Satellite-Based Prediction of Fire Risk in Northern California
Final Report

Caroline Famiglietti (SUID: 06272576; **CS229**),
Natan Holtzman (SUID: 06273767; **CS229**),
& Jake Campolo (SUID: 06165559; **CS229A**)

December 12, 2018

---

1 Introduction

In recent decades, climate change has drastically influenced key characteristics and patterns of wildfire across the global land surface[1]. In California, wildfires are increasing in magnitude, scale, frequency, and duration, thus heightening risks to human populations and ecosystems[2]. As temperature rises and water availability shifts with climate change, such effects are only expected to intensify. Vast sectors of the state encompassing dense population centers, key agricultural regions, and biodiversity hotspots are left in critical positions. A clearer understanding of the patterns of fire development and spread in California, including the spatial distribution of vulnerability to disturbance, is vital in adaptively managing land and resources. The danger and high economic and social cost of uncontrolled wildfires motivated us to pursue a strategy of fire risk assessment, rather than post-fire classification[3], which could potentially be used to aid fire prevention policy. Our goal in this project was to develop a regionally flexible and scale-able method to predict fire risk based on local climate and land surface conditions. Specifically, we input remotely-sensed measures of precipitation, temperature, soil moisture, evapotranspiration, drought, wind speed, land cover type, and vegetation characteristics to three models: logistic regression, gradient-boosted decision trees, and a multilayer perceptron. The output of interest is a probability of fire ranging from 0 to 1, which we interpret as fire risk from 0 to 100%. While we were able to predict non-fire more accurately than fire, predictions for the latter retained 75–80% overall accuracy. Our work demonstrates the strong potential of using remote sensing assets to preemptively identify fire risk and inform prevention efforts in the coming years and decades.

2 Related Work

Much of the prior work in this field has focused on mapping burned area and fire severity after a fire has already occurred[3,4,5]. Eidenshink et al. (2007) mapped burn severity, a measure of fire intensity and residence time, using differences in the Normalized Difference Burn Ratio (dNBR) index calculated from the Landsat Thematic Mapper at 30m resolution; this index approximates vegetation growth if negative and mortality if positive. GIS analysts then manually classified fire severity by comparing differences in dNBR to a synthesized database of historical fires in the United States. Hawbaker et al. (2017) improved upon this method by incorporating raw reflectance bands (red, blue, green, infrared, etc.) as well as additional reflectance indices (normalized difference vegetation index, wetness index, moisture index, tasseled cap greenness) in their analysis, and using a gradient boosted regression model for automated classification rather than manually estimating fire severity. They mapped 116% more burned area than in Eidenshink et al., and their method is more easily adaptable to other regions, while still being implemented wholly with Landsat data. Parks et al. (2016) modeled fire severity using boosted regression trees with the following fire-related climate variables as input data: evapotranspiration, water deficit, annual precipitation, soil moisture, and snow water equivalent. However, their analysis was aggregated to the level of contemporary fire severity (1984-2012) and used to estimate mid-21st century (2040–2069) fire severity based on simulated global climate models rather than predict daily fire risk. For our project, we adapted and combined some of the best aspects of these works. We used fire classifications derived from pre- and post-fire dNBR, following the method pioneered by Eidenshink et al., as the target variable in our model training. We compiled a comprehensive dataset of potential predictors of fire risk

based on the remotely-sensed reflectance and climate variables used by Hawbaker et al. and Parks et al., respectively, so that we have a greater chance of capturing the causes of fire risk while still being regionally adaptable. Additionally, we followed similar methods of these latter two studies by using machine learning to automate prediction, but adapted them to our goal of daily fire risk prediction.

3 Dataset

Our dataset was assembled in the Google Earth Engine platform[6], which allows for parallelized cloud computing on its large collection of geospatial data. The primary data sources used were the Moderate Resolution Imaging Spectroradiometer (MODIS) aboard NASA's Terra/Aqua satellites, the Global Land Data Assimilation System product (GLDAS), and the Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS). Using remote sensing data ensures that our model can be easily be extended to other regions and times (as long as similar remote sensing instruments are in operation at those times). We computed a collection of fire-relevant variables describing ecological, hydrological, and meteorological conditions from these gridded datasets for our study region of Northern California and for the time period of 2001–2017. These choices were informed by our review of the related literature. For the purposes of training and validation, we used a MODIS fire severity classification product developed from post-fire spectral signals to engineer a binary "fire/no fire" target variable. Table 1 contains the full list of variables fed to our models, including their sources and spatial & temporal resolutions; Figure 1 depicts an example input variable.

Table 1: Variables used for fire prediction. Climate variables are italicized; reflectance variables are not italicized.

Variable	**Source**	**Spatial Res.**	**Temporal Res.**
Evapotranspiration (ET)	GLDAS	27.5 km	3-hourly
Precipitation	CHIRPS	5.5 km	Daily
Palmer Drought Severity Index (PDSI)	U of Idaho	4.6 km	10-day
Soil moisture, 0-10cm	GLDAS	27.5 km	3-hourly
Soil moisture, 10-40cm	GLDAS	27.5 km	3-hourly
Soil moisture, 40-100cm	GLDAS	27.5 km	3-hourly
Soil moisture, 100-200cm	GLDAS	27.5 km	3-hourly
Wind speed (WS)	GLDAS	27.5 km	3-hourly
Land surface temperature (LST)	MODIS	1 km	Yearly
Reflectance (Red)	MODIS	1 km	Daily
Reflectance (Green)	MODIS	1 km	Daily
Reflectance (Blue)	MODIS	1 km	Daily
Reflectance (Near-infrared)	MODIS	1 km	Daily
Reflectance (Shortwave infrared 1)	MODIS	1 km	Daily
Reflectance (Shortwave infrared 2)	MODIS	1 km	Daily
Normalized difference vegetation index (NDVI)	MODIS	1 km	Daily
Green chlorophyll vegetation index (GCVI)	MODIS	1 km	Daily
Normalized difference moisture index (NDMI)	MODIS	1 km	Daily
Normalized difference wetness index (NDWI)	MODIS	1 km	Daily
Tasseled cap greenness (TCG)	MODIS	1 km	Daily
Tasseled cap wetness (TCW)	MODIS	1 km	Daily
Normalized burn ratio 1 (NBR1)	MODIS	1 km	Daily
Normalized burn ratio 2 (NBR2)	MODIS	1 km	Daily
Land cover classification (LC)	MODIS	1 km	Daily
Presence of fire	MODIS	1 km	Daily

Because we wanted to characterize short-term conditions preceding ignition days, we calculated a maximum of 4 variations (or 'lag periods') relative to the day of fire data for each variable: a 1-day lag, 1-week lagged average, 1-month lagged average, and 3-month lagged average. The number of lag periods calculated was dependent on temporal resolution of the data record for that variable. Each lag period was calculated for the current year as well as calculated and averaged over the years 1985 to 2005, in order to develop a 20-year climatology of the variable and thus differentiate anomalies from historical means. Finally, our sampling sampling strategy consisted of iterating through daily time steps and extracting each variable at every "fire" pixel, along with an approximately equal number of randomly selected "non-fire" pixels, into a tabular dataset. The resulting dataset had dimension $903,921 \times 112$ given all pixel-days and features considered.

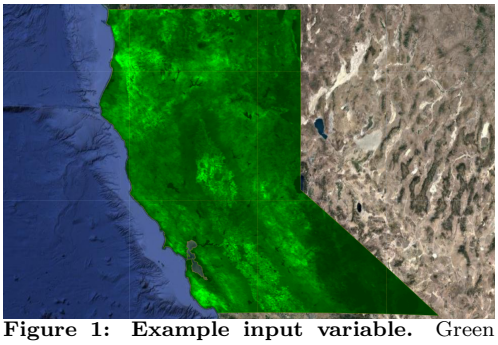

Figure 1: Example input variable. Green Chlorophyll Vegetation Index (GCVI) at 1 km resolution ranges from 0 to 6 (dark to light green). Shown for July 15, 2018.

For post-processing, we omitted records of already-burning fires, keeping only the first fire occurrence in any given year at each pixel. Had we not omitted these records, the associated lag variables would be "contaminated" by the ongoing fire. These data points are essentially duplicates of the initial day of fire at that pixel and thus add no new information to the data. More seriously, including these points in our training data might yield models trained to detect persistent burning rather than to learn likelihoods of ignition based on pre-fire conditions. The former is not our goal in this project. After removing already-burning fires, we were left with 561,662 examples.

4 Algorithms

4.1 Logistic regression with forward stepwise selection

Logistic regression is a linear classifier that outputs a probabilistic prediction $h(x) = g(\theta^T x)$ where g is the logistic function $g(z) = 1/(1 + e^{-z})$. The model is trained to maximize the likelihood assuming that $y \sim \text{Bernoulli}(p = h(x))$. We used the glm function in R, which uses the Newton-Raphson method for optimization[7]. We fit the model using all variables, but also experimented with forward stepwise feature selection, where one feature was added to the model at a time according to which added feature lowered the cross-entropy loss the most. We added 10 features one by one and then used the number of of features that maximized accuracy on a validation set as a final model. We used years 2001-2015 for training, 2016 for validation, and 2017 for testing.

4.2 Decision trees with gradient boosting

In gradient boosting, the n^{th} model in the ensemble is fitted on the residual of the previous $n-1$ models. The splits on each model are chosen to minimize the cross-entropy loss $-(y \log p + (1-y) \log(1-p))$ at every split. We used the XGBoost R package[8] to fit 100 gradient boosted trees with a maximum depth of 2. To avoid overfitting, we chose the number of trees to use in our final model based on which number of trees gave the highest validation accuracy (32 trees when both climate and reflectance subsets were included). We again used years 2001-2015 for training, 2016 for validation, and 2017 for testing.

4.3 Multilayer perceptron

A multilayer perceptron (MLP) is a feedforward neural network composed of an input layer receiving the signal, an output layer making a prediction about the input (in our case, assigning a 'fire' or 'no fire' label), and any number of hidden layers with non-linear activation functions. Our MLP uses the ReLU activation function ($f(x) = \max(0, x)$) and stochastic gradient descent for optimization, which we implemented using the scikit-learn Python library[9]. For classification, it minimizes the cross-entropy loss function. The model consists of two hidden layers with 10 and 2 neurons, respectively. We used the years 2001-2016 for training and 2017 for testing.

5 Experiments, Results & Discussion

We built machine learning models to predict the probability that a fire will ignite on any pixel on any given day, given prior climate and land surface conditions derived from remote sensing data. As

described in section 4, the three models considered perform binary classification with cross-entropy loss. We used models that output a probability, excluding methods such as support vector machines. We trained each model on three distinct sets of features: climate variables only, reflectance variables only, and both subsets.

We first considered logistic regression. Because the logistic regression models trained on each set of data had test accuracies of between 0.76 and 0.78, we found it more informative to compare their values of AUC (area under the Receiver Operating Characteristic curve, which plots true positive rate against false positive rate as the threshold for being predicted as a positive example is changed). Forward feature selection did not improve test performance; we achieved an AUC of 0.74 with all features, but 0.68 with the four selected features. This contrast suggests that the conditions that contribute to fire are too complex to be described with only a small number of variables. However, feature selection provides some insight into which specific variables are important. In particular, land cover classification, 1-week GCVI, 3-month SWIR2, and 1-week ET were selected. The fact that remote sensing indicators were chosen before climate variables suggests their abilities to capture the fire vulnerability state of the land system in a way that is not easily obtained from knowledge of the climate forcing. This finding is confirmed by comparing the AUCs of logistic regression models trained on the two types of variables: 0.68 and 0.72 for the climate and reflectance subsets, respectively. Reflectance alone can predict fire more accurately than climate alone, but we achieve the best results (AUC of 0.74) when including both sets of variables.

Table 2: Percent errors by model for training and testing. Model labels *(C)*, *(R)*, and *(C, R)* indicate which subset of data was used (climate, reflectance, or both). The baseline error achieved by classifying all test examples as "no fire" was 27.19%.

Model	**Train Error (%)**	**Test Error (%)**
Logistic regression (C)	20.62	22.96
Logistic regression (R)	20.93	23.24
Logistic regression (C, R)	21.33	22.71
Boosted trees (C)	21.07	23.11
Boosted trees (R)	19.57	22.62
Boosted trees (C, R)	20.38	22.37
MLP (C)	25.20	24.69
MLP (R)	20.51	22.72
MLP (C, R)	18.97	22.21

Both the gradient boosted decision trees and the MLP showed similar results to those of logistic regression *(Tables 2–4)*. In particular, the boosted trees yielded AUCs of 0.70, 0.72, and 0.75 for climate, reflectance, and both respectively. Notably, the minimum test error was found with the MLP *(C, R)*, but overall variations in test error are small across all models and subsets *(Table 2)*. Additionally, all models on all subsets surpass baseline performance of 27.19% error. Because two nonlinear algorithms, trees and neural networks, had only slightly higher accuracy than logistic regression with no interaction terms, we are led to believe that there is minimal nonlinearity in the processes that influence fire risk. The most important features found in in the boosted trees were land cover, climatological 1-month precipitation, 3-month LST, and climatological 3-month wind speed. The fact that these are different from the variables found in logistic regression feature selection except for land cover may be due to many features being collinear, which made forward selection sensitive to noise.

Table 3: Confusion matrices for (a) logistic regression *(C, R)*, (b) boosted trees *(C, R)*, and (c) MLP *(C, R)*.

(a)	*(Pred.)* No Fire	*(Pred.)* Fire
(Actual) No Fire	26,996	1,387
(Actual) Fire	7,170	2,134

(b)	*(Pred.)* No Fire	*(Pred.)* Fire
(Actual) No Fire	26,453	1,930
(Actual) Fire	6,502	2,802

(c)	*(Pred.)* No Fire	*(Pred.)* Fire
(Actual) No Fire	26,755	1,313
(Actual) Fire	7,249	2,055

Finally, using the coefficients from *(C, R)* logistic regression with feature selection, we have mapped fire risk across our domain for the summer of 2017 *(Figure 2)*. The fact that the Central Valley, an area dominated by cropland, stands out as most vulnerable to fire points to a key limitation of our dataset: wildfires are not distinguished from anthropogenically caused fires (*e.g.* crop residue burning, with which

Table 4: Additional statistics for the three models *(C, R)*.

	Precision	**Recall**	**F1**
No Fire *(Logistic regression)*	0.80	0.95	0.33
Fire *(Logistic regression)*	0.61	0.23	0.86
No Fire *(Boosted trees)*	0.80	0.93	0.86
Fire *(Boosted trees)*	0.59	0.30	0.40
No Fire *(MLP)*	0.80	0.94	0.86
Fire *(MLP)*	0.61	0.28	0.38

we expect many observed fires in the Central Valley are associated). This result echoes that of our feature importance determination; land cover was the single most important feature for both logistic regression and the boosted trees. In particular, whether or not a given pixel was located in cropland was crucial. Thus while our models have indeed identified fires, they have not differentiated between fire sources. A remedy to this problem could involve filtering out cropland pixels. However, for the purposes of this project, such filtering was not performed.

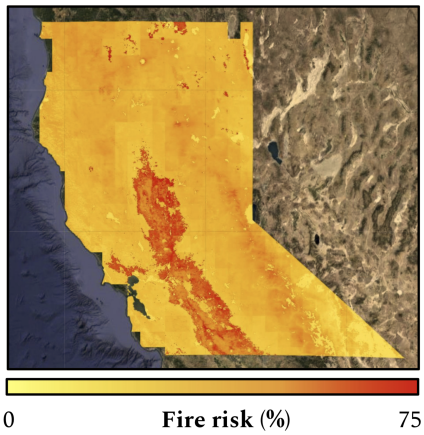

Figure 2: Map of average fire risk for the period June–August 2017. Risk was derived using coefficients from logistic regression *(C, R)*.

6 Conclusions & Future Work

Our work demonstrates the vast potential for using machine learning techniques to better understand vulnerabilities to fire based on features of climate and the land surface. We applied logistic regression with forward stepwise selection, decision trees with gradient boosting, and a multilayer perceptron to a robust dataset of ecological, hydrological, and meteorological variables derived largely from remote sensing. Performance across models was similar, though the multilayer perceptron using the full dataset (both climate and reflectance subsets) provided the overall minimum test error. In general, while we were able to predict non-fire conditions more accurately than fire conditions, predictions for the latter retained 75–80% overall test accuracy, measurably better than the 73% that would be achieved with random guessing. We believe that by coupling remote sensing assets with learning algorithms, preemptive identification of fire risk with the goal of adaptive resource management is possible.

Two future directions are particularly appealing to us. We are first interested in using projections of future climate to infer expected wildfire dynamics in the region in the coming decades. To address this, we could leverage climate model outputs under the RCP 4.5 or 8.5 emissions scenarios (which forecast climate change from the present to 2100), and modify our input variables to reflect these probable shifts. We could then apply our models to the predicted data to assess future changes in fire risk over the next century. The second future direction of interest to us involves predicting post-disturbance effects (in particular, economic impacts or costs of damages) on burned areas. We could compile spatially-explicit population, infrastructure, and economic data in order to extrapolate potential costs and damages from fires predicted by our model, which would be useful for policymakers in conducting cost benefit analysis of fire prevention.

Contributions

All team members contributed to the planning of experiments, determination of project foci, and writing of the report. Additionally, all team members collaborated in determining dataset structure (*e.g.* relevant variables and data sources), analyzing data, and creating the poster. Jake Campolo prepared data and created map-based visualizations; Caroline Famiglietti implemented the multilayer perceptron, tabulated outputs, and led overall design choices; Natan Holtzman implemented the logistic regression and boosted trees, and analyzed, interpreted, and visualized project results.

Code

Our code for this project is available at `https://github.com/cfamigli/229.git`.

References

[1]Pechony O, Shindell DT (2010). Driving forces of global wildfires over the past millennium and the forthcoming century. PNAS, http://www.pnas.org/content/107/45/19167

[2]Westerling AL, Hidalgo HG, Cayan DR, Swetnam TW (2006). Warming and Earlier Spring Increase Western U.S. Forest Wildfire Activity. Science, http://science.sciencemag.org/ content/313/5789/940.full

[3]Hawbaker TJ, Vanderhoof MK, Beal Y-J, Takacs JD, Schmidt GL, Falgout JT, Williams B, Fairaux NM, Caldwell MK, Picotte JJ, Howard SM, Stitt S, Dwyer JL (2017). Mapping burned areas using dense time-series of Landsat data. Remote Sensing of Environment, https://doi.org/10.1016/j.rse.2017.06.027.

[4]Eidenshink J, Schwind B, Brewer K, Zhu Z, Quayle B, Howard S (2007). A Project for Monitoring Trends in Burn Severity. Fire Ecology, https://doi.org/10.4996/fireecology.0301003

[5]Parks SA, Miller C, Abatzoglou JT, Holsinger LM, Parisien MA, Dobrowski SZ (2016). How will climate change affect wildland fire severity in the western US? Environmental Research Letters, https://doi.org/10.1088/1748-9326/11/3/035002

[6]Google Earth Engine, https://earthengine.google.com/

[7]R Core Team (2018). R: A Language and Environment for Statistical Computing. https://www.R-project.org

[8]Chen T. et al. (2018). xgboost: Extreme Gradient Boosting. R package version 0.71.2. https://cran.r-project.org/web/packages/xgboost/index.html

[9]Pedregosa F et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research.