Fire Prediction in Southeast Asia

Sally Zhen (sqzhen@stanford.edu) & Krishna Rao (kkrao@stanford.edu)

I. Introduction

Land fires in Southeast Asia have increased dramatically over the past 30 years due to changes in land use and population density, draining of swamp forests, etc.. In 2015, Indonesian fires alone are estimated to have emitted as much carbon dioxide as Indias annual fossil fuel usage, and to have caused around 12,000 premature deaths[1,2]. The massive negative environmental and health impacts have prompted increasing interest from regional governments in better quantifying fire risks and identifying land management strategies. One of the major contributing factors is the equatorial Asian peatlands, which is one of the worlds biggest carbon sinks[3]. However, regional-scale investigation of peatland hydrology is made difficult by the location's inaccessibility of peat forests[3].

Therefore, in this project, we will use machine learning to predict local fire risks from remote sensing satellite data over the tropical peatlands of Borneo, Sumatra, and Peninsular Malaysia in 2015, one of the worst years. The input to our algorithm is remote sensing data collected over seven features–soil moisture, vegetation optical depth, specific humidity, temperature, precipitation, potential and actual evapotranspiration over 0.25^o x 0.25^o spatial grids in Southeast Asian peatlands. We then use multinomial logistic (Softmax) regression to output a predicted fire risk class. The objective of the study will be to do better than the already existing fire index predictor which is said to have an accuracy of 50%[20].

II. Related Works

Remote sensing data, or satellite digital imagery, has been used in geophysical and geological studies since the 1970s for land use management and ocean and forest monitoring[4,5,6]. In recent years, with dramatically improved precision in satellite imaging, remote sensing has been combined with field reconnaissance, Geographical Information Systems (GIS), and increasingly machine learning, to aid in deeper understanding of hydrology, urban development, and biome stability[7,8]. Certain studies have focused on knowledge-base building: Huang et al. used an inductive learning algorithm to generate production rules for a expert image analysis system from GIS training data[9]. Other studies target specific geophysical feature inference and fire hazard mapping: Sunar et al. carried out forest fire analysis from satellite, topographical, and meteorological data using maximum likelihood classification and multilayer feed-forward neural networks [10]; Chuvieco et al. integrated Thermatic Mapper data with other layers of geographic information to derive a forest fire hazard map [11].

Recently, an increasing number studies are turning to the investigation on MLAs themselves: Cracknell et al. compared five MLAs including Naive Bayes, k-Nearest Neighbors, Support Vector Machines (SVM), Neural Networks, and Random Forests (an MLA built on a multitude of decision trees) in the task of supervised lithology classification from spatially constrained remote sensing data, concluding the latter to be superior in this particular task[12]; Ahmad et al. showed that SVM model better predicts and captures soil moisture variability compared to feed forward-back propagation Artificial Neural Networks (ANN) and Multivariate Linear Regression model (MLR)[13]; Pals results suggested Random Forests classifier performed equally well to SVM in terms of classification accuracy and training time, but required fewer and simpler user-defined parameters[14]; etc. With rapid technological and theoretical progress in both fields, remote sensing and machine learning will continue to grow as a powerful combination.

III.Data

A. Fire History

We obtained remote sensing fire history data over Borneo, Sumatra, and Peninsular Malaysia between January and December 2015, from GFED (Global Fire Emissions Database)[15]. The data come in the form of area with fire activity in hectares in each 0.25^o longitude x 0.25^o latitude geographical grid element, which we term a pixel or grid. The data was given by GFED one of three classes: no fire (zero fire activity in pixel), small fire (pixel area burned between zero and 1%), and large fire (pixel area burnt between 1% and 40%, the maximum burn fraction of any pixel). Therefore it makes sense to transform the fire response of each pixel to the area fraction with fire activity, or burn fraction.

The data was processed into a 103 (longitude) x 56 (latitude) x 12 (number of months) matrix that represents pixel burn fractions of the 25.75^o longitude x 14^o latitude rectangular earth surface containing Borneo, Sumatra, and peninsular Malaysia. Keeping peatland pixels only, removing ocean pixels and invalid entries, and flattening resulted in a response vector of approximately 304 burn fractions, which we then transformed into a response vector of fire classes–no fire (zero burn fraction), small fire(0.01-0.40 burn fraction), (Although we realize flattening the response space loses geographically discerning information such as spatial proximity correspondence, land-ocean edge effects, and altitude, we try to remedy this by including features that capture as much as such information as possible (see Features.

B. Features

We selected seven features that in our opinion best collectively capture the local-scale fire-activity related geophysical attributes. We draw our training feature data from satellite remote sensing, in combination with gauge measurements and land surface model data. Our feature space consists of:

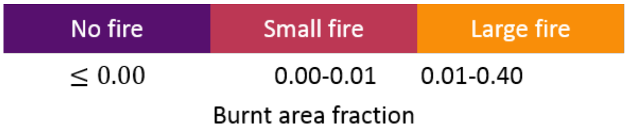


Fig. 1: Definition of fire classes. Color for each class used henceforth to label classes. Burnt area fraction is area of a cell affected by fire relative to total area of grid cell.

- Average monthly precipitation (precip) in mm/day in obtained from NASA covering the time period of interest. Precipitation data was collected via merged satellite and gauge measures and is in the desired temporal and spatial resolution[16];
- Average monthly temperature (temp), specific humidity (q), potential evapotranspiration (pet), actual evapotranspiration (et), and soil moisture (sm) obtained from GLDAS (Global Land Data Assimilation System)[17]. These data sets were collected via land surface models, specifically land data assimilation, which constantly updates the state parameters using observations, making them achieve very high accuracy. These data is in the form of desired temporal and spatial resolution;
- Vegetation Optical Depth (VOD) is an indicator of vegetation water content. We obtained VOD data from the Numerical Terradynamics Simulation Group at University of Montana[6]. The data was collected through remote sensing twice daily over elevation-adapted 0.25 decimal degrees grids. We transformed the data to monthly resolution re-scaled pixels to a uniform 0.25^o grid.

For each feature, we downloaded relevant data, retrieved and flattened the data subset according to indices of the time period and geographical area of interest, and performed 2D linear interpolation spatially to obtain features data at desired resolution. We then re-scaled each feature linearly between zero and one. After preprocessing, our training examples are represented in an m (304) x n (7) matrix, corresponding to a 304-value response vector. We randomly split the examples into 70% training set (213 examples), on which we trained our model, and 30% validation set (91 examples), on which we tested model performance. A training example may look that given in Fig. 1.

IV. Methods

In this analysis, we used multinomial logistic regression to classify the pixels as no fire, small fire or large fire. We chose logistic regression because after preliminary inspection of the features, it was clear that there is strong overlap between the labels in the feature space. Moreover, the response function seemed to be non-linear with respect to some of the features. Knowing that logistic regression is robust for capturing non-linear relationships and does not hold any assumption of normal distribution of features, it was the best choice. To perform the classification using the input features, we used scikit-learns logistic regression with cross validation module. The feature space represented a wide variety of climatic, geographic, and anthropogenic variables. So the magnitude and units of each feature is different. In order to compare the importance of different features, we scaled each feature between 0 and 1 using the formula:

$$X_i = \frac{X_i - min(X_i)}{max(X_i) - min(X_i)} \tag{1}$$

We split our data into training data, test data in the ratio 70-30. We chose this ratio because we wanted a large enough sample to train on before testing the model performance. We then performed multi-class logistic regression using 10-fold cross validation to tune the regularization parameter lambda. The goal of the model was to maximise the log-likelihood as defined by:

$$l(\theta) = \sum_{i=1}^{n} y_i log(p(x_i)) + (1 - y_i) log(1 - p(x_i)) \tag{2}$$

where,

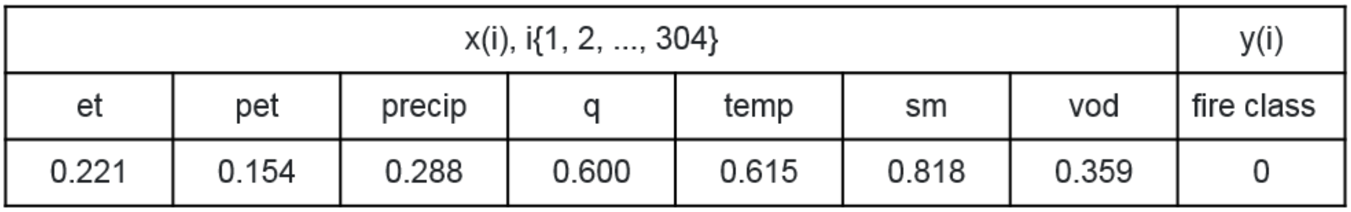
$$p(Y = c|\vec{X} = x) = \frac{e^{\theta^T x}}{\sum_c e^{\theta^T x}} \tag{3}$$

& $c \in \{$"no fire", "small fire", "large fire"$\}$.
We passed the outputs of the logistic regression through softmax function and predicted the label based on the maximum output. Once the model was trained and cross-validated, we tested the final performance of the model on the test (30% of data).

Additionally, we also tried Support Vector machines and Naive Bayes classification.

V. Results

The model predictions of the class labels were compared against the true labels and displayed in Fig. 3. The prediction accuracy was found to 0.71. Precision-recall curves computed after using one-hot function on the predicted labels. are shown in Fig. 4. The micro-averaged precision-recall area under the curve = 0.72. The comparison of weights from the model output after cross validation are shown in Fig. 4.



x(i), i{1, 2, ..., 304}							y(i)
et	pet	precip	q	temp	sm	vod	fire class
0.221	0.154	0.288	0.600	0.615	0.818	0.359	0

Fig. 2: Example of 1 row of data. Values are strictly for representation purposes only. $y^{(i)}$ fire class ∈ {0,1,2} is factorized version of true labels {”no fire”, ”small fire”, ”large fire”}

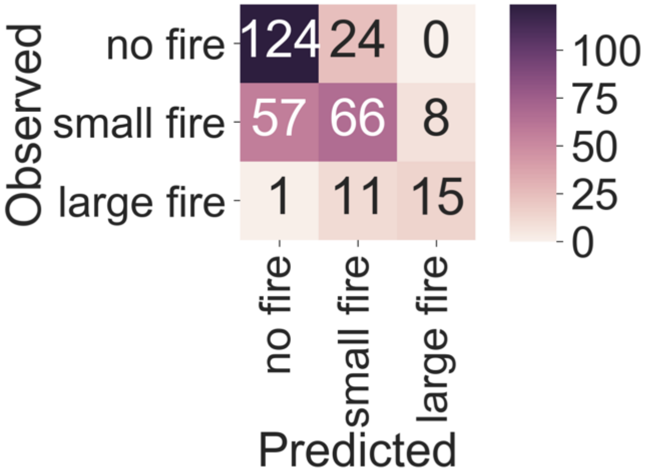


Fig. 3: Confusion matrix of actual labels and predicted labels. Strength of color is proportional to number of examples in each box. Number inside each box indicates number of examples.

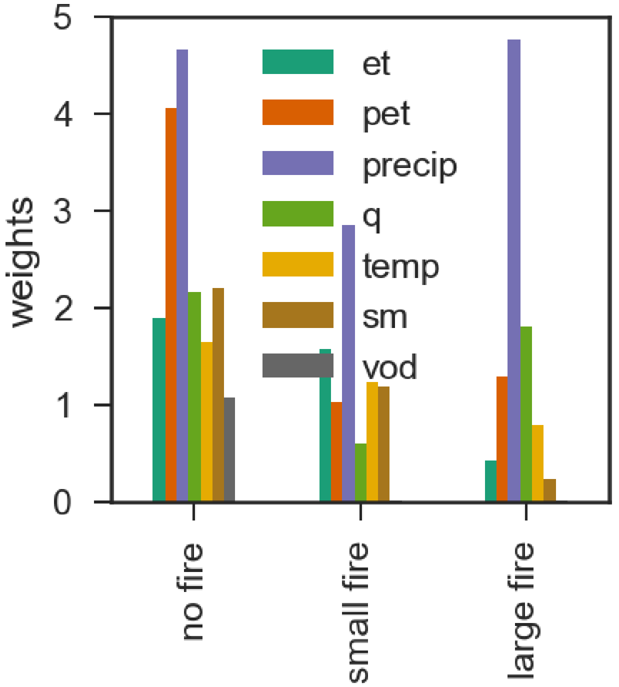


Fig. 5: Feature weights from classifier. Weights for each class used to calculate probability of output for respective class.

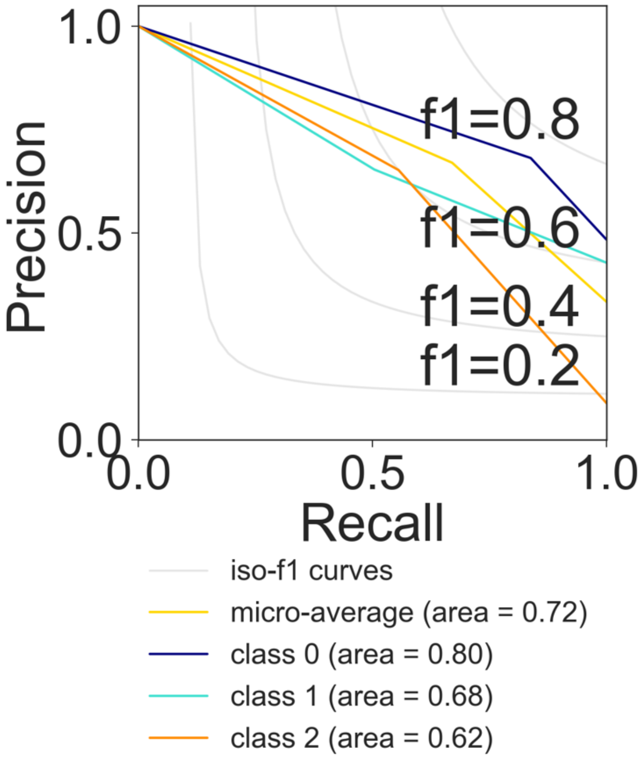


Fig. 4: Precision-recall curve of predicted labels. F_1 curve values indicate harmonic mean of precision and recall at respective value for area under the curve

The Naive Bayes classifier seemed to perform very poorly on the data with an accuracy of even less than 50% whereas the support vector machine algorithm was unable to perform better than the logistic regression classifier.

VI.Discussion

Even though there is a heavy overlap of class labels in the feature space, the classifier predicts fire with an accuracy of 71%. This is significantly better than the already existing classifier of the regional forest service which is said to predict fires with an accuracy of about 50%. The improvement may be partially attributed to the expansion of the feature space.

The existing model predicts fires based on a loop=up table of drought indices, thus basing the prediction off purely climatic factors.

The confusion matrix shown in Fig. 3 shows that the classifier is unable to efficiently separate ”small fire” pixels from ”no fire” pixels. This may be due to the similarity of climatic factors for these pixels. Also, the large resolution of 0.25^o could make the classifier unable to detect a separation boundary for these two classes.

The most important predictor as seen from Fig. 5, is precipitation for all the three classes. Physically, this follows from the mechanism of fire initiations. In our study area of peatlands in Borneo, the major cause of fire is drought driven.

The reason why logistic regression was able to outperform support vector machine based classifier could be the high overlap between the labels. Because of the high overlap, it was more difficult for model to choose support vector resulting in decreased accuracy. However, the accuracy of SVM in predicting large fires was substantially better than that of logistic regression which follows from our expectation. Due to the relatively low number of large fire pixels, the logistic regression is unable to maximize the log likelihood for the respective class efficiently whereas SVM is unaffected by the large number of ”no fire” or ”small fire” pixels. The separation boundary is solely dependent on the chosen support vectors.

VII. Contributions

- **Sally**: Data downloading, 2D spatial interpolation, temporal scaling, pre-processing, SVM, Poster
- **Krishna**: Linear regression, Locally weighted linear regression, Logistic regression, Cross-validation, Performance quantification, Plots

VIII. Conclusion

Land fires in Southeast Asia have increased dramatically over the past 30 years. Predicting fires have become more important than ever due to a sudden increase in fatalities, injuries and losses incurred to properties. Peatlands in Borneo play a major role in initiating these fires. We used geographic, climatic, and anthropogenic factors to build a predictive classification model using previously recorded fires. In spite of the constraints imposed by resolution of remote sensing data, uncertainty in climatic variables and multiple causes leading to fires, We were able to successfully predict the fires with an accuracy of 0.72. In the future, if access to higher resolution data is available through Synthetic Aperture Radars, potentially more accurate predictions can be made using machine learning.

IX. Data Accessibility

All scripts for analysis performed by us can be found at https://github.com/kkraoj/Forest-Mortality/blob/master/fire_model.py

References

[1] Lee, Hsiang-He, Rotem Z. Bar-Or, and Chien Wang. ”Biomass burning aerosols and the low-visibility events in Southeast Asia.” Atmospheric Chemistry and Physics 17.2 (2017): 965-980.

[2] Crippa, P., et al. ”Population exposure to hazardous air quality due to the 2015 fires in Equatorial Asia.” Scientific reports 6 (2016): 37074.

[3] Miettinen, Jukka, Chenghua Shi, and Soo Chin Liew. ”Land cover distribution in the peatlands of Peninsular Malaysia, Sumatra and Borneo in 2015 with changes since 1990.” Global Ecology and Conservation 6 (2016): 67-78.

[4] Shearman, P. L., et al. ”The state of the forests of Papua New Guinea.” Mapping the extent and condition of forest cover and measuring the drivers of forest change in the period 2002 (1972): 148.

[5] Rango, Albert. ”Applications of remote sensing to watershed management.” (1975).

[6] Wilson, W. H., and R. W. Austin. ”Remote sensing of ocean color.” Ocean Optics V. Vol. 160. International Society for Optics and Photonics, 1978.

[7] Huang, Xueqiao, and John R. Jensen. ”A machine-learning approach to automated knowledge-base building for remote sensing image analysis with GIS data.” Photogrammetric engineering and remote sensing 63.10 (1997): 1185-1193.

[8] Keane, Robert E., Robert Burgan, and Jan van Wagtendonk. ”Mapping wildland fuels for fire management across multiple scales: integrating remote sensing, GIS, and biophysical modeling.” International Journal of Wildland Fire 10.4 (2001): 301-319.

[9] Huang, Xueqiao, and John R. Jensen. ”A machine-learning approach to automated knowledge-base building for remote sensing image analysis with GIS data.” Photogrammetric engineering and remote sensing 63.10 (1997): 1185-1193.

[10] Sunar, F., and C. zkan. ”Forest fire analysis with remote sensing data.” International Journal of Remote Sensing 22.12 (2001): 2265-2277.

[11] Chuvieco, Emilio, and Russell G. Congalton. ”Application of remote sensing and geographic information systems to forest fire hazard mapping.” Remote sensing of Environment 29.2 (1989): 147-159.

[12] Cracknell, Matthew J., and Anya M. Reading. ”Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information.” Computers Geosciences 63 (2014): 22-33.

[13] Ahmad, Sajjad, Ajay Kalra, and Haroon Stephen. ”Estimating soil moisture using remote sensing data: A machine learning approach.” Advances in Water Resources 33.1 (2010): 69-80.

[14] Pal, Mahesh. ”Random forest classifier for remote sensing classification.” International Journal of Remote Sensing 26.1 (2005): 217-222.

[15] Randerson, J.T., G.R. van der Werf, L. Giglio, G.J. Collatz, and P.S. Kasibhatla. 2017. Global Fire Emissions Database, Version 4.1 (GFEDv4). ORNL DAAC, Oak Ridge, Tennessee, USA.

[16] https://pmm.nasa.gov/data-access/downloads/trmm

[17] https://hydro1.gesdisc.eosdis.nasa.gov/data/GLDAS/GLDAS_NOAH025_M.2.1/2000/

[18] http://files.ntsg.umt.edu/data/AMSR_Results/AMSR_E_2_v2/2002/

[19] https://ngdc.noaa.gov/eog/dmsp/downloadV4composites.htmlAVSLCFC

[20] Alexander, M. E. (1982). Calculating and interpreting forest fire intensities. Canadian Journal of Botany, 60(4), 349-357.