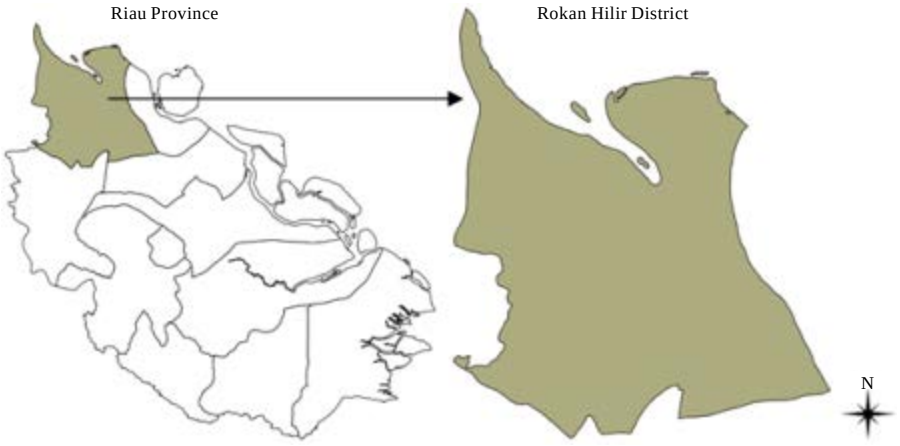

Fig. 1: Study area

Table 1: Data obtained from different sources

Data	Source
Spread and coordinates of hotspots 2006 (for satellite image processing)	FIRMS MODIS Fire/Hotspot, NASA/University of Maryland
Spread and coordinates of hotspots 2008 (for creating predictive models for hotspots occurrence)	FIRMS MODIS Fire/Hotspot, NASA/University of Maryland
Weather data 2008 (in the NetCDF format): maximum daily temperature, daily rainfall and speed of wind	Meteorological Climatological and Geophysical Agency (BMKG)
Digital maps for road, rivers, city centers, land cover and administrative border	National Coordinating Agency for Survey and Mapping (BAKOSURTANAL)
Digital maps for peatland depth and peatland types	Wetland International
Social and economic data: inhabitant's income source	BPS-Statistics Indonesia
Landsat TM, Resolution: 30×30 m^2	U.S. Geological Survey

- R for statistical computing (http://www.r-project.org/). This tool is used for creating logistic regression models
- Ilwis 3.7 (http://www.ilwis.org) for burn area processing
- ArcMap 9.3. This tool is utilized for spatial interpolation for weather data
- Weka 3.6.6 (http://www.cs.waikato.ac.nz/ml/weka/) for creating decision trees using non-spatial decision tree algorithms

Data preprocessing: This study conducted several preprocessing tasks on spatial data to prepare a dataset for the decision tree algorithms and logistic regression. The spatial data on forest fires are stored in a set of layers (the shp format) in the spatial database. There are two types of layers in the spatial database i.e., explanatory layers and a target layer. From the explanatory layers for physical, socio-economic and weather data, we determined the explanatory attributes for creating the models. Meanwhile, the target attribute containing classes of target objects was obtained from the target layer. Target objects in the dataset are hotspots of 2008 and non-hotspots that were randomly generated near hotspots. A buffer with the radius of 0.907374 km was created for each hotspot. Moreover, non-hotspot points were generated outside the buffers. Burn areas processing was performed in the study area to determine the radius of buffer for a hotspot. This task utilized a Landsat TM image as well as spread and coordinates of hotspots 2006. The Landsat TM image has the resolution of 30×30 m^2 with the acquisition date is 24 July 2006.

Preprocessing steps were also conducted on the objects in explanatory layers for weather data and income source of inhabitants who live in the study area. Weather data 2008 i.e., maximum daily temperature, daily rainfall and speed of wind, are represented in the NetCDF format. There are 62 points of weather data that spread on the Rokan Hilir area. Each point has values for weather data and location of the point (longitude and latitude). We applied the Ordinary Cokriging method (Goovaerts, 1998) to perform spatial interpolation for weather data. The purpose of this step is to calculate the values for weather data in the whole area of Rokan Hilir. The results of spatial interpolation for weather data were converted to the shp format such that the data can be integrated with other spatial data to prepare a task relevant dataset for modeling.

Additionally, several tasks was done to prepare the layer for income source including identifier matching, handling null value in polygon features and modifying categorical values for income source. These tasks are described as follows:

Identifier matching: Inconsistency between two layers occurred when a digital map for income source is created using the village border map. The digital map for village border is for 2007, whereas the income source data are selected from village potential data for 2008. There are 2 villages in the data that have different identifiers with those in the village border map. To overcome this inconsistency this work replaces identifiers in the map based on information from Statictics-Indonesia (BPS) such that spatial data from different years can be related one to each other.

Handling null values: There are two polygons in the village layer with no data for all socio-economic attributes. One polygon is located in forest and the other is a village (non-forest). This study assigns value 0 for attribute income source in forest area and non-zero new values for a village (non-forest) based on its neighbors. The topological operation ST_Touches in PostGIS was applied to find all neighbors that meet the village (non-forest). Moreover, income source of neighbor polygons was assigned to income source of the village.

Modifying categorical values for income source: Most of villages in Riau Provinces have income source Agriculture. The purpose of income source modification

is to detail the income source Agriculture. For villages with income source Agriculture, types of land cover are identified. A type of land cover which has the largest area is selected to modify the values for income source. This work combines a selected type and income source Agriculture to create a new value for income source. The topological operation ST_Intersection in PostGIS was applied to define all intersection areas between the land cover layer and the income source layer.

Creating non-spatial dataset: Decision tree algorithms require a relation as a task relevant dataset for the algorithm. The relation contains some explanatory attributes and one target attribute. The following steps were implemented to create the task relevant dataset from the spatial database using several spatial operations that are available in PostGIS:

Step 1: Calculate distance from target objects to nearest city center, river and road. The spatial operation employed is ST_Distance

Step 2: Apply the spatial operation ST_Within to relate explanatory layers to the target layer. This operation will produce several new layers. Each new layer is associated to an explanatory layer and the target layer

Step 3: Integrate all new layers in step 2 into a single layer by matching identifiers of objects in the target layer and those in an explanatory layer. This operation will produce another new layer

Step 4: The new layer in step 3 contains some non-spatial attributes and a spatial attribute. The spatial attribute stores the geometry type of spatial objects. In the database management system PostgreSQL, the_geom denotes this spatial attribute. The spatial attribute is removed because the dataset will be used as the input for the non-spatial methods.

Step 5: Remove duplicate objects in the non-spatial dataset

As mentioned in the step 1, this work computed distance from target objects (point features) to nearest road (line features), nearest river (line features) and nearest city centers (point features). To accomplish this task, the spatial operation ST_Distance in PostGIS 1.5 and the aggregate function min were applied to these layers. The results are three new layers i.e., dist_river, dist_road and dist_city. All objects in these three layers are points in which each point is associated to a target object and has minimum distance as the numerical value of its attribute. Because the ID3 algorithm works on categorical attributes, numerical values of distance from target objects to nearest rivers, roads and city centers were converted to categorical values based on the classes provided in Table 2.

Logistic regression model: A regression model formulates a relationship between x (called the explanatory variable, or predictor variable, or independent variable) and y (called the response variable, or dependent variable). Many applications include a dichotomous variable as a response variable. The variable, called a dummy variable, has only two possible discrete values (such as failure/success, male/female). These two possible values are denoted by 0 representing a "failure" and 1 representing a "success". A logistic regression model allows us to establish a relationship between a dummy variable as the dependent variable and a group of predictor variables (independent variables) (Triola, 2007). The logistic model has the form as follows (Everitt and Hothorn, 2006):

$$\text{Logit}(\pi) = \log\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q \quad (1)$$

where, π is the probability of the expected value taking the value one, β_0 is the intercept and β_i ($i = 1,2,\ldots,q$) are the slope parameters associated with independent variables x_i. β_i are coefficients of regression becoming the weights of each variable to produce a model for prediction. The logit of probability is the log of the odds of the response taking the value one (Everitt and Hothorn, 2006). The probability values can be quantitatively expressed in terms of independent variables by Everitt and Hothorn (2006):

$$\pi(x_1, x_2, \ldots, x_q) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q)} \quad (2)$$

The logit function can take any real value, but the related probability always falls in the interval (0,1). In a logistic regression model, $\exp(\beta_i)$ is the odds that the

Table 2: Classes for distance from target objects to city centers, rivers and roads

Distance (km)	Distance target object to nearest city center (x)	Distance target object to river (y)	Distance target object to road (z)
Low	$x \leq 7$	$y \leq 1.5$	$z \leq 2.5$
Medium	$7 < x \leq 14$	$1.5 < y \leq 3$	$2.5 < z \leq 5$
High	$x > 14$	$y > 3$	$z > 5$

dependent variable gets the value one when x_i increases by one, while the other explanatory variables remaining constant (Everitt and Hothorn, 2006).

The logistic regression has been applied to model forest fires occurrence. In the context of forest fire risk modeling, fires occurrence (hotspots) is the dependent variable. Meanwhile determinant factors (environmental and human factors) influencing fires occurrence are the independent ones. In predicting fires occurrence, π in Eq. 1 is the probability that a fire occurs. Thoha (2006) applied the logistic regression to develop a peat fire prediction model for Bengkalis district in Riau Province in Indonesia. The data used in the work of Thoha (2006) include land cover, vegetation, peatland depth, rivers, roads, villages, daily maximum temperature, daily rainfall and NOAA-AVHRR hotspots. The logistic regression was also used to construct a forest fire hazard model for West Kutai District in East Kalimantan Province in Indonesia (Danan, 2008). The data utilized in the work of Danan (2008) are land use, monthly precipitation, daily maximum temperature, daily rainfall and NOAA hotspots.

Decision tree algorithms: A decision tree is a model expressing classification rules. It is composed by three types of nodes: (1) a root node, (2) internal nodes and (3) leaf or terminal nodes. Each leaf node is assigned as a class label. Another node is either the root node or an internal node hold an attribute test condition to partition records that have different characteristics. Classification rules can be obtained by traversing the tree from the root node to the leaf nodes (terminals). Each rule consists of test attributes and their value.

There are several decision tree algorithms, such as ID3 developed by J. Ross Quinlan during the late 1970s and early 1980s, C4.5 as the successor of ID3 and CART (Classification and Regression Tree) proposed by Breiman *et al*. (1984). These algorithms have almost the same principle, where they build the tree in greedy manner starting from the root and selecting most informative features at each step (Marsland, 2009). Information gain is a measure to determine which the best feature will be selected for splitting the dataset. This measure is calculated based on the entropy. The idea in using entropy in the algorithm is to determine how much entropy of the whole training set will decrease if a particular feature is selected for the next classification step (Marsland, 2009).

Let S is the set of examples, F is a possible feature out of the set of all possible features and $|S_f|$ is number of member of S that have the value f for the feature F. Let a node N represents the tuples of S. The feature with the highest information gain is selected as the splitting feature for the node. Han and Kamber (2006) stated that "this attribute minimizes the information needed to classify the tuples in resulting partitions and reflects the least randomness or "impurity" in these partitions". Selecting a feature with the highest information gain minimizes the expected number of tests required to classify a given tuple and find a simple tree (Han and Kamber, 2006). The expected information needed to classify a tuple in S is given by:

$$\text{Info}(S) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad (3)$$

where, p_i is the probability that an arbitrary tuple in S belongs to the class C_i and it is estimated by $|C_{i,S}|/|S|$ (Marsland, 2009). A log function to the base 2 is used, because the information is encoded in bits. Info(S) is the average amount of information needed to identify the class label of a tuple in S. Info(S) is also known as the entropy of S.

Information gain is defined as the entropy of the whole dataset minus the entropy when a certain feature is selected (Marsland, 2009):

$$\text{Gain}(S,F) = \text{Entropy}(S) - \sum_{f \in values(F)} \frac{|S_f|}{|S|} \text{Entropy}(S_f) \quad (4)$$

Gain(S,F) tells us how much information would be gained by branching the tree on f. The highest Gain(S,F) represents that the amount of information still needed to complete classifying the tuples is minimal. Therefore splitting the dataset on the features f would give the best classification.

Figure 2 shows Quinlan's ID3 decision tree algorithm. The algorithm begins with the dataset test. If all objects in the dataset have the same class then the algorithm returns a leaf with the class as the label. The algorithm will determine the most common class and assign this class to a leaf if there are no features left to test. Otherwise the algorithm will calculate the information gain to select the best feature to split the dataset. The same procedure in the algorithm is then applied to smaller partitions to grow up the tree.

The C4.5 algorithm is the successor of ID3 that learns decision tree classifiers. The C4.5 algorithm uses information gain to select optimal splitting attributes and applies the post-pruning method to simplify the tree. There are three main tasks in the C4.5 algorithm: (1)

The ID3 algorithm
- if all examples have the same label:
  - return a leaf with that label
- else if there are no features left to test:
  - return a leaf with the most common label
- else:
  - choose the features $\hat{F}$ that maximizes the information gain of S to be the the next node using Eq. 4
  - add a branch from the node for each possible value f in $\hat{F}$
  - for each brach:
    - ⋆ calculate S_f by removing $\hat{F}$ from the set of features
    - ⋆ recursively call the algorithm with S_f, to compute the gain relative to the current set of examples

Fig. 2: The ID3 algorithm (Marsland, 2009)

generate the tree using the ID3 algorithm, (2) convert the tree to a set of if-then rules and (3) prune each rule by removing preconditions if the accuracy of the rule increases without it (Marsland, 2009).

RESULTS AND DISCUSSION

Several studies have been conducted in forest fire risk modeling using logistic regression and decision tree algorithms. A study by Vasconcelos *et al.* (2001) developed models to predict spatially distributed probabilities of ignition of wildland fires in central Portugal using logistic regression and neural networks. The logistic regression models have the producer's accuracy for ignition i.e., 77.8, 78.8 and 78.6% for three different datasets. Whereas, the user's accuracy for ignition are 51.3, 42.5 and 25.6% for three different datasets (Vasconcelos *et al.*, 2001). Moreover, a stepwise Ordinary Least Squares (OLS) regression was performed to construct a wildland fire occurrence model in Southern Europe (Koutsias *et al.*, 2005). The dataset used in the study by Koutsias *et al.* (2005) consists of socio-economic, demographic indicators and land cover (land use) statistics. The accuracy of the ordinary logistic regression model is 66% (Koutsias *et al.*, 2005). Stojanova *et al.* (2006) applied logistic regression and decision tree algorithms on forest fires datasets for different regions of Slovenia: Kras region in western Slovenia, Primorska region and the continental part of Slovenia. The average accuracy of logistic regression model is 81.53% while the average accuracy of the C4.5 decision tree is 79.4%. Additionally, a hotspots occurrence model for Rokan Hilir District in Riau Province Indonesia has been created using the C4.5 decision tree algorithm on the forest fires dataset in which the accuracy of the model is 63.17% (Sitanggang and Ismail, 2011). The dataset is composed by locations of hotspot occurrences and human activity factors including locations of city centers, road network, river network as well as land covers types. In another work by Sitanggang *et al.* (2012), the C4.5 decision tree model for hotspots occurrence in Riau Province has been developed with the accuracy 69.59%.

In this study, the decision tree algorithms and the logistic regression have been applied on the task relevant dataset that was resulted from the preprocessing steps. The dataset has one target attribute (target) and ten explanatory attributes. Explanatory attributes and its distinct values are provided in Table 3. Number of target objects in the dataset is 561 that consist of 235 hotspots as true alarm data and 326 non-hotspots as false alarm data. After relating all explanatory layers and the target layer, the number of hotspots in the dataset decreases from 517 to 235.

Decision tree algorithms: Experiments were conducted using the ID3 module and the J48 module as Java implementation of C4.5 that are available in the data mining toolkit Weka 3.6.6. The J48 package is a Weka's implementation of the decision tree learner. The package is a directory containing a collection of related classes that builds a C4.5 decision tree. The dataset is divided into two groups: a training set to develop a classification model and a testing set to calculate accuracy of the model. This work applied the 10-folds cross validation to determine accuracy of the classifier. The accuracy of ID3 decision tree is 49.02%, whereas the accuracy of C4.5 decision tree is 65.24%.

Moreover, the ID3 algorithm has 270 leaves with peatland type as the first test attribute. Meanwhile, the C4.5 algorithm produces a simpler decision tree with 35 leaves and the first test attribute of the tree is peatland type. The test attributes in the C45 decision tree are peatland type, distance to nearest road, distance to

Table 3: Number of features and distinct values in the dataset

Explanatory attributes	No. of distinct values
Physical	
Distance to nearest river (dist_river)	3 (low, medium, high)
Distance to nearest road (dist_road)	3 (low, medium, high)
Distance to nearest city center (dist_city)	3 (low, medium, high)
Land cover (land_cover)	12 (Dryland_forest, plantation, Water_body and so on)
income source (income_source)	7 (Forestry, Agriculture, Trading_restaurant and so on)
Weather	
Precipitation in mm/day (precipitation)	2, 3
Screen temperature in K (screen_temp)	297, 298, 299
10 m wind speed in m sec^{-1} (wind_speed)	0, 1, 2
Peatland	
Peatland type (peatland_type)	Saprists/min (50/50), Shallow and so on
Peatland depth (peatland_depth)	D1 (Shallow/Thin 50-100 cm), D2 (Moderate 100-200 cm), D3 (Deep/Thick 200-400 cm), D4 (very deep/very thick > 400 cm)

nearest city center, screen temperature, distance to nearest river and income source. The following are sample of rules generated from the C4.5 decision tree:

- IF peatland_type = non_peatland THEN Hotspot Occurrence = False
- IF peatland_type = Saprists/min(90/10), Moderate AND 2.5 km < dist_road ≤ 5 km THEN Hotspot Occurrence = True
- IF peatland_type = Hemists/Saprists(60/40), Moderate AND income_source = Plantation THEN Hotspot Occurrence = False
- IF peatland_type = Hemists/Saprists(60/40), Moderate AND income_source = Agriculture AND 2.5 km < dist_road ≤ 5 km THEN Hotspot Occurrence = True
- IF peatland_type = Saprists/min(50/50), Shallow THEN Hotspot Occurrence = False

The label in peatland type for example "Hemists/Saprists (60/40), Moderate" is described as follows: Hemists and Saprists are peatland types, the value 60 and 40, respectively represent 60% Hemists and 40% Saprists covering the area and Moderate (100-200 cm) is a category for peatland depth.

Logistic regression model: A logistic regression model for the forest fire dataset was created using the GLM (Generalized Linear Model) function provided in the open source tool for statistical computing R. The dataset contains 561 tuples, a dependent variable and independent variables. The dependent variable is the target with values: True (true alarm data i.e., hotspots occurrence) and False (false alarm data i.e., non-hotspots occurrence). Independent variables are nominal variables (unordered factors) i.e., dist_city, dist_river, dist_road, income_source, land_cover, peatland_type, peatland_depth, precipitation, screen_temp and wind_speed (refer to Table 3 for the description of variables).

Stepwise regression includes regression models in which the choice of independent variables is carried out by an automatic procedure. In R, the stepwise regression method uses the Akaike Information Criterion (AIC) for model selection. The Akaike information criterion is a measure of the relative goodness of fit of a statistical model. This work executed the stepwise regression function step in R on the forest fires dataset. The results are six logistic regression models.

Using the training dataset (the dataset for creating the regression models), accuracy of each model was calculated with the cut-off value for classification of 0.5. The cut-off value of 0.5 is recommended if there are two categories (classes) in the population in which an observation is assigned to the class if its probability of membership is the highest. In this case, a tuple is assigned to 'True' class (true alarm data) if the model predicts an outcome probability of greater than or equal to 0.5. Otherwise, a record is assigned to 'False' class (false alarm data). The best model gives the highest accuracy of 68.63% in which 385 records of 561 records are correctly predicted to True or False classes. This regression model excludes the variables income_source, land_cover and peatland_depth. The beta coefficients of the best model are given in Table 4.

According to Table 4, there is only one factor that is statistically significant affecting the hotspots occurrence with the significance level of 0.05. This factor is peatland type of Saprists/min(90/10),Moderate that has one star for its p-values (Pr(>|z|)) as the flag for the significant coefficient (Table 4).

The beta coefficients are represented in log-odds units. These coefficients indicate the values or the coefficients for the logistic regression model for predicting hotspots occurrence as the dependent variable from the independent (predictive) variables. The logistic regression coefficients provide the change in the log-odds of the outcome for a one unit increase in the predictor variable, holding all other predictors constant. For examples, having distance to nearest river with the

Table 4: Beta coefficients of the best logistic regression model

	Estimate	SE	z-value	pr (>|z|)
Intercept	-0.90016	0.42732	-2.107	0.0352
Distance to nearest city center (dist_city)				
Low (dist_city ≤ 7 km) - dummy variable of dist_city	-0.06829	0.24805	-0.275	0.7831
High (dist_city > 14 km) - dummy variable of dist_city	0.31603	0.22795	1.386	0.1656
Distance to nearest river (dist_river)				
Medium (1.5 km < dist_river ≤ 3 km) - dummy variable of dist_river	0.47953	0.25004	1.918	0.0551
High (dist_river > 3 km) - dummy variable of dist_river	0.40041	0.25098	1.595	0.1106
Distance to nearest road (dist_road)				
Medium (2.5 km < dist_road ≤ 5 km) - dummy variable of dist_road	0.36547	0.23104	1.582	0.1137
High (dist_road > 5 km) - dummy variable of dist_road	-0.49129	0.26573	-1.849	0.0645
Peatland type				
Saprists/ Hemists(60/40),Moderate - dummy variable of peatland type	-15.76648	764.65815	-0.021	0.9835
Saprists/ min(50/50),Moderate - dummy variable of peatland type	-1.55727	1.12605	-1.383	0.1667
Hemists/min(30/70),Moderate - dummy variable of peatland type	-16.39354	1681.81929	-0.010	0.9922
Hemists/Saprists(60/40),Very_deep - dummy variable of peatland type	0.10725	0.34567	0.310	0.7564
non_peatland - dummy variable of peatland type	-0.56774	0.32362	-1.754	0.0794
Saprists/min(90/10),Moderate - dummy variable of peatland type	1.02707	0.50323	2.041	0.0413
Hemists/Saprists(60/40),Deep - dummy variable of peatland type	0.88688	0.50158	1.768	0.0770
Saprists/min(50/50),Shallow - dummy variable of peatland type	-1.34998	0.82865	-1.629	0.1033
Saprists(100),Moderate - dummy variable of peatland type	0.91652	0.61753	1.484	0.1378
Precipitation				
2 (2 mm/day ≤ precipitation < 3 mm/day) - dummy variable of precipitation	-1.52038	1.08165	-1.406	0.1598
Screen temperature				
297 (297 K ≤ Screen temperature < 298 K) - dummy variable of screen temperature	0.54533	0.27838	1.959	0.0501
Wind speed				
0 (0 m sec^{-1} ≤ Wind speed <1 m sec^{-1}) - dummy variable of wind speed	-0.03025	0.24737	-0.122	0.9027

value medium (1.5 km < dist_river ≤ 3 km), versus distance to nearest river with the value low (dist_river ≤ 1.5), increases the log odds of hotspots occurrence by 0.47953 (Table 4).

Summary: Predictive models for hotspots occurrence are essential to develop so that the damage caused by forest fires can be minimized. Several spatial data supporting hotspots occurrence are geographical environment (land cover, roads, rivers, city centers and peatland), socio-economic data (income source) and weather data (precipitation, screen temperature and 10 m wind speed). This work was applied the decision tree algorithms i.e., ID3 and C4.5, as well as the logistic regression on the forest fires dataset. The experimental results show that the logistic regression has better performance in terms of accuracy than the two decision tree algorithms. The stepwise logistic regression that is available in R produces the best model with the accuracy on the training set is 68.63%. Meanwhile, the accuracy of ID3 decision tree is 49.02% and the accuracy of C4.5 decision tree is 65.24%. Furthermore, in terms of the number of rules generated from the trees, the C4.5 algorithm outperforms the ID3 algorithm with the number rules is 35 and the first test attribute of the tree is peatland type. The attributes in the C4.5 decision tree which are used to classified the objects include peatland type, distance to nearest road, distance to nearest city center, screen temperature, distance to nearest river and income source. Additionally, the logistic regression model consists of some variables to classify objects to True or False classes. The variables are distance to nearest city center, distance to nearest river, distance to nearest road, peatland type, precipitation, screen temperature and wind speed. Moreover, the logistic regression model shows that income source, land cover and peatland depth do not affect hotspots occurrence because these variables are excluded in the model. According to the p-values, the factor peatland type of Saprists/min(90/10), Moderate is statistically significant influencing hotspots occurrence with the significance level of 0.05. The peatland type is also significant in the C4.5 decision tree because it becomes the first test attribute in classifying objects to hotspots and non-hotspot points. The C4.5 decision tree and the logistic regression model for predicting hotspots occurrence may benefit the making of fire prevention plans. Therefore damages and losses because of forest fires can be minimized.

ACKNOWLEDGMENTS

The authors would like to thank Indonesia Directorate General of Higher Education (IDGHE), Ministry of National Education, Indonesia for supporting PhD Scholarship (Contract No. 1724.2/D4.4/2008) and Southeast Asian Regional Center for Graduate Study and Research in Agriculture (SEARCA) for partially supporting the research.

REFERENCES

Boonyanuphap, J., 2001. GIS-based method in developing wildfire risk model: A case study in Sasamba, East Kalimantan, Indonesia. Master's Thesis, Bogor Agricultural University, Indonesia.

Breiman, L., J. Friedman, R.A. Olshen and C.J. Stone, 1984. Classification and Regression Trees. 1st Edn., Chapman and Hall/CRC Press, California, USA., ISBN-13: 978-0412048418, pp: 368.

Danan, P.H., 2008. A RS/GIS based multi-criteria approaches to assess forest fire hazard in Indonesia (Case study: West Kutai district, East Kalimantan province). Master's Thesis, Bogor Agricultural University, Indonesia.

Darmawan, M., A. Masamu and T. Satoshi, 2001. 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. Proceedings of the 22nd Asian Conference on Remote Sensing, November 5-9, 2001, Singapore, pp: 1-6.

Everitt, B. and T. Hothorn, 2006. A Handbook of Statistical Analyses Using R. Taylor and Francis/CRC Press, Boca Raton, FL., USA., ISBN-13: 9781584885399, Pages: 304.

Goovaerts, P., 1998. Ordinary cokriging revisited. Math. Geol., 30: 21-42.

Hadi, M., 2006. Spatial modeling on susceptibility of fires in peatland, a case study in District of Bengkalis, Riau province. Master's Thesis, Graduate School, Bogor Agricultural University, Indonesia.

Han, J. and M. Kamber, 2006. Data Mining: Concepts and Techniques. 2nd Edn., Morgan Kaufmann Publisher, San Fransisco, USA., ISBN: 1-55860-901-6.

Herawati, H., S. Heru and F. Claudio, 2006. Forest fires and climate change in Indonesia. Background Document for the Southeast Asia Kick-Off Meeting of the Project Tropical Forests and Climate Change Adaptation (TroFCCA).

Hu, L., G. Zhou and Y. Qiu, 2009. Application of Apriori algorithm to the data mining of the wildfire. Proceedings of the 6th International Conference on Fuzzy Systems and Knowledge Discovery, Volume 2, August 14-16, 2009, Tianjin, China, pp: 426-429.

Iwan, S., A.R. Mahmud, S. Mansor, A.R.M. Sharriff and A.A. nuruddin, 2004. GIS-grid-based and multi-criteria analysis for identifying and mapping peat swamp forest fire hazard in Pahang, Malaysia. Disaster Prev. Manage. J., 13: 379-386.

Koutsias, N., J. Martinez, E. Chuvieco amd B. Allgower, 2005. Modeling wildland fire occurrence in southern europe by a geographically weighted regression approach. Proceedings of the 5th International Workshop on Remote Sensing and GIS Applications to Forest Fire Management: Fire Effects Assessment, June 16-18, 2005, Zaragosa, Spain, pp: 57-60.

Marsland, S., 2009. Machine Learning: An Algorithmic Perspective. Chapman and Hall/CRC Press, Boca Raton, FL., USA., ISBN-13: 9781420067187, Pages: 406.

Prasad, K.S.N. and S. Ramakrishna, 2008. An autonomous forest fire detection system based on spatial data mining and fuzzy logic. Int. J. Comput. Sci. Network Secur., 8: 49-55.

Razali , S.M., A.A. Nuruddin, I.A. Malek and N.A. Patah, 2010. Forest fire hazard rating assessment in peat swamp forest using landsat thematic mapper image. J. Applied Remote Sens., Vol. 4.

Sitanggang, I.S. and M.H. Ismail, 2011. Classification model for hotspot occurrences using a decision tree method. Geomat. Natl. Hazard. Risk, 2: 111-121.

Sitanggang, I.S., R . Yaakob, N. Mustapha and A.N. Ainuddin, 2012. Application of classification algorithms in data mining for hotspots occurrence prediction in Riau province Indonesia. J. Theor. Applied Inform. Technol., 43: 214-221.

Stojanova, D., P. Panov, A. Kobler, S. Dzeroski and K. Taskova, 2006. Learning to predict forest fires with different data mining techniques. Proceedings of the Conference on Data Mining and Data Warehouses, October 9, 2006, Ljubljana, Slovenia, pp: 255-258.

Thoha, A.S., 2006. Using remote sensing and geographical information system for detecting and predicting peatland fires in district of Bengkalis, Riau province. Master's Thesis, Bogor Agricultural University, Indonesia.

Triola, M., 2007. Elementary Statistics: With Multimedia Study Guide. 10th Edn., Pearson Education Limited, New York, USA., ISBN-13: 9780321460929, Pages: 868.

Uryu, Y., 2008. Deforestation, forest degradation, biodiversity loss and CO_2 emissions in Riau, Sumatra, Indonesia. WWF Indonesia Technical Rep., 27: 1-80.

Vasconcelos, M.J.P., S. Silva, M. Tome, M. Alvim and J.M.C. Pereira, 2001. Spatial prediction of fire ignition probabilities: Comparing logistic regression and neural networks. Photogram. Eng. Remote Sens., 67: 73-81.

Wahyunto, S. Ritung, Suparto and H. Subagjo, 2005. Peatland distribution and its C content in Sumatra and Kalimantan. Wetland International-Indonesia Programme and Wildlife Habitat Canada. Bogor, Indonesia.

Yu, L. and F. Bian, 2007. An incremental data mining method for spatial association rule in gis based fireproof system. Proceedings of the International Conference on Wireless Communications, Networking and Mobile Computing, September 21-25, 2007, Shanghai, China, pp: 5983-5986.