2019 International Conference on Computer, Control, Informatics and its Applications

Evaluation of Random Forest model for forest fire prediction based on climatology over Borneo

Arnida L. Latifah
Research Center for Informatrics
Indonesian Institute of Sciences
Bandung, Indonesia
arni001@lipi.go.id

Ayu Shabrina
Research Center for Informatrics
Indonesian Institute of Sciences
Bandung, Indonesia
ayus003@lipi.go.id

Intan N. Wahyuni
Research Center for Informatrics
Indonesian Institute of Sciences
Bandung, Indonesia
inta009@lipi.go.id

Rifki Sadikin
Research Center for Informatrics
Indonesian Institute of Sciences
Bandung, Indonesia
rifk001@lipi.go.id

Abstract—**Indonesia has entered an alarming condition related to forest fires. It is becoming a seasonal hazardous phenomenon in tropics. As the largest tropical forest in Indonesia, Borneo is the most susceptible area to fire especially in dry condition. Forest fires are threatened by climate, human activities, and ecosystem processes, but only climate variable can be quantified well in Borneo. This research aims to evaluate random forest model in predicting forest fires based on the climate variables and satellite data of burned area. Prediction of forest fires is expected to reduce the impact of forest fires in the future. Based on analysis of spatial and annual variability, the random forest model with all selected climate variables can represent the forest fires event over Borneo.**

Keywords—climate, forest fire, random forest, Borneo

I. INTRODUCTION

In the middle of 2019, Indonesia had entered an alarming condition related to forest and land fires. Some regions in Borneo determine the alert status of forest and land fire emergency. The alert status set due to the threat of drought with moderate to high risk. Based on the WHO and the Ministry of Health book, the direct impact of forest fires are loss of agricultural, plantation and forestry assets. Another worrying impact is fire fumes that trigger respiratory disease for residents.

The Indonesian Forum for Environment (WALHI) revealed that there were three causes of forest fires. First, slow recovery and protection of land. Then, land management is not strict in taking action of deforestation activities. Third, prolonged dry weather factors.

Exploring at the history of forest fires in Indonesia, in 2015, the fire situation in this region is exceptionally severe [2]. The 2015 fire episode in Indonesian Sumatra and southern Borneo is the largest fire event since 1997 surpasses the second largest year in 2006 [5]. The event is caused by the lack of control in fire agriculture, legally or illegally, during the dry season. The activity means the growing of crops by burning a forest and planting among the charred stumps. During drought years, fire can evade this control and burn a substantial acreage beyond what was intended of [3]. Earth Observatory, NASA, give a statement that as El Niño intensifies, peat deposits are making seasonal fires unusually difficult to control.

Prolonged dry weather and the El Nino events are one of the causes of forest fires in Indonesia. Dry weather can be measured by climate variables such as temperature and humidity, so the forest fires event and climate variables have a strong correlation. [6] explain that weather conditions, such as temperature and air humidity, are known to affect fire occurrence. In this research, forest fire event is measured by the percentage of burned area fraction and the climate variable is represented by a temperature, relative humidity, precipitation and wind speed.

Prediction of forest fires is a precaution or minimization of risk from the impact of forest fires in the future. The main purpose of this research is to evaluate the random forest (RF) model based on the climate variable and satellite data of burned area for predicting forest fire over Borneo. There are several studies that predict forest fires based on random forest model, but there is still a lack of forest fire study over Borneo using random forest model .

The study [9] used anthropogenic and geographical feature data with the random forest algorithm to highlight factors that most influence the fire-ignition and to identify areas under risk. Their study was implemented for forest fire over Canton Ticino (Switzerland).

In [13], Ripley's K(d) function and RF were applied to analyze the drivers, spatial distribution and risk patterns of fires in Yichun, a typical FC in China. The results revealed a clustered distribution of forest fire ignitions in Yichun, as well as identified the driving factors and their dynamic influence on fire occurrence. Fire risk zones were identified based on RF modelling. RF performed well, forecasting fire occurrence in Yichun with a high prediction accuracy (82.9%) using all factors combined (topography, vegetation types, infrastructure, meteorology, social-economic factors).

[11] using a RF classifier to develop model applied to the test site for classification of the burned area. An overall accuracy was obtained as 0.99. The results show that this approach is very useful to be used to determine burned forest areas in Adrasan and Kumluca regions in Antalya province.

II. METHODS

A. Study Area and Study Periods

The study area covers Borneo island, which is the third largest island in the world and crossed by the equator. The study domain is between 3.875S to 6.875N and 109.125 to 113.875 E. In the calculation process, this area is gridded

into spatial resolution of 0.25 latitude by 0.25 longitude. The data used in the training phase is collected from January 1998 to December 2013 (16 years). The testing phase uses data for two years from January 2014 – December 2015.

B. Forest Fire Data

Forest fire data is obtained from Global Fire Emission Database. They have combined satellite information of fire activity and vegetation productivity to estimate gridded monthly burned area and fire emission, as well as scalars that can be used to calculate higher temporal resolution emissions. Data files contains columns and rows that corresponded to the selected domain and has a 0.25 degree latitude by a 0.25 degree longitude spatial resolution. Data is available from 1997 through present. The data used in this experiment is the burned area without small fires.

Burned area is obtained from Global Fire Emission Database version 4 (GFED4) [8]. The GFED4 burned area is based on active fire detection from European Remote Sensing Satellite Along-Track Scanning Radiometer (ATSR) World Fire Atlas, Tropical Rainfall Measuring Mission Visible and Infrared Scanner (VIRS) and the Moderate Resolution Imaging Spectroradiometer (MODIS) burnt area product (MCD64A1) [8]. The GFED4 burned area set provides global, monthly burned area from 1997 through the present and higher temporal resolution daily burned area. Burned area is represented by burned fraction with fraction of grid cell as the unit of calculation. The burned fraction in August, September and October had higher proportion than other months in every year. Based on historical data, the burned fraction had maximum proportion at 33% of grid cell.

C. Climate Data

In this paper, we used several climate factors that have correlation with weather condition. The elements that important in spreading property of forest fire are temperature, relative humidity, wind speed and precipitation.

The climate data is provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). ECMWF uses its forecast models and data assimilation system to 'reanalyse' archived observations, creating global data sets describing the recent history of the atmosphere, land surface, and oceans. ERA-Interim is a frequently used global atmospheric reanalysis produced by ECMWF [12]. The ERA Interim version 2.0 climate data set hosted by the ECMWF is used in this research. The ERA Interim is delivered in a netCDF format in the geographical coordinate system with a 0.25 degree resolution. The data is available for the years January 1979-August 2019. The air temperature and wind speed from the years 1998 -2015 are taken from this data set. The climate data from ERA Interim that is used in this study are in the following:

- 2 meter dew point temperature (d2m) in Kelvin.
- 2 meter temperature (t2m) in Kelvin.
- 10 meter wind speed (si10) in ms^{-1}.

In further calculations, the dewpoint temperature and temperature variables are transformed in Celsius unit using subtraction by 273.15. Using Magnus formula, these variable is used to calculate the relative humidity variable [10].

The precipitation variable is obtained from the Tropical Rainfall Measuring Mission (TRMM), a joint mission of NASA and the Japan Aerospace Exploration Agency to study rainfall for weather and climate research. TRMM was a research satellite designed to improve our understanding of the distribution and variability of precipitation within the tropics as part of the water cycle in the current climate system. The Tropical Rainfall Measuring Mission (TRMM) Multi-satellite Precipitation Analysis (TMPA) is intended to provide a "best" estimate of quasi-global precipitation from the wide variety of modern satellite-borne precipitation-related sensors. Estimates are provided at relatively fine scales (0.25° × 0.25°, 3-h) in both real and post-real time to accommodate a wide range of researchers. The 3B43 dataset is the monthly of merged microwave-infrared precipitation rate (in mm / hour) [4]. This variable is transformed from per hour unit to per day unit. The preprocessing was performed for all variable due to inconsistency of range value, metric and unit. The climate and forest fire variable is normalized using the mean and standard deviation of each data.

D. Random Forest

Based on [7], random forests algorithm are an ensemble learning method for classification and regression. This algorithm creates the forest with the number of trees. The random forest classification is determined based on the results of voting from the tree formed. The winner of the tree formed is determined by the most votes.

Random forest uses a decision trees for the selection process. The tree is divided recursively from data in the same class. Split is divide data based on the type of attribute. When determining classification, a bad tree will make conflicting random predictions. Thus, some decision trees will produce good answers.

According to [1] the random forest algorithm both for classification and regression is as follows:

- Draw n-tree bootstrap samples from the original data.
- For each of the bootstrap samples, grow an unpruned classification or regression tree, with the following modification: at each node, rather than choosing the best split among all predictors, randomly sample m-try of the predictors and choose the best split from among those variables.
- Predict new data by aggregating the predictions of the n-tree trees.

E. Simulation

The random forest model was built by regressing forest fires data with 4 climate data variables in the training period. Forest fire data, burned area, as independent variable and climate data as the dependent variable. Sklearn Ensemble Random forest library is used to develop random forest model. The random forest regressor estimates the number of decision trees on the training data on several parameters. Estimation results from decision trees are averaged to improve accuracy and control over fitting. A number of parameter values is adjusted in the regressor to build a model that fits the data. Bootstrap samples are used when building trees. The number of trees is set to 50 with the maximum depth of the tree is 10 and the minimum number

of samples required to split an internal node is 3. The mean squared error is chosen to measure the quality of a split. There is no maximum number of leaf nodes. At a leaf node, the minimum number of samples required is 4 and there is no minimum weighted fraction of the sum total of weights (of all the input samples) required to be. A node will be split if this split induces a decrease of the impurity greater than or equal to zero, as in default. The model which is developed by the random forest regressor with the parameters described is used for the prediction process in testing period. Based on the Sklearn library, the process of predicting burned areas in 2014-2015 use a random forest model with input in the form of climate data in the same year. The results of these predictions used to calculate the accuracy of the model from 16 years training period.

III. RESULTS AND DISCUSSION

Forest fire prediction based on climate data in Borneo is displayed in spatial variability and annual variability. Spatial variability shows the burned area at spatial location with different colors based on the chance of forest fire. Annual variability shows the average of burned area every month in one year. Forest fire prediction using the Random Forest model will be compared with GFED data in 2014 and 2015.

Figure 1. Features' Significance Importance.

Figure 1 shows the monthly average burned area of each climate variable on the prediction of forest fires over Borneo in 2014-2015. Climate data consists of 1) monthly precipitation, 2) temperature, 3) relative humidity, and 4) wind speed. An analysis of the effect of each climate variable on the predicted results will focus on suitability with the GFED pattern. The achievement of prediction results with GFED value will be more focused in September, where there is an extreme increase in the percentage value of the burned area.

For one variable model, the predictive value pattern does not correspond to GFED pattern in 2014 and 2015. The peak potential for forest fires is in August-November. But the predicted value of the model with one variable did not increase in September. It means the model with one climate variable cannot represent the value of fires.

For two variables model, predictive value patterns have been following the GFED pattern, except for variable 14 (monthly precipitation, wind speed), variable 13 (monthly precipitation and relative humidity), and variable 34 (relative humidity, wind speed). In September 2014, the predicted value of the variable models 23 (temperature, relative humidity) and 24 (temperature, wind speed) almost reached the highest value of the average GFED. Same as the predicted value of one variable model, the average value of two variable models cannot reach the highest value in September 2015.

For three and four variables model, the predictive value pattern has followed the GFED pattern, except for variable 134 (monthly precipitation, relative humidity, wind speed). Model with variable 234 can reach the highest average value in September 2014 but the value is still below the value of GFED in 2015. The model with the variable 1234 shows an average value is close to the highest value in GFED 2015 in September, but the predicted average value in September 2014 exceeds the GFED value.

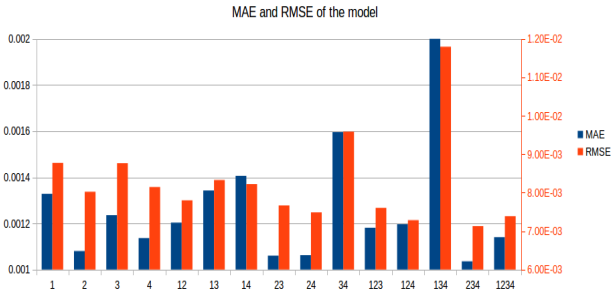

Figure 2. Spatial Distribution of Multi-Model.

Figure 2 shows the calculation of the error from model RF with all combinations of climate variables. The smallest mean absolute error (MAE) is obtained from the model with variable 234 then model with variables 23, 24, 2, and 1234. Variables 234, 124, 1234 are three variable combinations that produce the smallest root mean square error (RMSE) value. From these results, candidates who have small MAE and RMSE values are RF models with variables 234 and 1234.

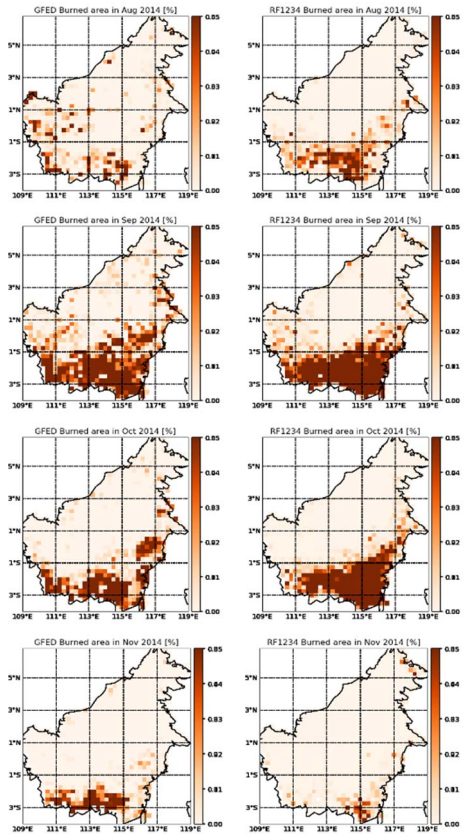

Figure 3. Spatial Variability over Borneo in August – November 2014

In Figure 3, we analyse the fire spots map from the model RF with the four climate variables were compared to the GFED map. Analysis of spatial variability for domain Borneo in the most significant testing period, in August - November 2014:

- In August 2014, there is a difference in distribution location of fire spots from model RF with GFED map. The GFED map shows the fire spots spread to northern and eastern Borneo but the RF model map shows the fire spots gather in southern Borneo. Therefore, over fitting prediction of fires occurred in South Borneo.
- September and October 2014 were the months with the most fires. This happens a lot in southern Borneo. RF model maps have been able to accurately estimate the location of fires according to GFED data.
- In November 2014, the RF model map could not capture all the fires in South Borneo. Fire spots prediction can only describe a small part of the GFED fire location, so that the under fitting prediction occurs on the predicted results of the RF model.

In August - November 2015, the RF model map was able to predict the similar position of fire spots according to the GFED map. Although the predicted results of the RF model have not been able to capture several fire spots with a small percentage such as in August and November 2015.

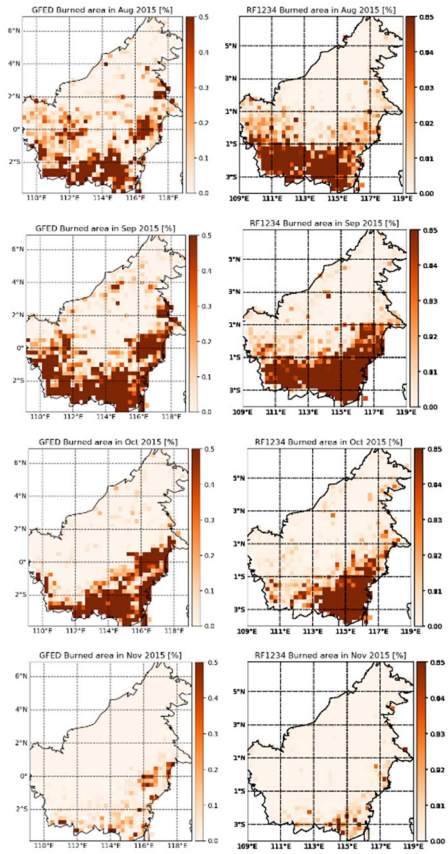

Figure 4. Spatial Variability over Borneo in August – November 2015

Figure 5 shows the comparison of percentage the average of burned areas over Borneo in 2014 and 2015 from model RF with all selected variables. It shows that there is a matching trend between GFED and RF in the testing period. The highest value of the burned area occurred in September due to a rapid increase compared to other months in every testing year. Comparing the predicted results for the month, in September 2014 the prediction value was below the GFED value, meanwhile the prediction value was much higher than the GFED in 2015.

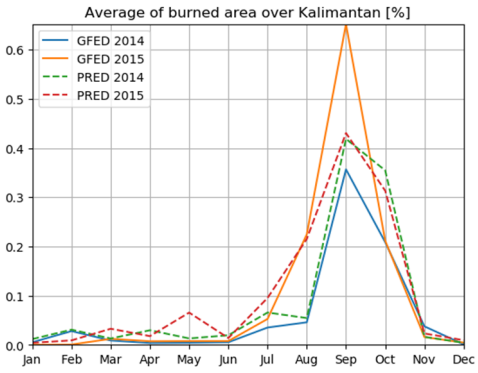

Figure 5. Comparison average of Burned Area over Borneo between RF model and GDEF

IV. CONCLUSIONS

Based on the comparison of annual variability analysis, RF model using four climate variables produce the predictive results that compatible with the GFED trend. In September, it has the best predictive value of all combination climate variables. The results of MAE and RMSE calculations also show the model with variable 1234 is one of the models with the smallest error. The spatial analysis of the selected model shows that the RF model can predict the similar position of fire spots according to the GFED map in the testing period. It can be concluded that the RF model with all selected climate variables is the chosen model to represent the forest fires event in the Borneo region.

ACKNOWLEDGMENT

We thank to Dr. Ardhasena Sopaheluawakan, Dr. R. Kartika Lestari, and Zaenal Akbar, Ph.D. for the initiation and discussion related to the forest fire research. The computation in this work has been done using the facilities of HPC LIPI, Indonesian Institute of Sciences (LIPI). The provider of TRMM data, GFED and ECMWF are highly appreciated.

REFERENCES

[1] A. Liaw, M. Wiener, "Classification and Regression by Random Forest", ISSN 1609-3631, December 2002.

[2] A. Voiland, “Heavy Smoke Blankets Borneo”, earthobservatory. nasa.gov, 2015. (http://earthobservatory.nasa.gov/IOTD/view.php?id=86847&eocn=image&eoci=related_image)

[3] C. Chen, H. Lin, J. Yu, M. Lo, “The 2015 Borneo fires: what have we learned from the 1997 and 2016 El Ninos?”, Environmental Research Letters, vol. 11,no. 10, 2016.

[4] G. J. Huffman, R. R. Adler, D. T. Bolvin, G. Gu, E. J. Nelkin, K. P. Bowman, Y. Hong, E. F. Stocker, D. B. Wolff, “The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-Global, Multiyear, Combined-Sensor Precipitation Estimates as Fine Scales”, Journal of Hydrometeorology, vol. 8, February 2007.

[5] G. R. Van der Werf,,” Indonesian fire season progression”, globalfiredata.org, 2015. (http://globalfiredata.org/updates.html)

[6] J. Terradas J. Pinol and F. Lloret. “Climate warming, wildfire hazard, and wildfire occurrence in coastal eastern Spain”. Climatic Change, vol. 38, pp. 345–357, 1998.

[7] L. Breiman,”Random forests”. Mach. Learn., vol. 45, pp 5 – 32, October 2001.

[8] L. Giglio, J. T. Randerson, G. R. van der Werf, “Analysis of daily, monthly, and annual burned area using the fourth-generation global fire emissions database (GFED4)”, Journal of Geophysical Research: Biogeosciences, vol. 118, issue. 1, March 2013.

[9] M. Leuenberger, M. Kanevski, C. D. V. Orozco, “Forest Fires in a Random Forest”, Geophysical Research Abstract, vol. 15, 2013.

[10] NPL, “A Guide to the Measurement of Humidity”, The Institute of Measurement and Control, 1996, pp. 53-54.

[11] R. Comert, D. K. Matc, U. Avdan, “Object Based Burned Area Mapping with Random Forest Algorithm”, International Journal of Engineering and Geosciences, vol. 4, issue. 2, pp. 078 - 087, June 2019.

[12] P. Berrisford, P. Kallberg, S. Kobayashi, D. Dee, S. Uppala, A. J. Simsons, P. Poli, J. Sato, “Atmospheric conservation properties in ERA-Interim”, Quarterly Journal of the Royal Meteorological Society, vol. 137, issue. 659, July 2011.

[13] Z. Su, H. Hu, G. Wang, Y. Ma, X. Yang, F. Guo, “Using GIS and Random Forests to identify fire drivers in a forest city, Yichun, China”, Geomatics, Natural Hazards and Risk 2018, vol. 9, no. 1, pp. 1207 - 1229, 2018.