Predicting Wildfire Risk Based On Land Cover Classification and Past Wildfire Data in California

Akshada Guruvayur[1], Atharva Kulkarni[2], Cristina Marculescu[3], Ananya Chakravarthi[4], Andrew Liu[5]

[1] *Edison Academy Magnet School, 100 Technology Drive, Edison, NJ 08837*

[2] *Chantilly High School, 4201 Stringfellow Rd, Chantilly, VA 20151*

[3]*Westlake High School, 4100 Westbank Dr, Austin, TX 78746*

[4]*Plano East Senior High School, 3000 Los Rios Boulevard Plano, TX 75074*

NASA STEM Enhancement in Earth Science, Earth System Explorers, 2024

Wildfires have the capacity to destroy forests, homes, and the health of ecosystems. As a result of rising global temperatures, wildfires have experienced increased frequency and severity, leading to devastating outcomes throughout sections of North America. All of these factors have led to an increase in demand for tools that can accurately and efficiently identify potential risk areas, particularly after the devastating California wildfires of the 2010s and 2020s. This project aims to analyze pictures taken by NASA's GLOBE Observer app to identify land cover types in order to classify their potential contribution to wildfire risk in any given region. Additionally, past wildfires in California from 2000-2018 were also considered as another factor for future wildfire risk. A python model was then created using the FEMA Wildfire Risk Database and the GLOBE Observer database. The two features of land cover type and historical wildfires are paired with the FEMA wildfire risk database to determine the wildfire risk. The random forest algorithm helped ensure that each decision tree in the algorithm contributes to the final prediction, with the most frequent risk level chosen as the output. This approach ensures robustness and accuracy by combining the insights from multiple trees. The model uses these features to make predictions about wildfire risk, assigning a high, moderate, or low risk level based on the patterns it learned during training. In the end, the results showed that forests were least susceptible to wildfire spread while Urban areas presented a significant threat to wildfire risk. Scientists can use this algorithm, which provides real-time, ad-hoc data to analyze the risk of wildfires at times when satellites may not accurately capture land cover imagery.

Keywords: wildfire risk, remote sensing, NASA, GLOBE, land cover types

Research Questions

This work aims to answer the following research question: **What key structural components of land cover most significantly influence fire danger, and how can they be efficiently identified using the GLOBE Observer App and historical fires?** The GLOBE Observer App includes pictures taken directly from eye level and, as such, can provide valuable insight that satellites cannot capture, while historical data provides concise data to help narrow our research's temporal and geographic scope.

As global temperatures continue to rise due to climate change, the conditions conducive to wildfires—such as higher temperatures, prolonged droughts, and dry vegetation—are becoming more prevalent. This combination of data will help with understanding the structural components of land cover that influence fire danger, as they are essential for *predicting* and mitigating these increasingly common and destructive events.

1. Introduction

Once considered seasonal anomalies, wildfires in the United States are now raging year-round, devastating landscapes and communities as climate change and human activity intensify their frequency and ferocity (Burke et al., 2021). An analysis by the National Park Service has shown that throughout the nation, with California in particular, more than 80% of wildfires are started by people. Campfires burning for longer than recommended are heading this trend; when campers fail to fully extinguish a fire, leaving smoldering embers behind, those embers can quickly reignite and spread to surrounding vegetation.

Factors such as wind can carry sparks from these embers to dry grass or leaves, quickly igniting a larger fire. In drought conditions or during the dry season, even the slightest spark can lead to a major blaze. Even controlled burns can quickly spiral out of control if not conducted properly. Farmers and landowners often use fire to clear fields, dispose of agricultural waste, or manage underbrush. However, these fires can easily spread beyond their intended boundaries, especially under windy conditions (Balch et al., 2017).

Without human intervention, there are a few conditions that can result in a wildfire. Fuel, oxygen, and a heat source, paired with dry weather, drought, strong winds, or even lightning, can change a singular spark to a raging fire that consumes thousands of acres of land in its vicinity. The fire itself is the result of a combustion reaction, wherein some type of fuel is heated to the lowest temperature, and it needs to ignite and mix with the oxygen found in the air (McLauchlan et al., 2020).

The fuel needed consists of any type of flammable material, including shrubs, trees, grass, and even man-made structures. When exposed to heat, these materials undergo a process called pyrolysis, where the complex molecules within the material break down into smaller, volatile

compounds. This decomposition releases vapors, and when ignited, the heat generated by the combustion process continues to break down more of the material, releasing more vapors and sustaining the fire. This continuous cycle keeps the fire burning as long as there is enough fuel and oxygen.

Along with its potential to decimate everything around it, a wildfire's characterizing trait is its ability to spread incredibly fast. However, the intensity and movement of a wildfire depend on the weather, fuel, and topography of the area (Moore, 2021). A fuel's characteristics, including its moisture content, chemical properties, and density, play a crucial role in determining how rapidly a fire spreads and at what temperature it burns. Therefore, when vegetation is dry with low moisture content, it ignites and burns more swiftly because there is less water for the fire to evaporate, which allows the heat to combust the material directly, leading to more intense wildfires.

A fire will only ignite when the fuel's moisture level decreases significantly as the heat can then evaporate the remaining water, leading to the fuel becoming susceptible to burning. This shows the importance of considering moisture content in different types of vegetation, as they can help predict fire behavior more accurately.

The size and type of fuel impact wildfire behavior significantly. Smaller fuels, like grasses, ignite quickly but burn with less heat compared to larger fuels, such as trees, which sustain longer and hotter fires. Lush vegetation accelerates wildfire spread, while plants with high oil content, like eucalyptus and pine, are more flammable. Weather conditions, including wind, temperature, and humidity, also play critical roles; wind fuels the fire's spread, high temperatures and low humidity dry out fuels, and afternoon conditions are particularly conducive to rapid fire growth (Castro Rego et al., 2021). Additionally, topographical features, such as slope, elevation, and aspect, influence fire dynamics. Fires on steep slopes climb quickly due to rising heat, and south-facing slopes or lower elevations, which receive more sunlight, tend to have drier, more flammable fuels. Understanding these factors is essential for predicting fire behavior and developing key management strategies.

Recent research done by students at USC has found a new model that can accurately predict the spread of a wildfire using artificial intelligence USC scientists claim that this model can use satellite imagery to track a wildfire's progression in real-time and then feed this data into another algorithm that predicts the wildfire's intensity, growth rate, and the predicted path it will take.

They utilized historical wildfire data from satellite imagery and were able to "track how each fire was started, spread, and eventually contained by studying the behavior of past wildfires. Their analysis revealed patterns influenced by different factors like weather, fuel (for example, trees, brush, etc.) and terrain," (Shaddy et al.,

2024). The model then was taught to recognize patterns in the satellite images that helped it predict how wildfires spread in nature. The purpose of this study was to create a fully functioning model that anticipates how future fires might spread by observing how previous fires reacted.

Understanding the various ways wildfires start and spread is needed to identify the key factors that must be considered when collecting data to create predictive models that can estimate the likelihood of a wildfire occurring in a specific area. To enhance this understanding, the GLOBE Observer Database was used as a remote sensing mechanism.

The app is an online tool available to the general public that allows citizens and scientists of all ages to make observations about their environment through pictures taken through their devices, all of which contribute to the open GLOBE database (Figure 1) which tracks how the environmental has changed in an area to enhance any gaps in satellite data and offer more accurate scientific information that can be used in research ("Global Learning," 2023).

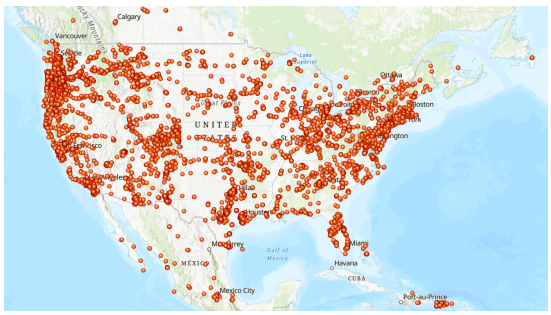

FIG 1: GLOBE Observers Land Cover Map

Citizen science significantly aids in this process by providing real-time data from diverse conditions, such as rain, snow, or drought, which enhances land maps and predictive models. By capturing images and reporting observations, citizen scientists fill gaps in satellite data, making environmental information more current and accessible. This collective effort not only improves our understanding of land cover and wildfire dynamics but also supports more timely responses to environmental challenges.

However, while research is able to focus on how existing fires are able to gain more ammunition to spread after already being started, little research has been done to predict *where* the fires can start in the first place. In this paper, we present a random forest analysis of classification and algorithms that find the correlation between land cover type and the corresponding risk of wildfire in California. This approach aims to fill a critical gap in predictive wildfire management, providing valuable insights that can lead to more essential prevention strategies. Understanding these correlations not only aids in early detection and response but also supports long-term planning and resource allocation to mitigate wildfire risks.

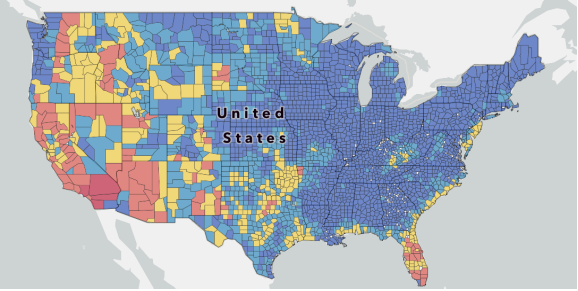

FIG 2: FEMA Wildfire Map

2. Methodology

California was the area of interest, not only due to its large land mass but also due to the variety of ranges of wildfire risk in the state, as denoted by the FEMA Wildfire Risk Map (Figure 2), which details the risk of wildfire in a certain location. California had locations ranging from very low to very high risk of wildfire, which provided us with enough variability to conduct our research.

Another distinct factor of California is its Mediterranean climate, with dry and warm summers and mild winters. The climate throughout the state can vary significantly due to the presence of the Sierra Nevada Mountains and Pacific Ocean coastline. Average temperatures remain within 70 degrees Fahrenheit but can go up to 80 degrees (Brown P.T, *et al*). Much of inland California is arid and filled with brush, desert, and dusty areas, as it is encompassed by the Mojave desert. This is a strong contrast to the fertile, green Central Valley, where much of the land is covered in greenery, both cultivated and wild.

To obtain the dataset needed for our research, we consulted the GLOBE Observer Database, which contains thousands of data points collected nationwide. This database is downloadable as a CSV file, but we specifically required data points located in California. To achieve this, we had to implement a filtering process to isolate the relevant data. Given the vast number of entries, manual filtering was impractical, so we turned to Python for an automated solution.

We developed a Python script that accessed the CSV file and filtered out any data points outside a defined range of latitude and longitude values specific to California. This involved setting precise margins for acceptable coordinates to ensure accuracy. The script then removed any data points falling outside these boundaries, reducing the margin of error and ensuring relevance. Finally, the script generated a new CSV file containing only the GLOBE Observer coordinates within California. This refined dataset ensured that our research focused exclusively on relevant geographical data.

We also meticulously scanned the dataset to ensure that no data from other countries made its way into our dataset. Additionally, we ignored any points that did not provide complete information on the land cover in the area, such as those lacking images for the north, south, east, and west directions. This was crucial as we needed to ensure the availability of detailed land cover information to accurately identify the type of land cover.

Any entries with missing or incomplete data were discarded to maintain the integrity of our dataset. Finally, the script generated a new CSV file containing only the GLOBE Observer coordinates within California. This refined dataset was crucial for ensuring that our research focused exclusively on relevant geographical data, providing a solid foundation for our analysis and model training.

2.1 Remote Sensing Data:

The remote sensing data used in this project included measurements of the Wildfire Risk Index, land cover type, as well as size and shape of historical wildfires. The Wildfire Risk Index was obtained from the Federal Emergency Management Agency's National Risk Index, which aimed to provide a view of natural hazard risk across the United States, addressing the diverse likelihood and consequences of natural hazards and the social factors influencing community risk. Initiated under FEMA's Natural Hazards Risk Assessment Program (NHRAP), the Index combines hazard likelihood, impact, and community vulnerability to deliver a holistic assessment. This tool leverages available data to develop baseline risk measurements for each county and Census tract, with an interactive map and data interface allowing users to investigate community risks.

The size and shape of a wildfire that occurred between the years of 2000 and 2018 were contributed by the National Interagency Fire Center, which fed its database into the ArcGIS software with a list of points where previous fires occurred as well as details on the location, size, and shape of the wildfire, the year in which it happened, the date and time when the fire perimeter was last mapped or updated, and the unit responsible for the point of origin.

2.2 Data Preparation:

To utilize the majority of our data, we downloaded each dataset in the form of Excel CSV files, which were subsequently fed into our data model. In the end, three datasets were created and then combined into a master file. This master file was then filtered into several different categories, all the while maintaining data quality and ensuring that no useful data points were disregarded.

The dataset that included points from the GLOBE Observer platform provided us with the latitude, longitude, elevation, and source codes for the images linked to each observation a user had taken at that site. However, as this set had more than 10,000 points from all over the world, we needed to ensure that we were considering just the California region. To do so, we imported the file into Jupyter Notebook, where we dropped all of the columns that were not being used, such as the elevation values, identification numbers, as well as the date when the observation was taken (as the data set included only values taken in the past couple months). In the end, the only columns that were left were in the format of

latitude, longitude, country name (just to ensure that each point lies in the United States), as well as four columns labeled

landcoversNorthPhotoUrl
landcoversSouthPhotoUrl
landcoversEastPhotoUrl
landcoversWestPhotoUrl

which contain the source code for each picture taken at the point. After the new CSV file was created, we conducted a statistical analysis of land cover.

2.3 Statistical Analysis:

The filtered CSV file had over 1200 points located in the California region, and in order for us to come up with a correlation between the type of land cover and the wildfire risk in the area, we needed to first identify the majority wildfire risk for a certain set of these points. Using the FEMA Wildfire risk map, we were able to plug in each set of coordinates into the software and receive the wildfire risk index for each point. In the end, about 325 coordinates were labeled to have a very low risk for wildfire, 350 had a relatively moderate risk, and 578 had a very high risk for wildfire.

These were considered to be our “strata” or categories that we initially clustered the dataset into and were the main component of our stratified random sampling method. Stratified random sampling is a type of statistical sampling technique in which researchers are able to obtain a sample population that best represents the entire population being studied. After the entire population is divided into homogenous groups, random samples are then selected from each stratum in either proportion or disproportion to the population.

This method differs from simple random sampling, which involves the random selection of data from an entire population so each possible sample is equally likely to occur. The primary benefit of stratified random sampling is its ability to capture key population characteristics within the sample, similar to a weighted average. This method ensures the sample's attributes are proportional to the overall population, making it better for diverse populations with distinct subgroups. Stratified sampling offers greater precision and smaller estimation errors compared to simple random sampling, with increased precision as differences among strata grow.

In our case, the population we used was the entire dataset of 1253 points of observations located in California. These points were then divided into three strata (very high, very low, and relatively moderate categories), with each stratum being homogeneous as the points it contained all had the same type of wildfire risk. After this, we used a disproportionate stratified method, in which instead of choosing the number of samples based on the population in each category, we simply chose 33 random points from each stratum. Often, disproportionate stratification is used to give larger than proportionate sample sizes in one or more subgroups so that separate analyses by sub-groups are

possible. As we want to analyze each group for the type of land cover found in each category, we chose to use disproportionate stratification over proportionate stratification.

After we inputted the row number of the values in each stratum into a random number generator, we were able to generate 33 random numbers that would be used for further analysis. This is where we began our land cover analysis (Figure 3 below). Each row contained the source code for each image, so we had to manually visit each of the images to identify the land cover found at that location. The type of land cover was split into 5 categories: urban, forest, wetland, grassland, and shrubland.

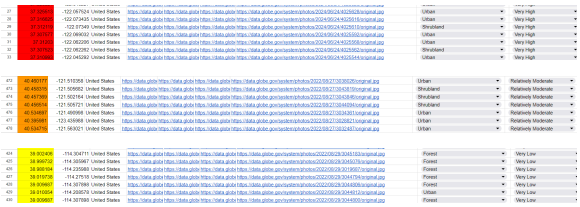

Figure 3: Spreadsheet illustrating strata organization: red indicates very high wildfire risk, orange represents moderate risk, and yellow denotes very low risk. Points from each section were analyzed for land cover type.

2.4 Random Forest Regressor:

Random Forest Regression is an ensemble learning method that constructs a multitude of decision trees during training and outputs the average prediction of the individual trees to make predictions. Each decision tree in the ensemble is built from a random subset of the training data, employing a technique known as bootstrapping, and at each split, a random subset of features is considered. Combined with averaging, this randomization helps reduce overfitting and variance, making the model more generalizable to unseen data. The method leverages the power of multiple trees to enhance predictive accuracy and control over-fitting, ensuring that the model captures the essential patterns in the data without being overly sensitive to noise.

The core principle behind Random Forest Regression lies in its ability to create an aggregated model that benefits from the strengths of multiple weak learners, in this case, decision trees. Each tree in a random forest operates as an independent estimator, and its predictions are aggregated to form the final prediction. This aggregation can be achieved by averaging the predictions (in regression tasks) or by majority voting (in classification tasks). The randomness introduced by bootstrapping and feature selection ensures that the trees are de-correlated, providing a diverse set of predictions that, when averaged, result in a more accurate and stable prediction. This ensemble approach inherently deals with the limitations of individual decision trees, such as high variance and overfitting, making Random Forest Regression a powerful and versatile tool for predictive modeling.

FIG 4: The machine learning model our team created looks at GLOBE Observer pictures as well as satellite imagery to determine relative probabilities of a fire at that position

The Random Forest Classifier in our ML model (Figure 4) was used to predict wildfire risk by analyzing land cover and historical fire data. Initially, categorical data was encoded into numerical values to facilitate the model's processing. The data was then divided into features (input variables) and the target variable (wildfire risk). This separation allowed us to train the model on 80% of the data while reserving 20% for testing its performance.

The Random Forest algorithm, consisting of multiple decision trees, was trained to classify wildfire risk based on the input features. During training, the model built numerous decision trees, each making predictions that were aggregated to improve accuracy. After training, feature importance was assessed to determine which variables were most influential in predicting wildfire risk. The trained model was then used to make predictions on new data, offering valuable insights into areas at risk of wildfire.

2.5 Evaluation Metrics:

One of the key performance metrics for Random Forest regression is the Mean Squared Error (MSE), which measures the average squared difference between the predicted and actual values. It quantifies the variance of the residuals and helps in assessing the accuracy of the model. Lower MSE values indicate better model performance.

$$MSE = \frac{\sum[(actual\ value - predicted\ value)^2]}{n} = \frac{\sum_{i=1}^{n}[y_i - \widehat{y_i}]^2}{n}$$

The Root Mean Squared Error (RMSE) is the square root of MSE, providing a measure of the average magnitude of the prediction errors in the same units as the target variable. These metrics are crucial for evaluating the effectiveness of the Random Forest regressor, and are an interpretable measure of model error since it is in the same units as the original data, making it easier to understand the magnitude of the error. By minimizing MSE and RMSE, the model ensures that the predictions are as close as possible to the actual values, enhancing its reliability and precision in real-world applications.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum[(actual\ value - predicted\ value}{n}}$$

$$= \sqrt{\frac{\sum_{i=1}^{n}[y_{,_i} - \widehat{y_i}]^2}{n}}$$

R-squared (r^2) is a statistical measure that indicates the proportion of the variance in the dependent variable that is predictable from the independent variables.

$$R^2 = \frac{variation\ explained\ by\ the\ model}{total\ variation\ in\ the\ data}$$

$$= 1 - \frac{\sum_{i=1}^{n}[y_{,_i} - \widehat{y_i}]^2}{\sum_{i=1}^{n}[y_{,_i} - \widehat{y}]^2}$$

It is calculated as shown above, where the $\overline{y}$ is the mean of the actual values. R^2 values range from 0 to 1, with an R^2 of 1 indicating that the model perfectly predicts the dependent variable, while an R^2 of 0 means that the model does not predict the dependent variable at all. R^2 is a key indicator of the model's explanatory power, showing how well the independent variables explain the variability of the dependent variable.

Mean Absolute Error (MAE) measures the average of the absolute errors between predicted and actual values, providing a straightforward measure of error magnitude. It is defined as

$$MAE = \frac{\sum_{i=1}^{n}|y_{,_i} - \widehat{y_i}|^2}{n}$$

Unlike MSE and RMSE, MAE does not square the errors, making it less sensitive to outliers. This makes MAE a good measure for understanding the average magnitude of prediction errors in a dataset without the disproportionately large impact of significant deviations.

In the context of a Random Forest classifier, the **accuracy_score** function from the **sklearn.metrics** module is commonly used to calculate accuracy. The Random Forest model is trained using the training data. During this process, multiple decision trees are built based on different subsets of the training data.

When making predictions, each decision tree in the forest predicts the class label for a given sample. The final prediction of the Random Forest is determined by majority voting: the class label that receives the most votes from all the trees is chosen as the final prediction. After making predictions on the test set, we compare these predicted labels to the actual labels (ground truth).

The accuracy_score function takes two arguments: the actual labels (y_test) and the predicted labels (y_pred). It counts the number of correct predictions by comparing these two arrays element-wise.

$$Accuracy = \frac{number\ of\ correct\ \ predictions}{total\ number\ of\ predictions}$$

The function then calculates the ratio of correct predictions to the total number of predictions, as described in the formula above.

Results:

The following section presents the results obtained from applying the Random Forest classifier to our dataset. This analysis includes an evaluation of the model's performance through various metrics, providing insights into its predictive accuracy and feature importance. By examining these results, we aim to understand the efficacy of the Random Forest model in capturing the underlying patterns within the data and its ability to generalize to new, unseen samples. Additionally, we will explore the contribution of different features to the model's predictions, shedding light on the key drivers influencing the classification outcomes.

Algorithm Results				
Model	*MSE*	*RMSE*	R^2	*MAE*
RF Regressor	0.40	0.630	0.290	0.300
	Accuracy			
RF Classifier	0.75			

TABLE 1. Results from random forest regression and classification.

The results from the random forest models in our project indicate substantial predictive capabilities. Table 1 above shows that the RF Regressor achieved a Mean Squared Error (MSE) of 0.40, a Root Mean Squared Error (RMSE) of 0.630, an R-squared (R^2) value of 0.290, and a Mean Absolute Error (MAE) of 0.300. These values reflect the model's ability to estimate fire risk with reasonable precision, though there is room for improvement. The RF Classifier, with an accuracy of 0.75, demonstrates a strong capability in correctly predicting fire risk categories based on land cover types. These metrics collectively suggest that our approach of using land cover classification significantly enhances the model's performance in fire risk prediction, providing a strong tool for future fire mitigation efforts.

3.1 Confidence Interval

To create a confidence interval for our Random Forest Classifier's accuracy, we will use the formula for the confidence interval of a proportion. First we need to ensure that all of the conditions for a confidence interval are satisfied:

1. The data used is from a random sample and is therefore representative of the population
2. Each observation in the sample data is independent of every other observation.
3. The size of the sample is greater than 30 (33 > 30)
4. The sample size should be less than or equal to 10% of the population size. (33 < 125.3
5. The sample size should be large enough so that both np and n(1-p) > 10

The Sample Proportion (Accuracy):

$$\hat{p} = 0.75$$

Z value for 95% confidence = 1.96

Sample size= 33

SE (Standard Error) =

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.75(1-0.75)}{33}} = 0.0754$$

ME (Margin of error) =

$$Z \times SE = 1.96 \times 0.0754 = 0.1478$$

Confidence interval:

$$\hat{p} \pm ME = 0.75 \pm 0.1478$$

So, the 95% confidence interval for the accuracy is:

(0.6022,0.8978)

The 95% confidence interval for our model's accuracy, ranging from 60.22% to 89.78%, indicates a broader range due to the smaller sample size of 33. While this interval is wider, it still provides useful information about the model's performance. The accuracy is likely to fall within this range 95% of the time, highlighting the need for more data to narrow the interval and improve precision. This shows the importance of citizen science in collecting additional data points to enhance the reliability of our wildfire risk predictions.

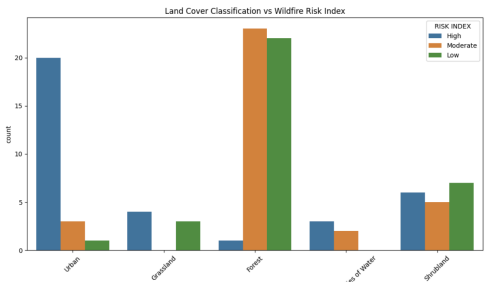

FIG 5: Bar graph that denotes the wildfire risk index per the type of land cover classification

Our model had a 75% accuracy rate, and through our analysis, we established a clear and direct relationship between different types of land cover and their associated wildfire risk. By leveraging the Random Forest classifier, we were able to discern the varying degrees of wildfire risk across distinct land cover types, confirming several established scientific observations. Specifically, our findings (shown by figure 5 above) indicate that forests generally exhibit low to relatively low wildfire risk, shrublands and grasslands demonstrate relatively high to very high wildfire risk, urban areas show varying levels of risk from moderately high to very high depending on location and population density, and wetlands present moderately high to relatively high wildfire risk.

The results suggest that forests generally exhibit a lower wildfire risk compared to other land cover types. This finding aligns with existing research indicating that the dense canopy and moisture retention in forest ecosystems contribute to their relative resistance to wildfire. However, this does not imply that forests are immune to wildfires; factors such as prolonged drought, accumulation of combustible material, and human activity can still elevate the risk.

Our analysis confirmed that shrublands and grasslands have a relatively high to very high wildfire risk. These areas are characterized by lighter, more flammable vegetation that can ignite and spread quickly. This observation is supported by studies showing that the fine, dry fuels present in these ecosystems are highly susceptible to ignition, especially during dry, windy conditions. Additionally, the open structure of these landscapes facilitates the rapid spread of fire (Radeloff, 2023).

The wildfire risk in urban areas varies significantly, showing moderately high to very high levels depending on factors such as location and population density. Urban areas located near wildland-urban interfaces (WUIs) are particularly at risk due to the proximity to

flammable vegetation and potential ignition sources from human activities. High population density can also increase the risk, as more potential ignition sources and vulnerabilities exist. Urban planning and fire mitigation play crucial roles in managing this risk.

Despite being less commonly associated with wildfires, wetlands also present a moderately high to relatively high wildfire risk. This might seem counterintuitive given their typically moist conditions, but during periods of drought, wetland areas can dry out and become susceptible to fire. Additionally, the organic-rich soils and dense vegetation in wetlands can provide ample fuel for wildfires once they ignite.

Discussion:

While the Random Forest classifier provided valuable insights, we faced several challenges with the data and methodology. One significant issue was data quality and completeness. Variations in data collection methods, missing values, and inconsistencies across different datasets required extensive preprocessing to ensure the integrity of our analysis. Moreover, the resolution of land cover data varied, potentially affecting the accuracy of our predictions for specific areas.

For some locations, classifying a land cover proved to be difficult due to the clarity of the images, which added another layer of complexity to our work. Every individual may interpret a picture of land cover differently; where one person sees 60% grassland, another might only see 40%. These variations in judgment can significantly impact conclusions about wildfire risk, potentially leading to the assignment of more risk to one land cover type over another, even when such differences may not be statistically significant.

The Random Forest mechanism itself also posed some challenges. Although it is capable of handling large datasets with multiple features, it can be computationally intensive and may require fine-tuning to optimize performance. Overfitting was a concern, particularly given the diverse nature of our land cover types and the varying scales of data. To mitigate this, we implemented techniques such as cross-validation and parameter tuning to enhance the model's generalizability and accuracy.

Park Williams, an associate research professor and a 2016 Center for Climate and Life Fellow, is conducting a comprehensive study on climatology, drought, and wildfires in the Western United States. His research involves compiling data on tens of thousands of fires that occurred over the last 30 to 40 years and cross-referencing this information with data sets on human population distribution and vegetation patterns.

The goal is to develop a computer program that can project how these variables influence the probability of large fires and simulate vegetation responses to fire (A. Park Williams et al. 2020). This tool aims to provide seasonal forecasts of wildfire

probabilities, which could be extremely useful for landowners in planning preventative measures such as prescribed burns or forest thinning. Additionally, it could enhance public awareness of wildfire risks and encourage proactive measures for yard maintenance and evacuation planning.

After reviewing Park Williams' research, several improvements could be made to enhance our own study on the correlation between land cover types and wildfire risk. Firstly, cross-referencing wildfire data with detailed vegetation and population distribution data could help refine the understanding of how these factors influence wildfire risk in different land cover types. Secondly, integrating advanced modeling techniques like those used in Williams' research, such as computer simulations that simulate vegetation responses, could significantly enhance the predictive capabilities of our study. Developing a tool that offers seasonal forecasts and assesses how different variables affect wildfire probabilities could provide valuable insights for policymakers.

Lastly, addressing the issue of data quality and resolution is crucial. Williams' research shows the importance of high-quality, consistent data sets. Improving the resolution of land cover data and ensuring consistency in data collection methods could mitigate the challenges of subjective interpretation and enhance the accuracy of wildfire risk assessments. By adopting this, our research could yield more reliable and actionable insights into wildfire risks associated with various land cover types.

Conclusion

Based on our model, we confirmed a correlation between land cover type and fire risk, with both methods showing similar accuracy. Using land cover classification significantly improved our model's accuracy, which was 75 %, enhancing future fire prediction. A 75% accuracy rate means that the model is reliably identifying and predicting fire risk in a significant majority of cases. Interestingly, historic fires in the area did not significantly impact the predictions, suggesting that other factors play a more crucial role in determining fire risk.

However, there is a lack of data points in specific areas of California, highlighting the importance of citizen science. Engaging the public in data collection can fill these gaps, improving the model's precision and aiding in better wildfire response strategies. Citizen science has the potential to play a huge role in preventing the loss of life, nature, and properties, provided that data is filtered according to a standard protocol (Dodson, 2022). To further bolster our model’s capabilities, we could integrate a Convolutional Neural Network (CNN) into our algorithm. CNNs are known for their ability to perform efficient data extraction with minimal human intervention, which could accelerate the analysis process while maintaining high accuracy. Implementing a CNN could significantly enhance our model’s performance, providing more rapid and reliable predictions that are essential for

timely wildfire response and prevention efforts.

The future applications of this research extend far beyond immediate wildfire risk assessment, offering a foundation for several impactful advancements. By refining predictive models based on land cover types, we can develop more sophisticated early warning systems that provide localized and timely alerts for potential wildfire outbreaks. This capability will allow for proactive measures, such as community evacuation and emergency response plans, to be implemented more practically. Additionally, the integration of citizen science into our data collection process opens avenues for real-time monitoring and updating of land cover information, further enhancing the model's accuracy. In the long term, this research could be instrumental in shaping policies for sustainable fire prevention.

Data and Code

Data and code to replicate the results of this experiment are available at the following public Github repo:

https://github.com/CoderManChild/NASA_SEES

GLOBE Observer data were obtained from NASA and the GLOBE Program and are freely available for use in research, publications, and commercial applications. GLOBE Observer data analyzed in this project are publicly available at globe.gov/globe-data (accessed on 5 July 2023). The Python code to read, analyze, and visualize GLOBE data for this article as well as the analyzed datasets are available on github.com/IGES-Geospatial.

Acknowledgements:

Working with project mentors was extremely impactful on our research, as they all provided their own insight into how the project could be approached. We acknowledge our mentors, Russanne Low, Peder Nelson, Cassie Soeffing, Andrew Clark, and Erika Podest, for their support and guidance in the research and publication process. We would also like to thank our peer mentors Benjamin Herschman and especially Andrew Liu, who greatly assisted in developing our project idea, utilizing online resources, and providing valuable feedback on our methodology. Their expertise and support were instrumental in refining our approach and enhancing the overall quality of our research.

The authors would like to acknowledge the support of the 2024 Earth System Explorers Team, NASA STEM Enhancement in the Earth Sciences (SEES) Virtual High School Internship program. The NASA Earth Science Education Collaborative leads Earth Explorers through an award to the Institute for Global Environmental Strategies, Arlington, VA (NASA Award NNX6AE28A). The SEES High School Summer Intern Program is led by the Texas Space Grant Consortium at the University of Texas at Austin (NASA Award NNX16AB89A).

References:

(1) A. Park Williams et al. ,Large contribution from anthropogenic warming to an emerging North American megadrought.Science368,314-318(2020).DOI:10.1126/science.aaz9600

(2) Balch, Jennifer K., et al. "Human-Started Wildfires Expand the Fire Niche across the United States." Proceedings of the National Academy of Sciences, vol. 114, no. 11, 27 Feb. 2017, pp. 2946–2951, www.pnas.org/content/114/11/2946.short, https://doi.org/10.1073/pnas.1617394114.

(3) Burke, Marshall, et al. "The Changing Risk and Burden of Wildfire in the United States." Proceedings of the National Academy of Sciences, vol. 118, no. 2, 12 Jan. 2021, www.pnas.org/content/118/2/e2011048118, https://doi.org/10.1073/pnas.2011048118.

(4) Brown, P.T., Hanley, H., Mahesh, A. et al. Climate warming increases extreme daily wildfire growth risk in California. Nature 621, 760–766 (2023). https://doi.org/10.1038/s41586-023-06444-3

(5) Dodson, J. Brant, et al. "Do Citizen Science Intense Observation Periods Increase Data Usability? A Deep Dive of the NASA GLOBE Clouds Data Set with Satellite Comparisons." Earth and Space Science, 6 June 2022, https://doi.org/10.1029/2021ea002058.

(6) Francisco Castro Rego, et al. "Fire Regimes, Landscape Dynamics, and Landscape Management." Springer Textbooks in Earth Sciences, Geography and Environment, 1 Jan. 2021, pp. 421–507, https://doi.org/10.1007/978-3-030-69815-7_12. Accessed 31 July 2024.

(7) Moore, Andrew. "Explainer: How Wildfires Start and Spread." College of Natural Resources News, 3 Dec. 2021, cnr.ncsu.edu/news/2021/12/explainer-how-wildfires-start-and-spread/.

(8) McLauchlan, Kendra K., et al. "Fire as a Fundamental Ecological Process: Research Advances and Frontiers." Journal of Ecology, vol. 108, no. 5, 8 June 2020, pp. 2047–2069, https://doi.org/10.1111/1365-2745.13403.

(9) Radeloff, Volker C, et al. "Rising Wildfire Risk to Houses in the United States, Especially in Grasslands and Shrublands." Science, vol. 382, no. 6671, 10 Nov. 2023, pp. 702–707, https://doi.org/10.1126/science.ade9223.

(10) Shaddy, Bryan, et al. "Generative Algorithms for Fusion of Physics-Based Wildfire Spread Models with Satellite Data for Initializing Wildfire Forecasts." Artificial Intelligence for the Earth Systems, vol. -1, no. aop, 23 Apr. 2024, journals.ametsoc.org/view/journals/aies/aop/AIES-D-23-0087.1/AIES-D-23-0087.1.xml, https://doi.org/10.1175/AIES-D-23-0087.1. Accessed 31 July 2024.

(11) Visit California. "How to Plan for the Weather in California | Visit California." Www.visitcalifornia.com, 23 Oct. 2014, www.visitcalifornia.com/experience/weather-timing-your-visit/.

International Virtual Science Symposium Badges

Data Science: This badge is being applied due to the large amount of data collected from the GLOBE Observer App from California. We used multiple data filtration techniques to process our data. Additionally, we integrated historic wildfire data from the National Interagency Fire Center (nifc) with our GLOBE Observer Data to our data points. This data was used with our Random Forest Algorithm to predict wildfire risk from our processed data.

Engineer: This badge is being applied for because we evaluated our research question through machine learning and hyperparameter tuning to improve the performance of our models. Our research addresses the real-world problem of wildfire risk and allows other researchers to understand the effects of the environment on how it may influence future wildfires.

Collaborator: This badge is being applied as a result of our team's collaboration skills. Atharva Kulkarni - from Chantilly High School - was the data scientist and created the ML Model. Akshada Guruvayur - from Edison Academy Magnet School - was the data gatherer/interpreter as well as the statistical analyst. Ananya Chakravarthi.from Plano East Senior High School, was our information organizer as well as Cristina Marculescu from Westlake High School. These roles helped clearly define everyone's part of the project based on our skills so that we could have a completed project on time.