Analyzing Feature Importance of Random Forests to Define Key Variables in Post-Wildfire Regeneration

Written by Phil Hilt and Sean Lord

University of Colorado Colorado Springs
1420 Austin Bluffs Pkwy,
Colorado Springs, CO 80918

Abstract

Wildfires are common each year across the world and can have a significant impact on an ecological area. Some ecological regions can see vegetation regrow quickly while in some areas, vegetation can take decades to regrow. Machine Learning (ML) is used in various facets of wildfire science and supervised ML has recently been used as an application in assessing the ecological impact of wildfires. In this paper a method was proposed to quantify the ecological effect of wildfires by identifying key variables that contribute to vegetation regrowth. Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) data collected after the 2013 California Rim Fire was used to build Random Forests (RF) to predict the Normalized Burn Ratio (NBR) of the effected area. 13 other spectrometer bands were used as a feature set. The RF model performed relatively well with a Mean Square Error of 1.333 and a coefficient of determination of 0.866. Feature analysis on the Decision Trees of the RF indicated that Moisture Stress Index (MSI) was the most significant variable from the feature set. By utilizing larger datasets of time-series remote sensing data, we believe the framework used in this project can be extended to practical applications to better understand the driving forces behind post-wildfire recovery.

Introduction

Each year, wildfires are caused from a multitude of causes. With increasing global temperatures the number of annual catastrophic wildfires has increased. These fires can pose a significant threat to wildlife, the environment, property and even human life. In recent decades researchers have started to use ML methods to assist in wildfire science. ML methods have been used to better understand domains such as fire susceptibility, severity mapping, and wildfire preparedness. For this paper we will focus on supervised ML methods to quantify and and better understand variables that drive an ecological area's ability to regenerate and regrow in the time period following a wildfire.

A variety of supervised ML models have been successfully used in the application of post-fire regeneration prediction, including Artificial Neural Networks (ANN) and Random Forests (RF). For general ecological applications, RF has become a preferred model due to its flexibility and ability to handle both categorial data and linear data. And after reviewing similar projects and viable datasets we decided to move away from ANN and focus entirely on RF. The approach uses a RF model in order to extract relative feature importance of different stress factors that may impact vegetation growth. Our model predicts the Normalized Burn Ratio (NBR) of a spatial area. By analyzing the frequency of features in top level nodes of the Decision Trees (DT) within the RF model, and by analyzing the Mean Decrease of Impurity (MDI) of the DTs our framework can provide insight into which variables contribute the most to vegetation recovery. The primitive method can be extended to larger time-series datasets to provide greater understanding of a post-wildfire area's ability to regrow. With this approach agencies could take better care to practice responsible forestry and agricultural practices, as well as giving us better metrics by which to measure the overall health of biomes.

1 Background

As technological advances, researchers have improved access to data. Further advancements in computer hardware have opened up new possibilities for applications of ML. Remote monitoring allows for advanced data collection with which ML can be utilized to draw new insights.

1.1 Remote Monitoring Techniques

Remote monitoring techniques provide a large amount of data to help understand the surface of the Earth. There is a great need to understand how to effectively analyze Earth surface data to draw conclusions that can help better understand ecosystems, best agricultural practices, and stewardship of the Earth's climate as a whole.

Satellite data is perhaps the most prevalent and has the advantage of providing consistent and constant time series measurements. The shortcoming of satellite data is that it typically only allows for linear analysis and the bands can have relatively low resolution when compared to methods like LiDAR. LiDAR sensing techniques are also capable of providing categorical data that can give insight into the specific species of plant vegetation in an area by accurately measuring canopy depth (Debouk et al., 2013). Additional field measurements like soil samples and in person classification of plant species are also incredibly useful, however, the downside of field samples and LiDAR is that datasets are not continuous and data collection is often more expensive.

There are also a number of agencies that perform remote monitoring using flyover data collection. While these methods do not have the continuous time series measurements as satellite data, they provide greater spatial resolution and often can provide greater accuracy. Common flyover techniques measure wave lengths reflecting off of the Earth's surface by measuring and collecting visible infrared data, LiDAR data, and MODIS data.

1.2 Random Forests in Ecology

For domain specific applications in ecology, RF have become a favored ML model for researchers. Ecological studies often require data sets that are both categorical and linear in nature. RF are useful due to their abilities to handle multiple datasets and analyze variable significance, all while having high accuracy (Cutler et al. 2007). Furthermore, RF can be used for both supervised and unsupervised ML analysis which can be used in a wide range of applications.

2 Related Work

A large number of studies have been done to better understand potential relationships between forest fires and environmental variables. Areas of focus often include severity of burn, fire susceptibility, fire intensity, and landscape controls on a fire. There is a general need to develop frameworks that can merge parametric time-series data with categorical data and while there are a growing number of studies on post fire regeneration, this area is lacking.

A review from *Environmental Reviews* that identified 300 publications relevant to wildfire science, determined that RF was the most widely used model for post-wildfire regrowth prediction, and RF also saw the most general success (Piyush et al. 2020). The report also indicated that ANN has been known to be adequate in similar applications, but RF have had excellent historical success with basic ecological problems dealing with classification and regression (Cutler et al., 2007).

2.1 Studies Incorporating Regression and Random Forest

A publication from *Ecological Indicators* implemented RF to assign three indicator indices to quantify the regrowth potential of an area following a wildfire. A long-term, mid-term, and short-term index was generated using RF with the long-term and mid-term indices proving to be more accurate by including the features of fire traits, post-fire climatic conditions and plant life-history traits (Torres, et al. 2018). The mid-term index used was Recovery Trend Index (RTI). The RTI implementation utilized a Theil-Sens estimator to map Normalized Difference Vegetation Indices (NDVI) from time-series satellite data to a regressive slope value. For the purpose of clarity RTI and Theil-Sens will be used interchangeably throughout the rest of this paper. The research from this paper also stresses the importance of including a variety of independent variable features in the data collection process, but proves that RF is an optimal method for quantifying post-wildfire regrowth using remote sensor data.

A separate study from *Frontiers in Fire Ecology* analyzed the recovery of three separate tree species from different fires across the Western United States (Bright et al.,2019). This study used Normalized Burn Ratio (NBR) to label the RF model, which was ultimately used to predict and better understand how the three species were recovering. The end result provided insight into useful features and data collection techniques.

3 Methodology

Due to the robust nature of RF and its known success rate with similar application, RF was the model used for this project. The original scope of this project sought to compute a Theil-Sens RTI from NDVI time series data to use as a label. Theoretically, a number of supervised ML models would be able to predict an RTI base on feature attributes, but they all require a large dataset that is taken consistently for a period of time to be viable. After spending some time finding valid datasets, this approach was abandoned and the model was scaled back. A dataset that included flyover measurements was used for the model and instead of using an RTI, we used NBR to label the data as this label was used successfully in the study done by Bright et al. in *Frontiers in Fire Ecology*.

3.1 Finding a Significant Wildfire Event

To find a valid dataset we needed a to find a significant wildfire that had been well studied. The California Rim Fire was a wildfire that started on August 17, 2013 and would burn over 400 square miles of forest (Rim Fire Information, 2013). At the time of its occurrence, it was the third largest wildfire in California's recorded history. We chose this wildfire as the primary source for our project, due to its size, age, and the high degree of interest it has received from the scientific field since. There has been a great deal of research on the effects of the Rim fire, the ecological causes behind the fire, and the area's recovery. The high degree of interest in the fire translated to a wide breadth of accessible data and information that we used to train our model, and better optimize the project.

3.2 Finding a Dataset

We originally looked into two primary datasets to be utilized for this project. The first was a set of data acquired off Kaggel by multiple authors that all utilize the same source formatting. Unfortunately, the Kaggel open source data sets would not prove sufficient for our needs due to a lack of features and inconsistency with the time ranges. Our second set of data was utilizing Landsat data provided through NASA. This data is available through several client searches open to the public. Two client searches we used for exploratory analysis was the USGS EarthExplorer and NASA EarthData. Both client searches allow users to find granularity geographic areas from specified satellites. For our project we decided to try to use measurements from Moderate Resolution Imaging Spectroradiometer (MODIS) satellites due to their ability to measure vegetation indices.

The NASA EarthData also proved to be inconvenient for the scope of our project. Not only did it not have enough variables that were relevant to our RF for feature recognition, but performing formatting on the data to account for the distortions between datasets proved to be too time intensive. MODIS passes do not always pass over the same spot on the globe each time, causing the datasets themselves to always be distorted by the Earth's curvature.

Upon further investigation, we came open source data provided by Oak Ridge National Laboratory's Distributed Active Archive Center (ORNL DAAC) which met the needs of our application.

3.3 ORNL DAAC Dataset

The ORNL DAAC dataset includes a single before and after flyover collection of MODIS, LiDAR, and Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) measurements and contains more than 200 relevant numerical measurements. Spectrometer bands included measurements like moisture levels, the presence of stress pigments in the leaves, and the degree of burn the ground itself suffered. The dataset also provided uniform geopositional resolution between the different measurement types, saving us the effort in having to perform positional corrections.

We chose to use AVIRIS Level 3 measurements taken on November 11, 2013, approximately three weeks after the last day of the fire. AVIRIS Level 3 data was chosen because it includes NBR measurement and most of the other band measurements were ratios and indices that are normalized across AVIRIS Level 1 and Level 2 measurements. So although our dataset did not include cumulative time series data, the examples would be normalized in some way. The AVIRIS data set was downloaded in the GeoTIFF file format, where each file included pixel values for a specific spectrometer band. Each pixel on the image corresponded to the AVIRIS resolution of 14.8 square meters, and the images had 2926x2926 number of pixels. Table 1 shows the 13 bands chosen for the feature set, as well as NBR, and their abbreviations.

Table 1: AVIRIS Bands Used for Feature Set

AVIRIS L3 Product	**Band Name**
Anthocyanin Reflectance Index 1	ARI1
Anthocyanin Reflectance Index 2	ARI2
Carotenoid Reflectance Index 1	CRI1
Carotenoid Reflectance Index 2	CRI2
Char Soil Index	CSI
Enhanced Vegetation Index	EVI
Mod. Chlorophyll Absorption Ratio Index	mCARI
Mod. Soil Adjusted Vegetation Index	mSAVI
Moisture Stress Index	MSI
Normalized Burn Ratio	NBR
Normalized Difference Nitrogen Index	NDNI
Normalized Difference Vegetation Index	NDVI
Plant Senescence Reflectance Index	PSRI
Water Band Index	WBI

3.4 Preprocessing & Cleaning the Data

Upon downloading the data the first step was getting the data from the 14 GeoTIFF files consolidated into a format where we could perform analysis and cleaning. The open-source *Geospatial Data Abstraction Library* provides a Python binding with which we used to extract raw values from the GeoTIFF files. The pixel matrix was flattened into a 1D array for each file and then placed into columns of a Comma Separated Values (CSV) file labeled by the band name. This intermediate step was done so that we always had the raw values of our dataset easily accessible for troubleshooting purposes. The consolidated AVIRIS data was loaded into a dataframe using the Python library *Pandas*.

After consolidating the ORNL DAAC data into a format we could work with, we began analyzing the data to validate the measurements and check if normalization was necessary. Many of the data points were imported as null, NaN, or infinite values, so we first standardized all of the values to NaN to use consistent computations in Python. To determine if any of the features required normalization, the raw data for each feature was plotted as a heat map to represent the spatial image for that measurement.

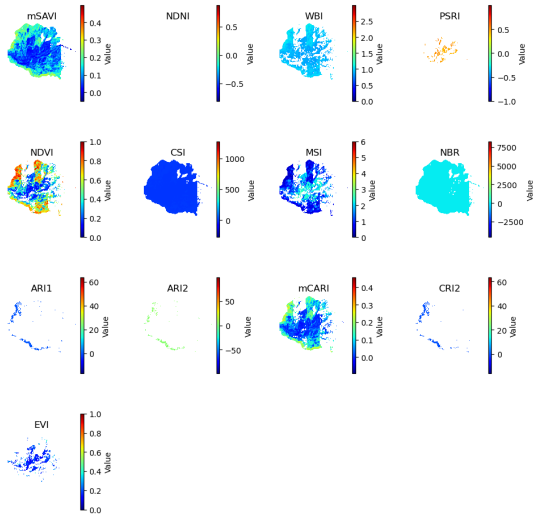

Figure 1: Spatial Heat Map of Raw Band Measurement

Because the NDNI measurements contained no valid data, this feature was dropped from the dataset. According to the ORNL DAAC documentation this can occasionally happen due to pollutants and aerosols in the air during data collection.

The minimum, maximum, mean and standard deviations were also analyzed. If a measurement had a wide range of values but low resolution on the heat map, it was normalized. Features that included negative and positive values were normalized using min max normalization to a range of [-1:1]. Features that only included positive values were min max normalized to a range of [0:1]. The bands "NBR" and "ARI2" were normalized to [-1:1] while the bands "ARI1", "CRI2" and "CSI" were normalized to [0:1].

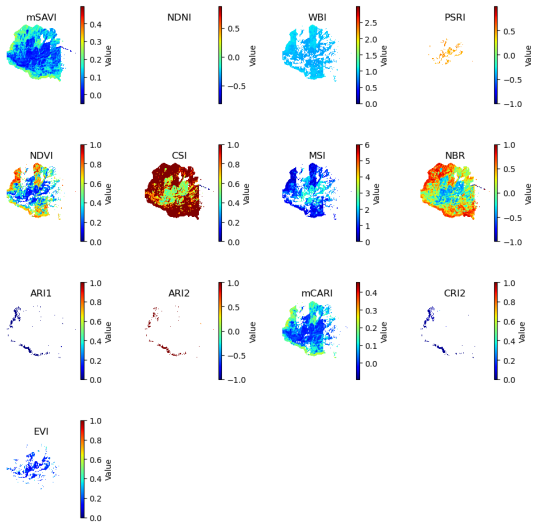

Figure 2: Heat Map of Features Normalized

The final step in preparing the data for the RF model would be performing imputation. The heat maps show that not every pixel in a measurement has a valid value. This was by design since we replaced invalid values with NaN. To account for this a simple imputation was performed on every feature, where the mean pixel value was calculated and then placed into the invalid pixels.

3.5 Random Forest Model Approach

With a dataset preprocessed and ready for training we began to train and fit the RF model. RF uses random data to generate multiple decision trees. The result of RF is an average from all decision trees. Algorithm 1 illustrates RF as defined in *Pro Machine Learning Algorithms*. The RF accounts for over-fitting and incorrect path selections and the implementation can be performed with Python's *scikit-learn* library (Ayyadevara, 2018). After building a RF model with 100 DTs, then the same model was trained with 150 DTs and 200 DTs to validate the results. Once adequate performance metrics were achieved, feature importance was analyzed to make assumptions on the feature set. Figure 3 shows the overall project architecture.

Algorithm 1: Random Forest

Let $n = 0$.
Let $output = 0$.
while $n <$ number of trees **do**
    Subset data to build a decision tree
    Subset features
    Build decision tree based on table with instances as rows and features as columns and
    Predict on test dataset
    $output+ = treeOutput$
    $n++$
end while
return $output/n$

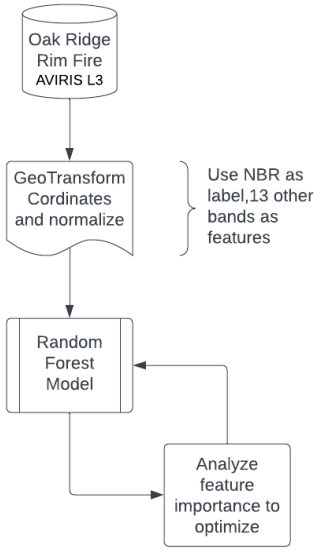

Figure 3: Project Architecture

4 Evaluating the Model

To evaluate the performance of the RF, Mean Square Error (MSE) and coefficient of determination (R-squared) were used. With a valid RF model, feature importance was analyzed by counting the frequency of features present in top level node's of the RF's individual DTs. Mean decrease of impurity was also used to inspect feature importance.

4.1 Random Forest Performance Metrics

MSE calculates the average squared difference between the predicted values and the actual values of the label. A lower MSE value indicates that the regressor is more accurate, with a MSE of 0 indicating a perfect fit. R-squared is a useful metric that computes the sum of the level of variance in the error terms divided by the the true sum of squares, subtracted from one. A R-squared value of 1 indicates a perfect fit. Python libraries provide a straight forward way to measure MSE and R-squared of a regression model.

4.2 Measuring Feature Importance

The first metric used to quantify feature importance was a manual calculation of the frequency of features present in the top level, or root nodes of each individual DT. If a feature is more pervasive in the root nodes, it may suggest that this feature is driving performance of the RF.

Another method used was MDI, which measures the average decrease of impurity throughout all of the nodes a DT. Since the goal of DT is to minimize impurity or heterogeneity in the subset generated by the node split, MDI is a useful way to analyze important features in the RF. The Python library *sklearn* provides a "RandomForestRegressor" class with a built in feature importance attribute and method for calculating MDI.

5 Results

To validate the performance of the RF three different models were built, keeping all hyperparameters constant except for the number of trees. The data was split into subsets of 80% for training and 20% for testing. There appeared to be no clear improvement in performance from increasing the number of DTs. Table 2 shows the MSE and R-square across each run. The average R-square value was 0.886 and the average MSE was 1.403. From these metrics, we can see the RF model used achieved relatively high performance in predicting NBR.

Table 2: RF Model Performance

# of DTs	MSE	R-squared
100	1.344	0.885
150	1.532	0.887
200	1.333	0.886

5.1 Feature Importance

The first metric taken to quantify feature importance was the frequency of top level nodes in the individual trees of the RF.

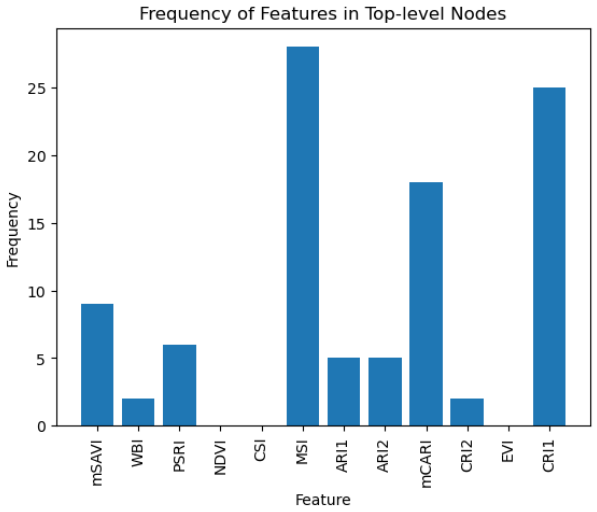

Figure 4: Frequency of Features in Top Level Nodes

From this testing, we see that the bands "MSI", "CRI1" and "mCARI" were the most frequent features in top level nodes. The high rates of these features may indicate that they drive the performance of the RF since the root node split has a large influence on a DT. The bands "NDVI", "CSI", and "EVI" are not present in any of the top level nodes, suggesting that these features may have a lesser impact on RF performance.

Mean Decrease of Impurity was the second metric used to analyze feature importance. The mean of this measurement was collected as well as the standard deviation across each feature. Figure 5 plots the mean as bars and the standard deviation as vertical lines.

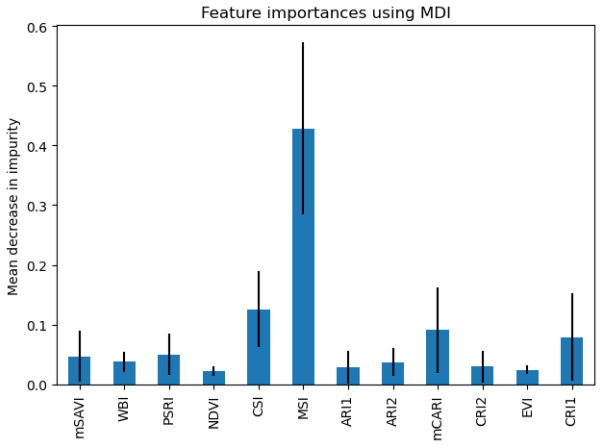

Figure 5: RF MDI

To observe the overall importance between the two metrics, we took an average of both to see which bands had the highest and lowest scores.

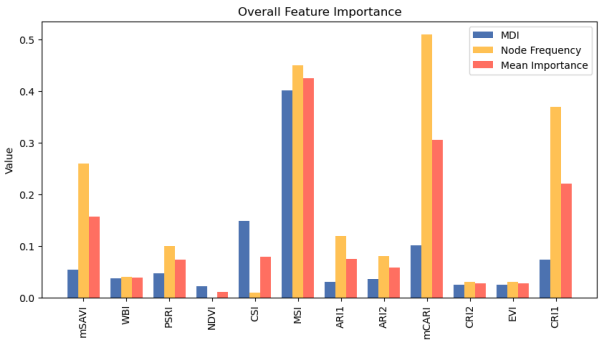

Figure 6: Overall Feature Importance

Some observations of note are the fact that NDVI had a frequency of 0 and the lowest overall importance of all the features for determining the model. NDVI is essentially a measure of an area's "greenness", a value determined by the ratio of green light reflecting from an area compared to the amount of red light being reflected by the same area, and is one of the most common metrics available from satellite imagery. From similar projects, NDVI is typically a significant indicator for vegetation health, but in our application, the measurement was likely skewed as insignificant since we took a single measurement post-fire, rather than a cumulative measurement before and after the fire. This is partially a limitation of the abilities of the AVIRIS data, and highlights the importance of consistent time series data collection to inform proper models.

The variable with the highest importance was Moisture Stress Index (MSI) which not only had the highest overall score, but also had the lowest spread between the two metrics amongst high scoring variables. MSI measures the amount of water being stored by the plants and was the greatest factor in predicting NBR, which suggests it is a key

variable in vegetation regrowth. Interestingly, another band that measures water content – Water Band Index (WBI) – had very low importance scores.

Modified Chlorophyll Absorption Ratio Index (mCARI), and Carotenoid Reflectance Index 1 (CRI1) had high overall scores, but there was a significant deviation between the two metrics. Performing further analysis may help explain the variance of the two metrics in these bands. The Carotenoid Reflectance Index is a measure of a pigment that plant leaves express when they are undergoing stress, with the CRI1 value being more tuned to low-to-medium concentrations and the CRI2 more accurately measuring high carotenoid concentrations. This finding suggests that taking high resolution measurements of low concentrations of caratenoid may be more useful than high concentration measurements.

It is important to note that these relationships should not be considered conclusive in determining variable importance due to the nature of the dataset. Given a larger dataset with measurements normalized overtime, the feature importance could have different results.

6 Suggestions for Future Applications

While the framework used in this project is primitive, we believe it could be extended to provide valuable insight in real world applications by using large data sets that cover measurements taken over a continuous time period before and after a fire. To label this dataset we would recommend using a basic regression method such as Stochastic Gradient Descent (SGD) or Theil-Sens regression. Thiel-Sens regression is not sensitive to bias in the feature set and has been used on similar projects.(Yang et al., 2021).

$$\theta = median \frac{NDVI_i - NDVI_j}{i - j}, \forall i < j \qquad (1)$$

In a Thiel-Sens regression (equation 1), the i and j values are timestamps for the data being operated over in the set. So, when considering the most recent timestamp j, the Thiel-Sens θ considers all previous timestamps for which there is data, $\forall i$, and uses all of them to compute the value. By utilizing this technique on a time cumulative measurement like NBR or NDVI, a label could be used for the RF.

Different types of linear data and categorical data should also be considered for future projects. Linear variables like topographical and weather data could be useful. Canopy height measurements should also be considered to justify the values of vegetation indices. LiDAR is a remote sensing technique that can collect handle canopy height. Finally, field measurements can be interpolated with time series data to include features like plant species and soil content present in a spatial area.

7 Conclusion

In this paper, we proposed a ML method to quantify and better understand the driving variables contributing to post wildfire vegetation regrowth. 14 AVIRIS bands from the ORNL DAAC 2013 Rim Fire data set were used. The data was preprocessed and cleaned and NBR was used as a predictor for a RF model. Feature frequency and Mean Decrease in Impurity were used to analyze feature importance of the RF model. This framework can be extend to other applications. But the work doesn't end here. These trends may vary over time, with some variables being more important to short term regrowth, and others more important to long term regrowth. Learning these trends may require more variables collected across a longer time frame. There may be even more types of variables to be considered, like wildlife populations and temperature ranges, that this model could further be developed to incorporate. But, so far, our model has worked, and achieved its goals, even if there exist more peaks to surmount. We input data, and from it, patterns emerged. We turned assumptions into demonstrable proofs, and figures into trends. There is more work to be done, but our model is a step along that path, a rung on the ladder.

References

Ayyadevara, V. 2018. Random Forest. In *Pro Machine Learning Algorithms*, 105–116. Berkeley, CA.: Apress.

Stavros, E.; Tane, Z.; Kane, V.; Veraverbeke, S.; McGaughey, R.; Lutz, J.; Ramirez, C.; and Schimel, D. 2016. Remote Sensing Data Before and After California Rim and King Forest Fires, 2010-2015.

Barsi, J. A.; Lee, K.; Kvaran, G.; Markham, B. L.; and Pedelty, J. A. 2014. The Spectral Response of the Landsat-8 Operational Land Imager. *Remote Sensing*, 6(10): 10232–10251.

Bhat, Nagaraj. 2020. Australian Bush fire satellite data (NASA), Version 1. https://www.kaggle.com/datasets/nagarajbhat/australian-bush-fire-satellite-data-nasa. Accessed: 2023-02-21.

Debouk, H.; et al. 2013. Assessing post-fire regeneration in a Mediterranean mixed forest using LiDAR data and artificial neural networks. *Photogrammetric Engineering and Remote Sensing*, 79(12): 1121–1130.

João, T.; et al. 2018. Indicator-based assessment of post-fire recovery dynamics using satellite NDVI time-series. *Ecological Indicators*, 89: 199–212.

Piyush, J.; et al. 2020. A review of machine learning applications in wildfire science and management. *Environmental Reviews*, 28(4): 478–505.

Stavros, E. N.; Coen, J.; Peterson, B.; Singh, H.; Kennedy, K.; Ramirez, C.; and Schimel, D. 2018. Use of imaging spectroscopy and LIDAR to characterize fuels for fire behavior prediction. *Remote Sensing Applications: Society and Environment*, 11: 41–50.

USDA. 2013. Rim Fire Information. https://www.fs.usda.gov/detail/stanislaus/home/?cid=stelprdb5433438. Accessed: 2023-03-31.

USGS. n.d. MOD13A3 v006. https://lpdaac.usgs.gov/products/mod13a3v006/. Accessed: 2023-04-02.

USGS. 2000. Earth Explorer. https://earthexplorer.usgs.gov/. Accessed: 2023-02-21.

USGS. 2017. What are the band designations for the Landsat satellites. https://web.archive.org/web/20170122043515/https://landsat.usgs.gov/what-are-band-designations-landsat-satellites. Accessed: 2023-02-22.

Yang, C.; Fu, M.; Feng, D.; Sun, Y.; and Zhai, G. 2021. Spatiotemporal Changes in Vegetation Cover and Its Influencing Factors in the Loess Plateau of China Based on the Geographically Weighted Regression Model. *Forests*, 12: 673.

Ye, Andre. 2020. NASA Wildfire Satellite Data, Version 1. https://www.kaggle.com/datasets/washingtongold/wildfire-satellite-data. Accessed: 2023-02-21