Data-Driven Wildfire Risk Prediction with augmented spatial accuracy

Team 4

Megha Rajam Rao
Venkata Anil Kumar Thota
Qiao Liu
Nandini Puppala
Prathusha Koouri

SWiPS

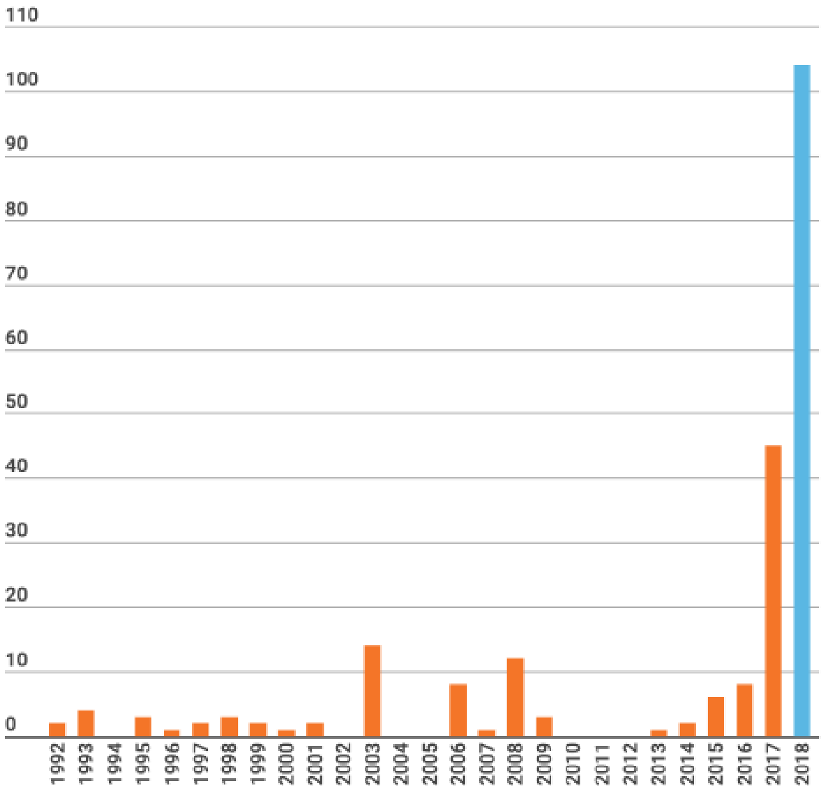
Guide:
Dr. Jerry Gao

Advisor:
Dr. Lee Chang

Outline

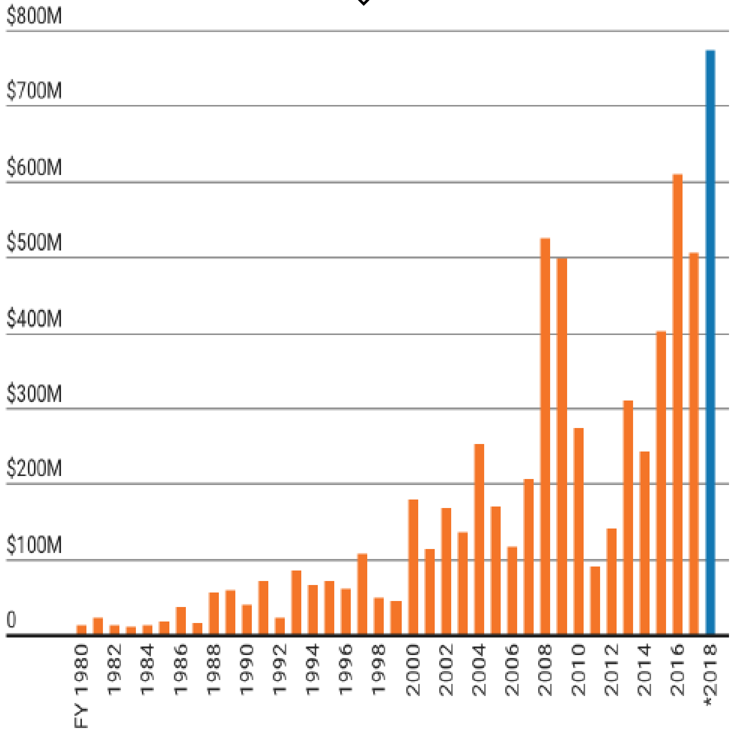
- Fire Statistics
- Project Overview
- Comparison of different systems
- Study Area
- Data-related Statistics
- Weather Data
- Powerline Data
- Terrain Data
- Fire History Data
- Vegetation Data
- Proposed Solutions
- Model experiments
- Model Architecture
- Threshold Tuning
- Evaluation Metrics
- Conclusion
- Lessons Learned
- Recommendations
- Demo

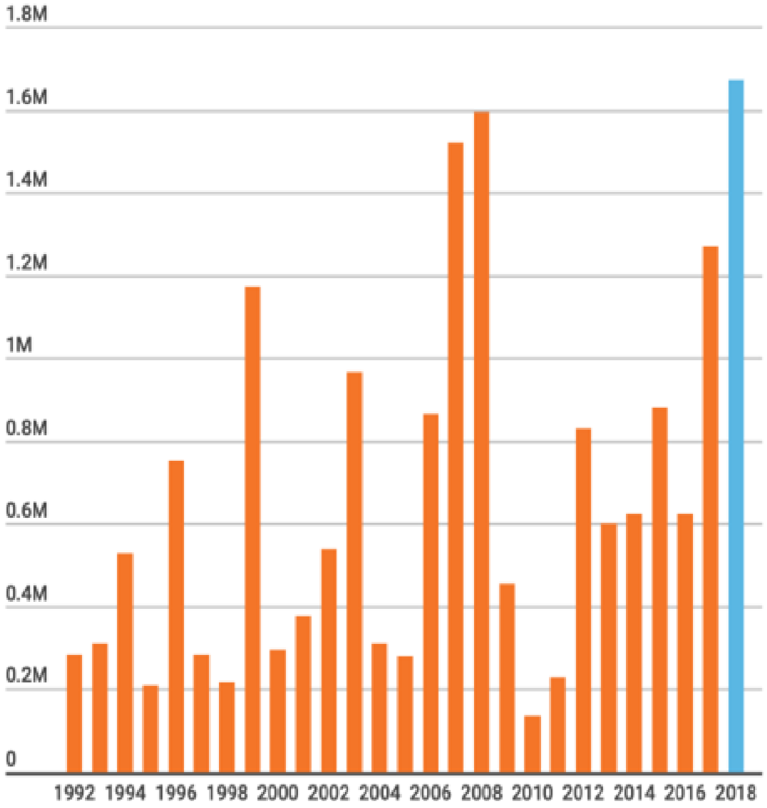
Fire Statistics



Yearly death toll in California

Cost of Wildfires from the year 1980 to 2018





Yearly Estimation of acreage burned in California

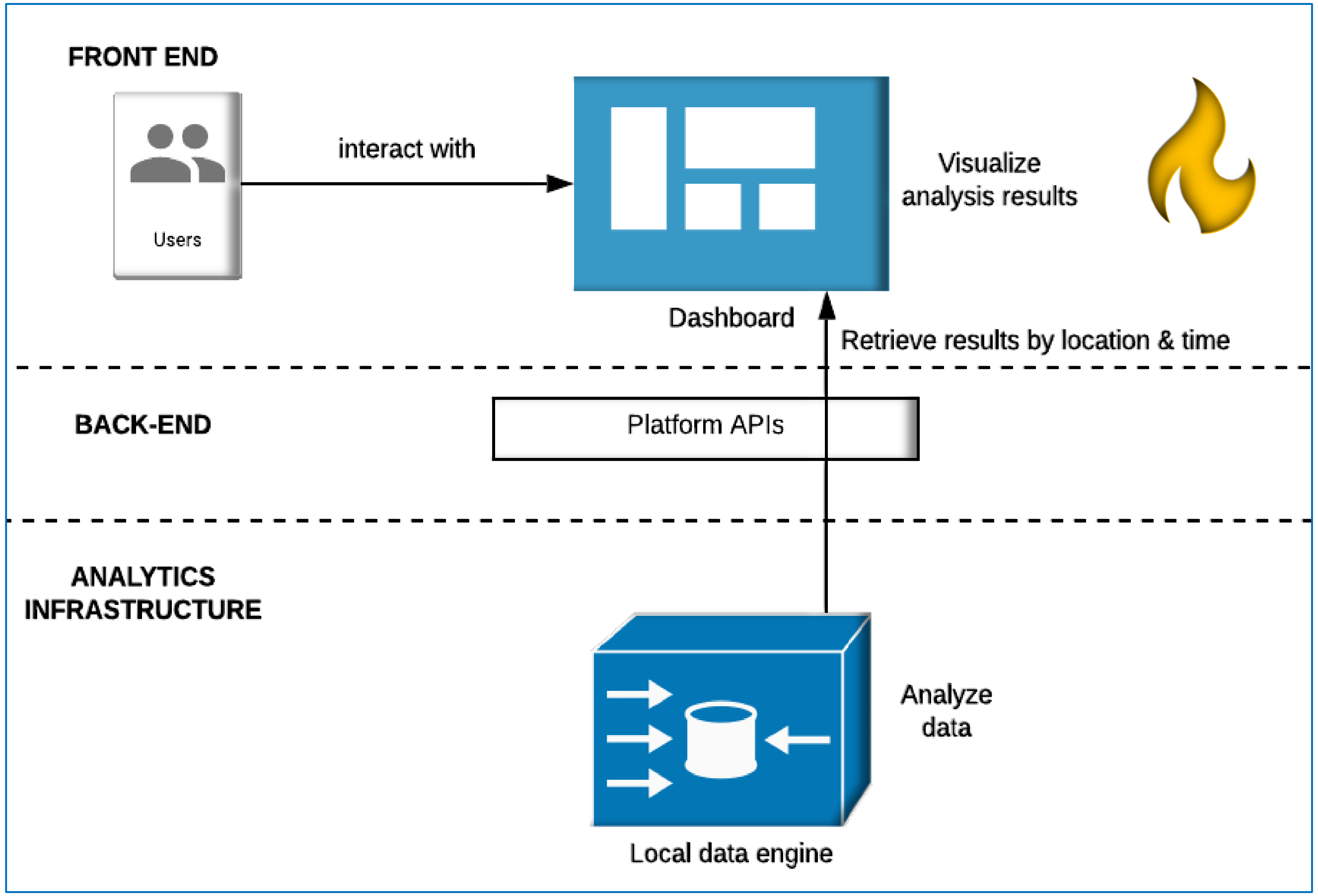
Fire statistics show the rapid increase of wildfire damage in California. Blue bar shows the largest numbers. The year 2018, deemed one of the worst years in California history, witnessed 7571 fires that burned across 1.6 million acres of land and claimed more than 100 lives.

Project Overview

- **Objective**: Build a model that improves accuracy for predicting wildfire risk in Northern California.

- **Purpose**: Create a wildfire risk prediction system that would help concerned agencies such as Cal Fire.

- **Outcome**: A ML model with augmented sensitivity that can predict local wildfire probability to a 92% accuracy, along with an interactive user Interface.



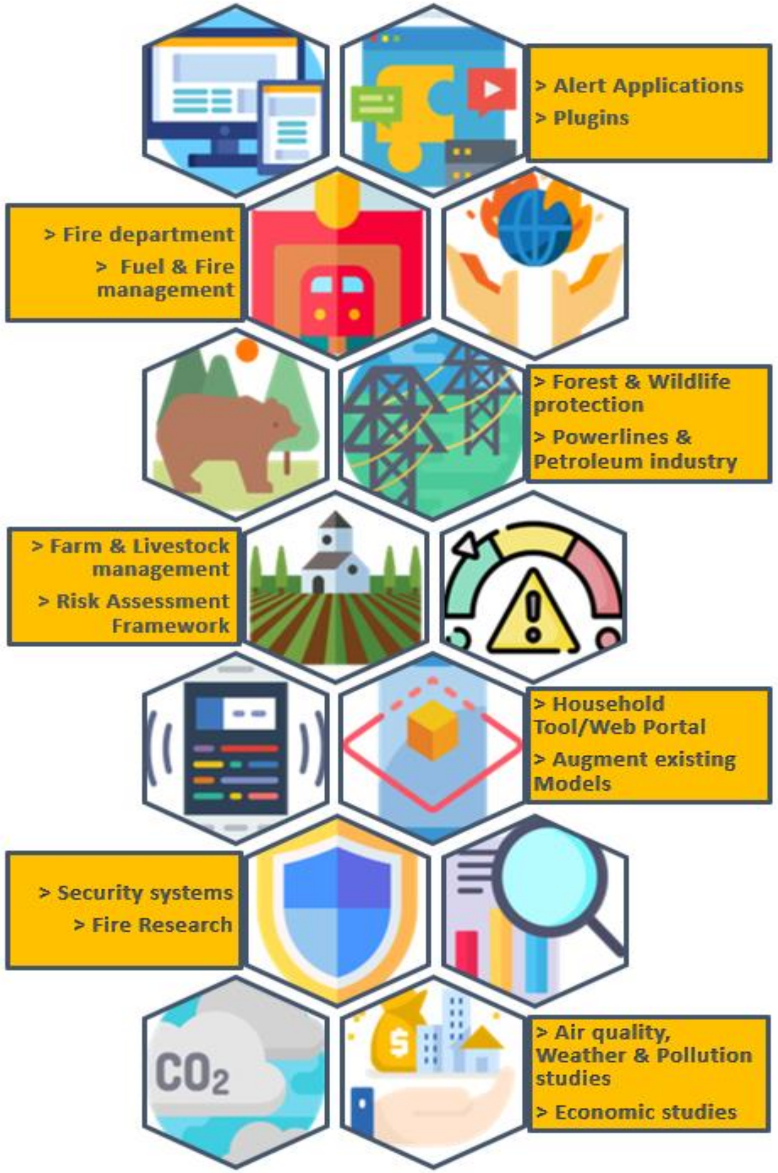
Need for this project

SWiPS

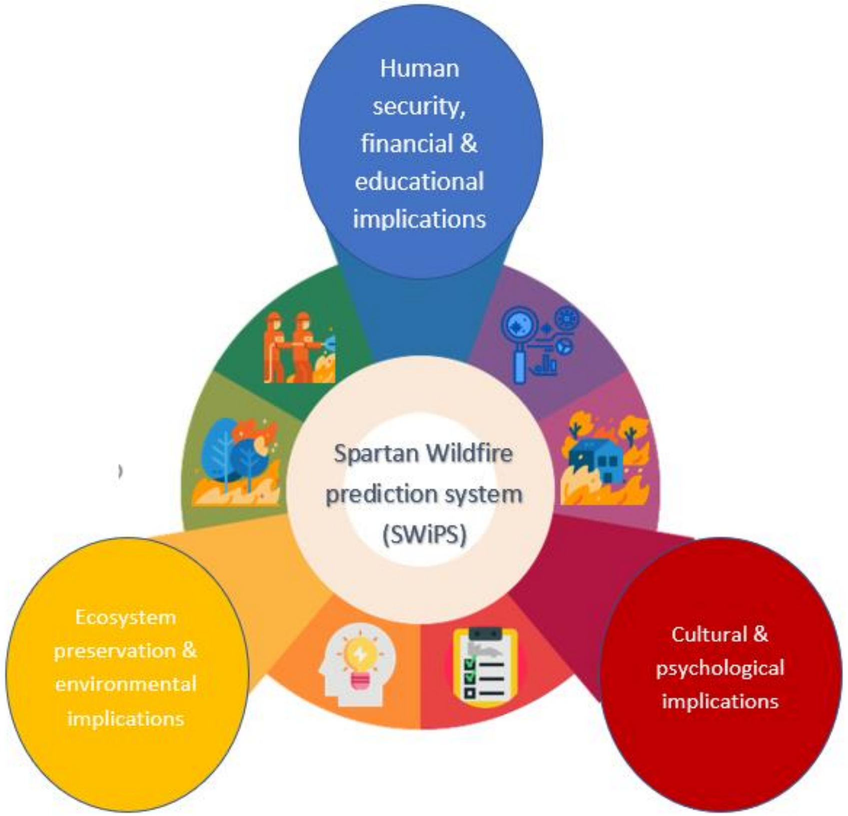
Applications

Wildfire Prediction System

Applications



Implications



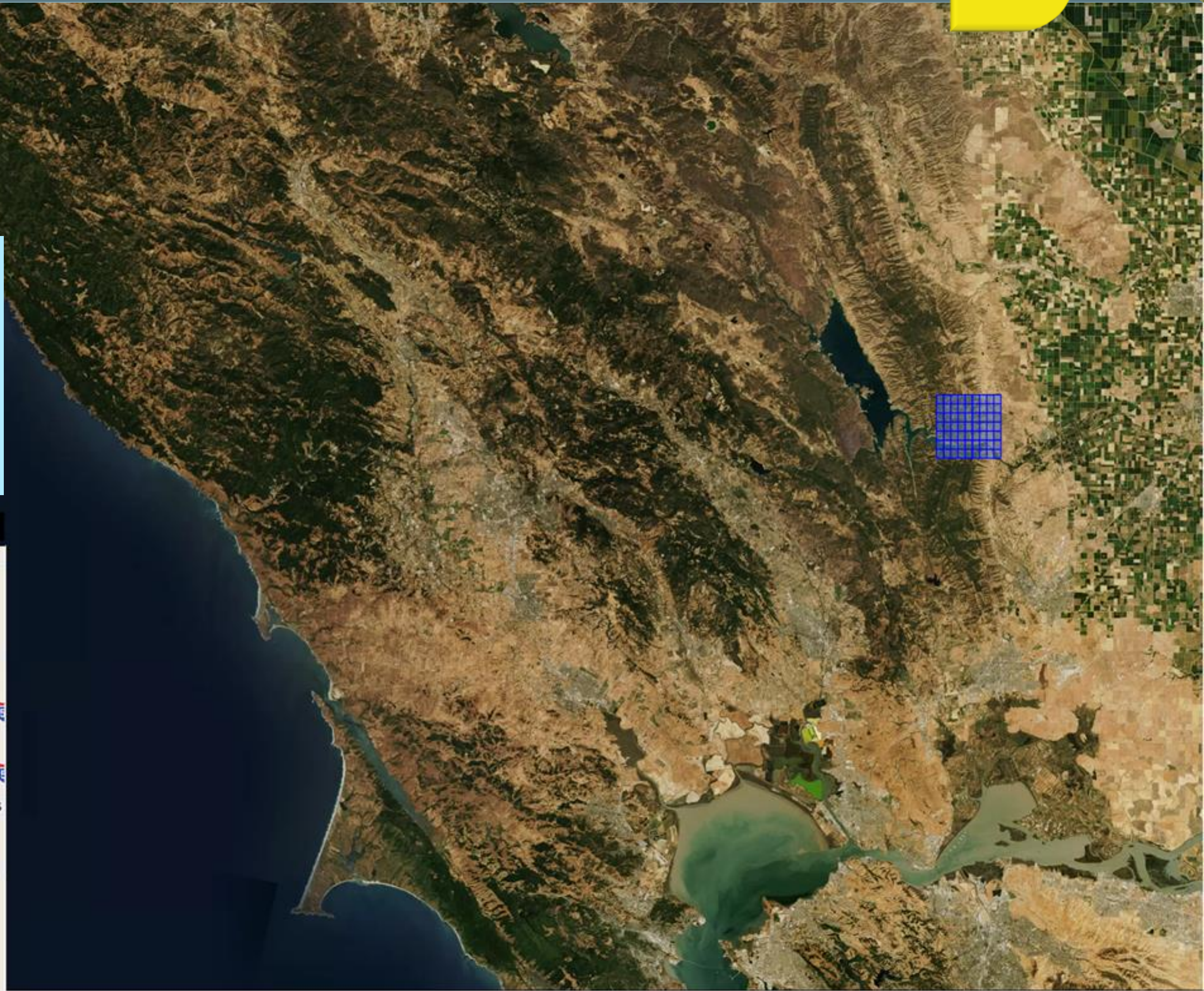
Comparison of different systems

System name	Country	Purpose				Data Acquisition				Approach			Methods	
		Prediction	Detection	Simulation	Management	Satellite	Sensor	Manual	Camera	Data-driven	ML	IR	Mathematical	ANN
MFFDI	Australia	√	×	×	×	×	√	×	×	√	×	×	√	×
NFDRS	US	√	×	×	×	√	√	√	×	√	×	×	√	×
CFFDRS	Canada	√	×	×	√	×	√	√	×	√	×	×	√	×
FFRFS	Japan	√	×	×	×	√	√	√	×	√	√	×	×	√
NCMSSD	Russia	×	×	×	×	√	×	×	×	√	×	×	√	×
SWiPS	(Ours)	√	×	×	×	√	√	×	×	√	√	×	√	√

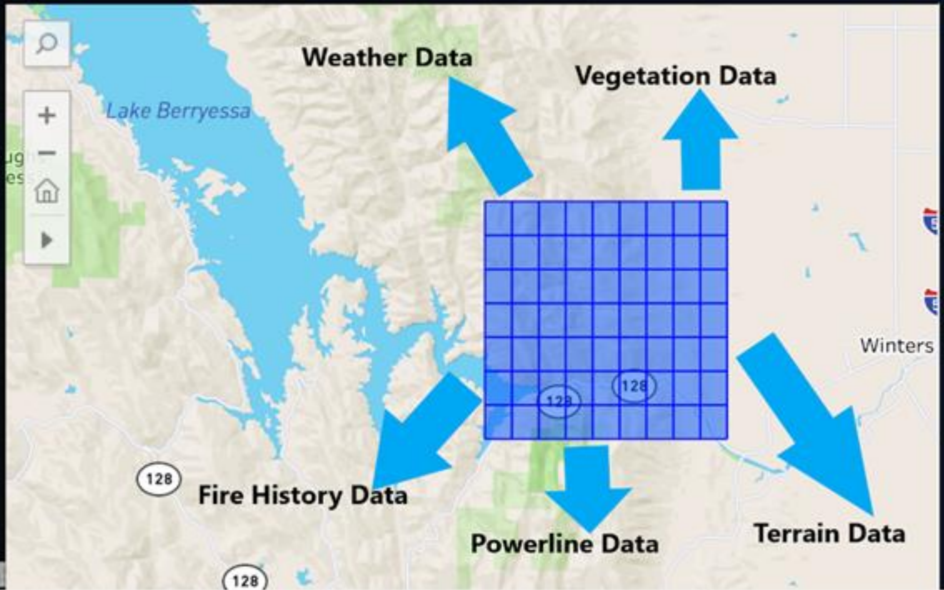
Study Area

Fire-prone Area divided to 63 grids.

- Dimension: 7 rows and 9 columns.
- Size: 1km*1km
- Area: 3969 km ^2

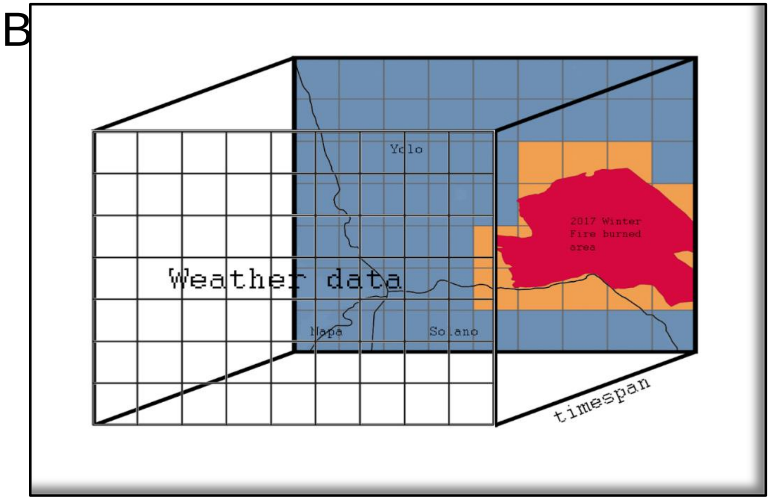


Street View of the Grids



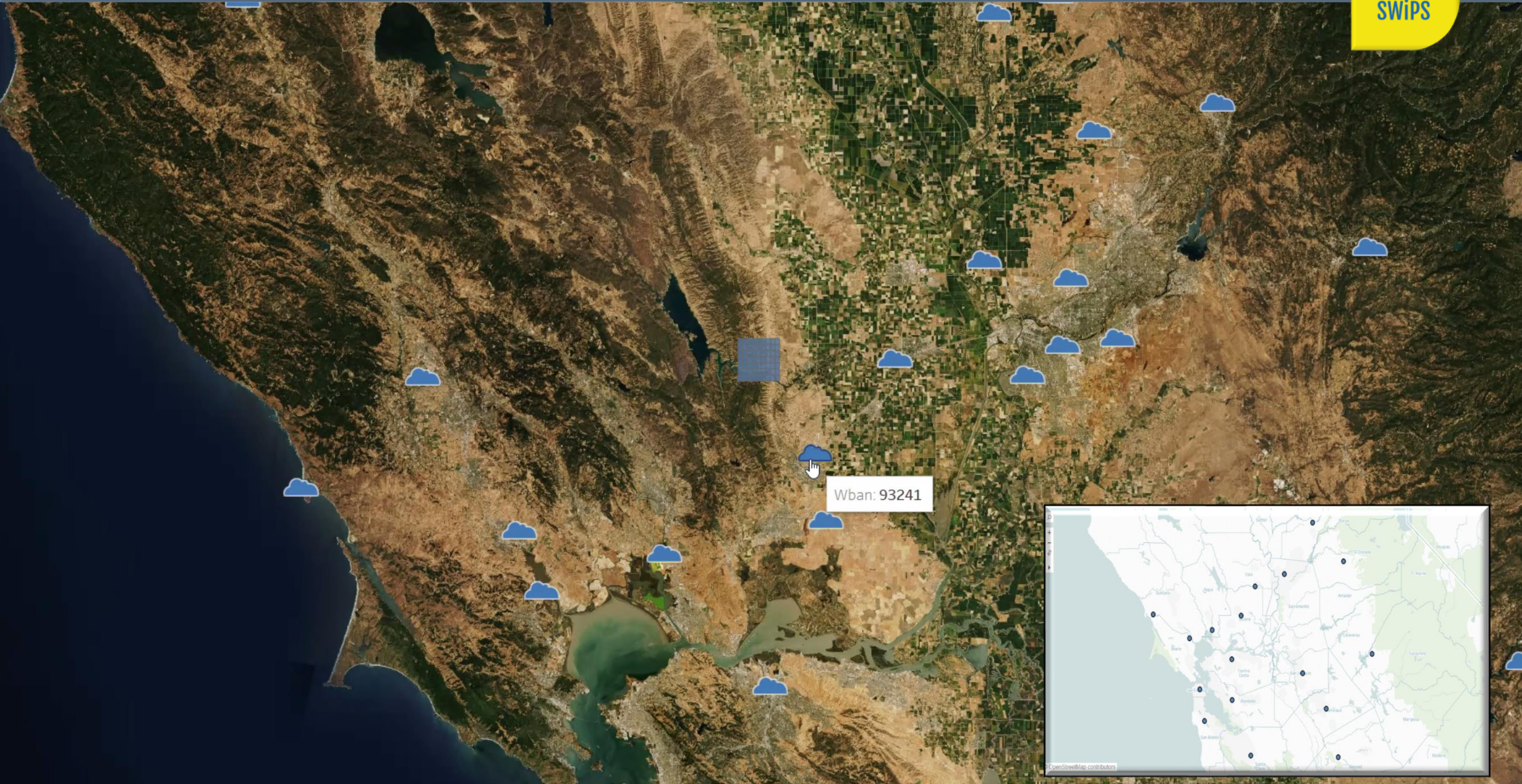
Weather Data

- Weather data from Local Climatology Data (LCD) is maintained by National Centers for Environmental Information (NCEI) which provides atmospheric and geospatial data across the United States.
- **Format** – csv file
- **Date range** - 01/01/2015 to 12/31/2019.
- **Frequency** - Hourly
- **Source** - [https://www.ncdc.noaa.gov/cdo-web/datatools/lcd.](https://www.ncdc.noaa.gov/cdo-web/datatools/lcd)
- **Size** – 168 MB

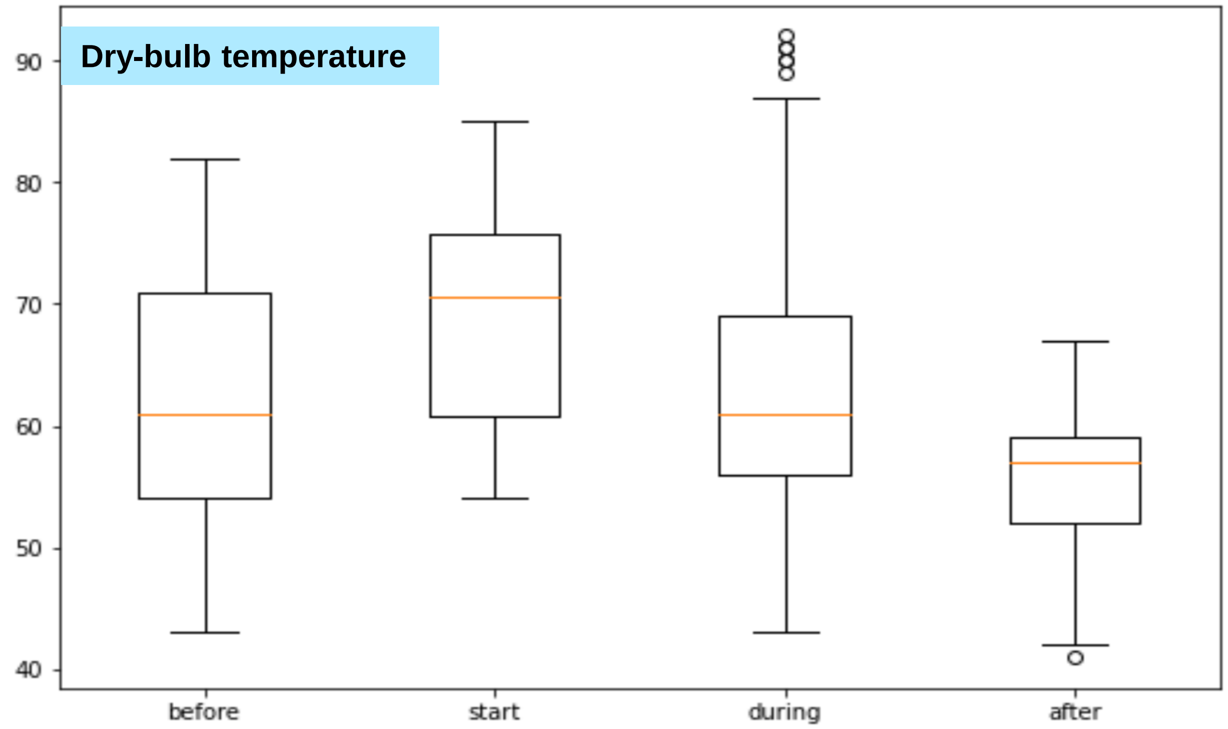


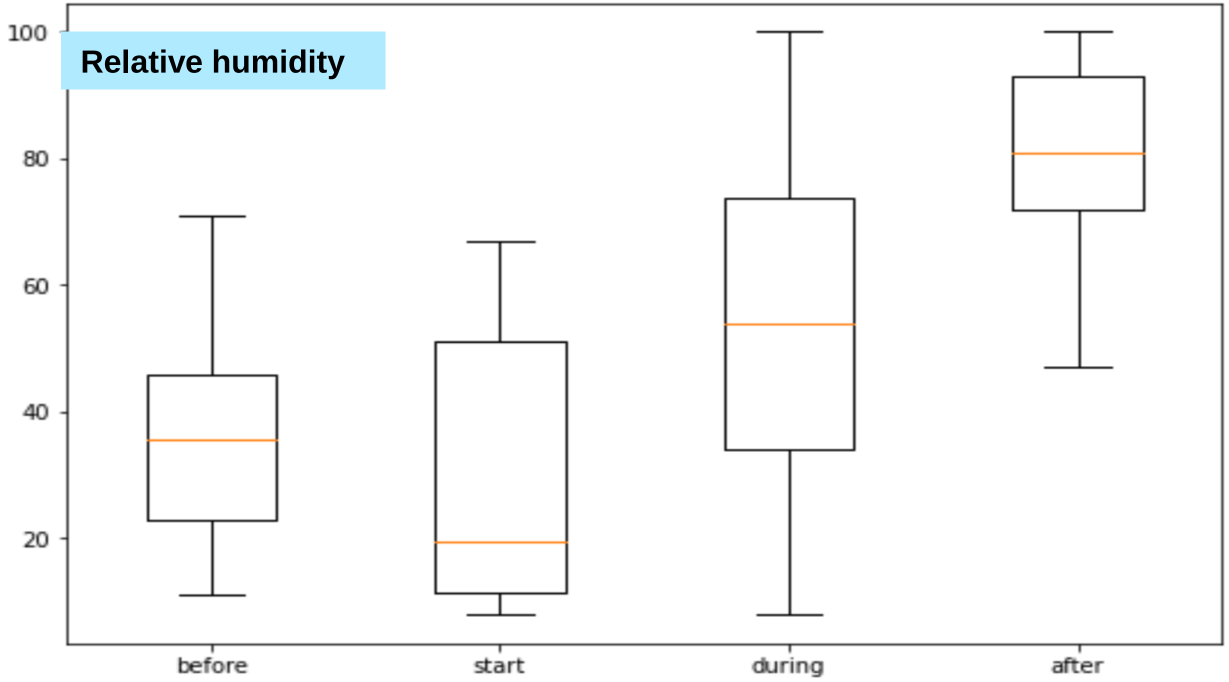
Fields	Description	Range	Unit
WBAN	WBAN is a five-digit number used for digital data storage and weather station identification.	(117, 94299)	NA
Latitude	Latitude of the weather station	(32.81667, 41.78139)	Degree
Longitude	Longitude of the weather station	(-115.57861, -124.23667)	Degree
Elevation	Elevation of the weather station	(-17.7, 2172.6)	Meter
Date	Time at which different weather parameters are captured.	(01/01/2016, 12/31/2018)	NA
Hourly DryBulb Temperature	Dry Bulb Temperature of air at the weather station.	(-22.0, 115.0)	Fahrenheit
Hourly Relative Humidity	Relative humidity at the weather station.	(1.0, 100.0)	Percentage
Hourly WindSpeed	Wind speed at the station.	(0.0, 57.0)	Miles per Hour (mph)
Hourly Precipitation	Precipitation at the station.	(0.0, 10.31)	Inches to Hundredths

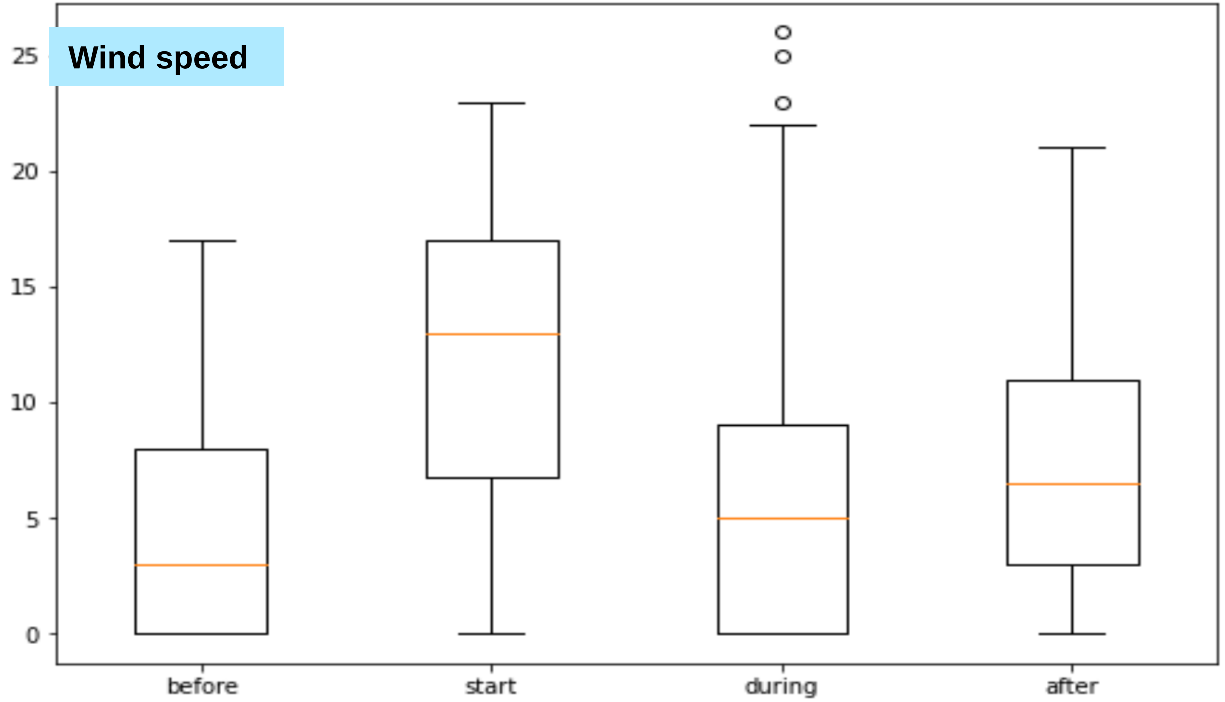
Weather Stations



Hourly Dry-bulb temperature, relative humidity and wind speed
- Comparison of Before, Start, During and After Fire

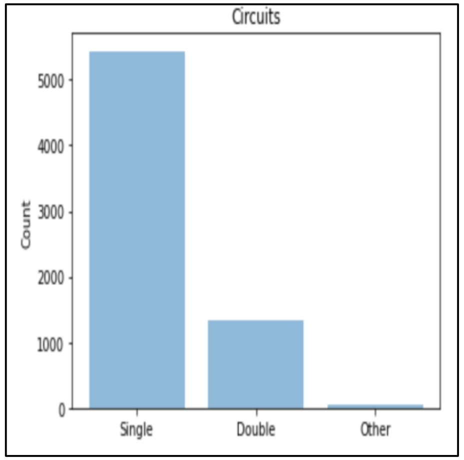


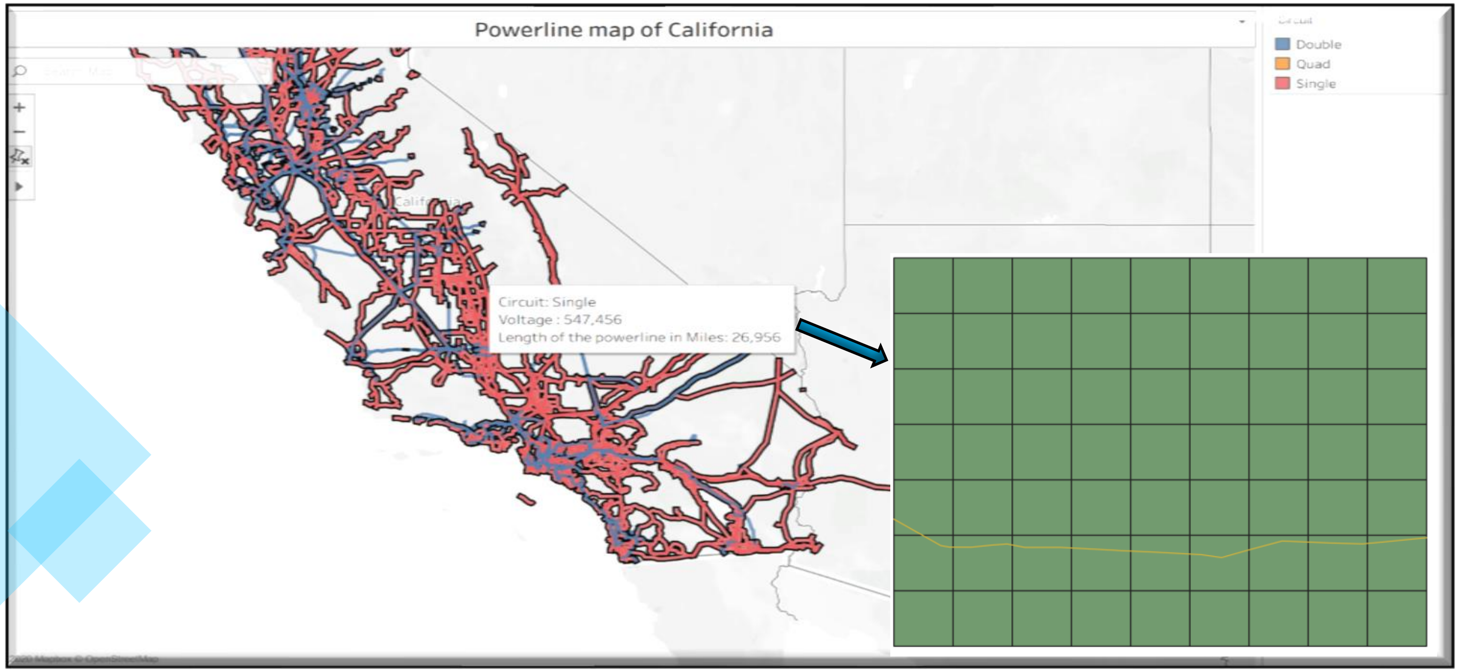




Powerline Data

- California Energy Commission created an Electric transmission line geospatial data layer to display the electric transmission grid in California.
- **Format** – Shape file
- **Size -** 58 KB
- **Source** - https://www.energy.ca.gov/
- Converted Shape file to data frame using QGIS tool

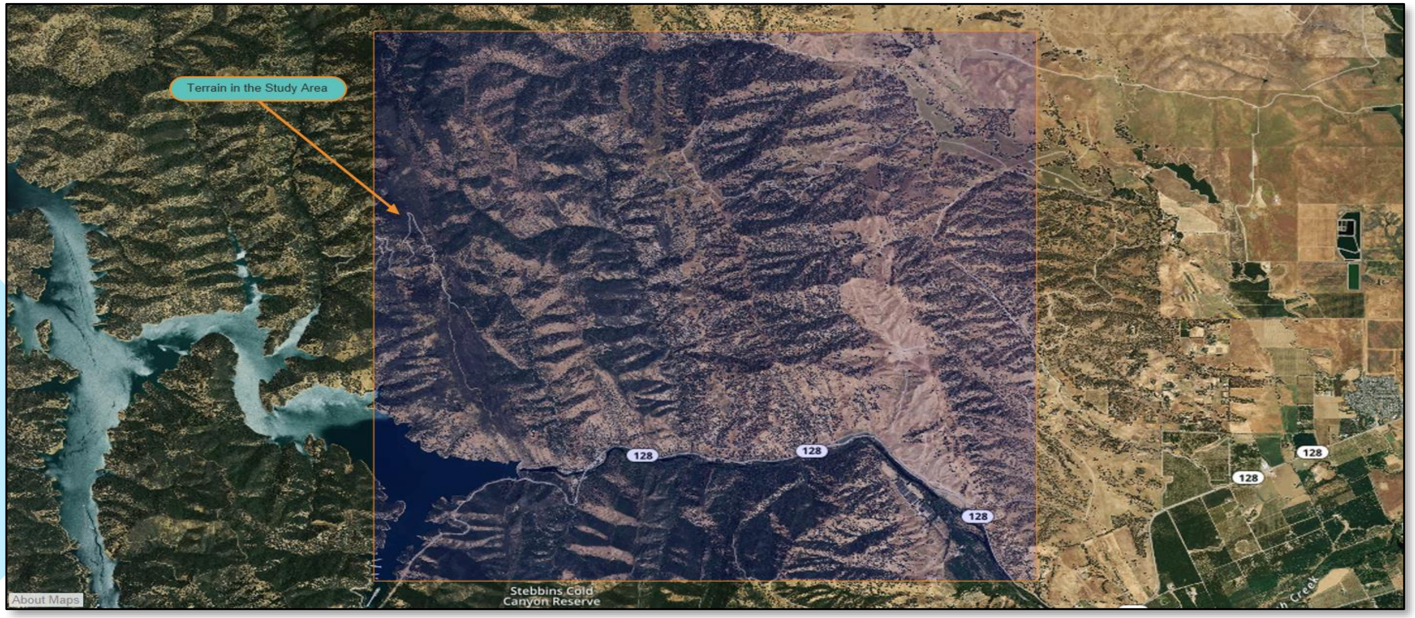




Parameter	Description
kV	Voltage of the powerline
Owner	Owner of the powerline
Status	Whether the line is operating or not
Circuit	Type of circuit
Length (Mile)	Length of the line in miles
Length (Feet)	Length of the line in Feet

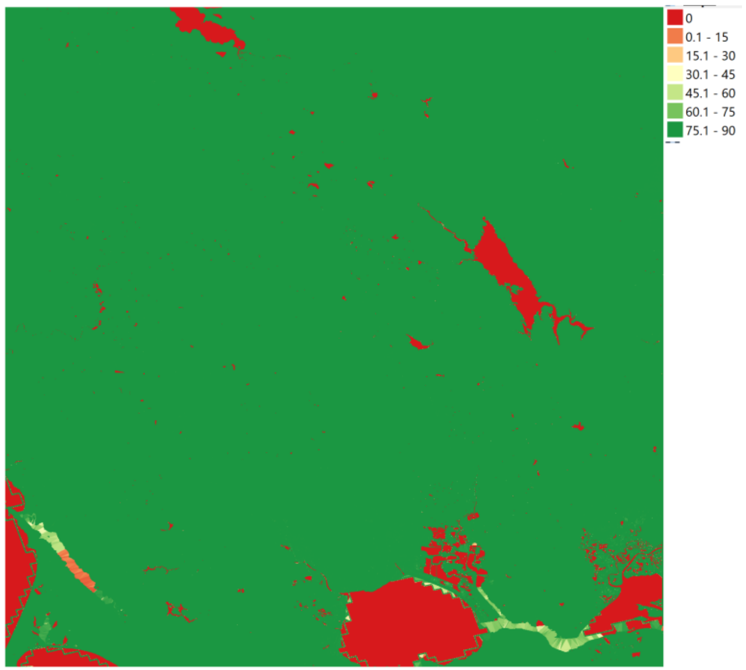
Terrain Data

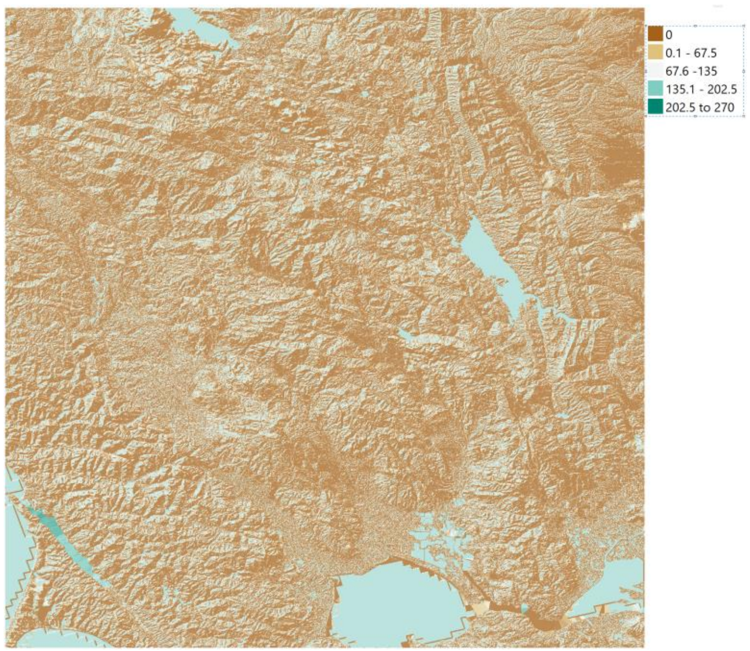
- United States Geological Survey provides topography information, which is available in the form of Digital Elevation Models (DEM).
- **Format** - DEM file
- **Size** – 1.6 GB
- Calculated parameters from DEM using QGIS tool.
- **Source** - https://www.usgs.gov/

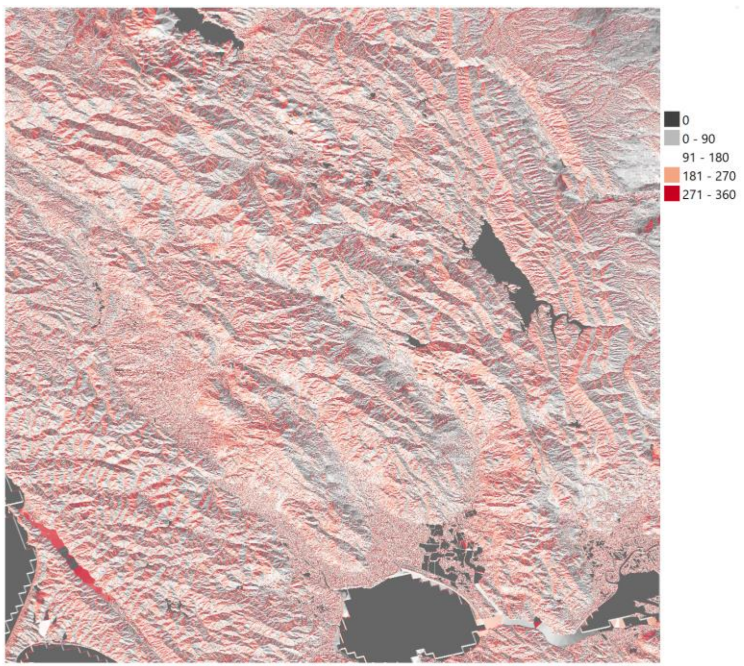


Parameter	Description	Range	Unit
Slope	Slope of the terrain	(0, 90)	degree
Aspect	It is a compass direction that a slope face	(0, 360)	azimuth
Hill shade	Terrain surface with sun relative position.	(0,360)	azimuth

Slope, Hill Shade & Aspect

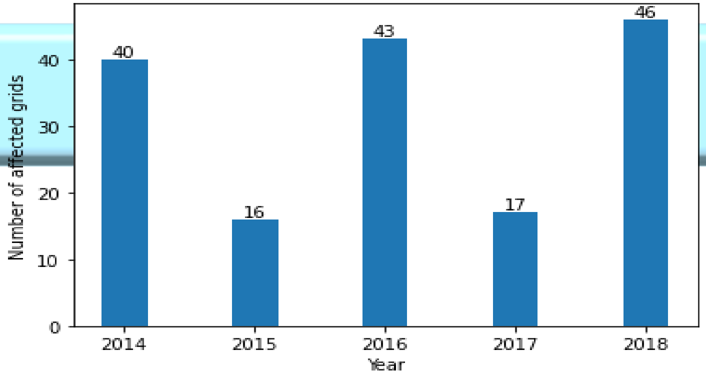


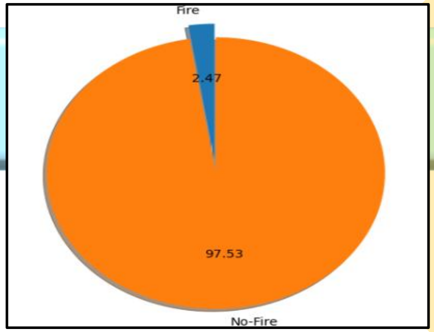


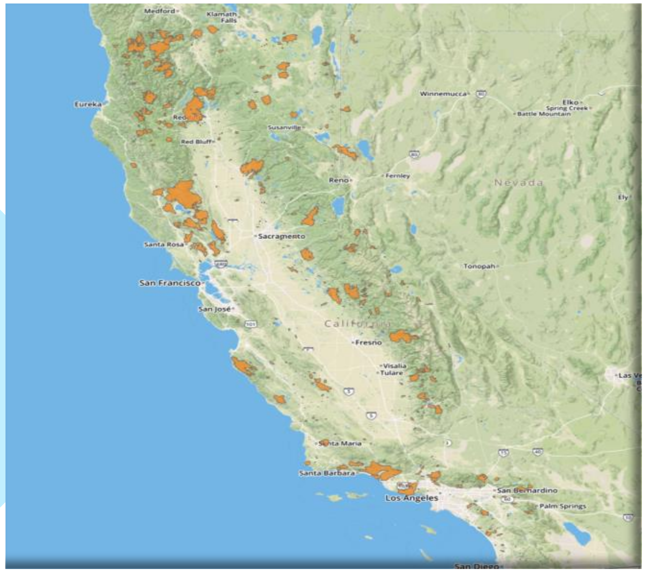


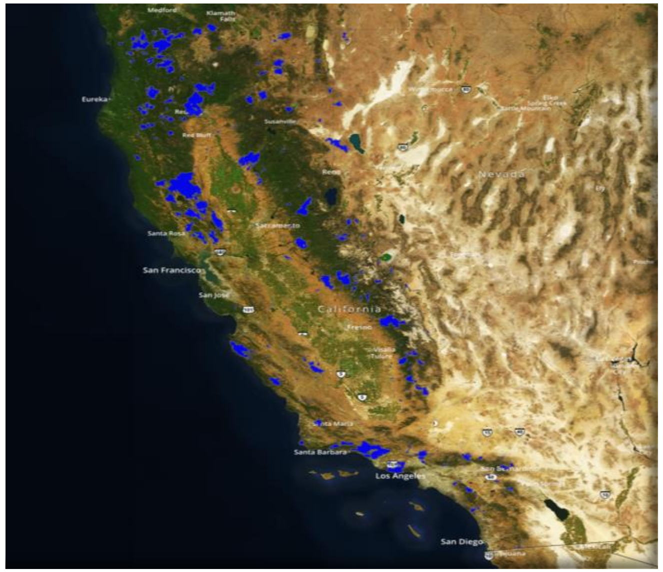
Fire History Data

- Fire and Resource Assessment Program (FRAP) maintained by CAL Fire, United States Forest Service region, Bureau of Land Management and National park service.
- **Format** - Shape file
- **Date range** - 01/01/2014 to 12/31/2018.
- **Source** - https://frap.fire.ca.gov/frap-projects/fire-perimeters/
- **Size** – 80 KB







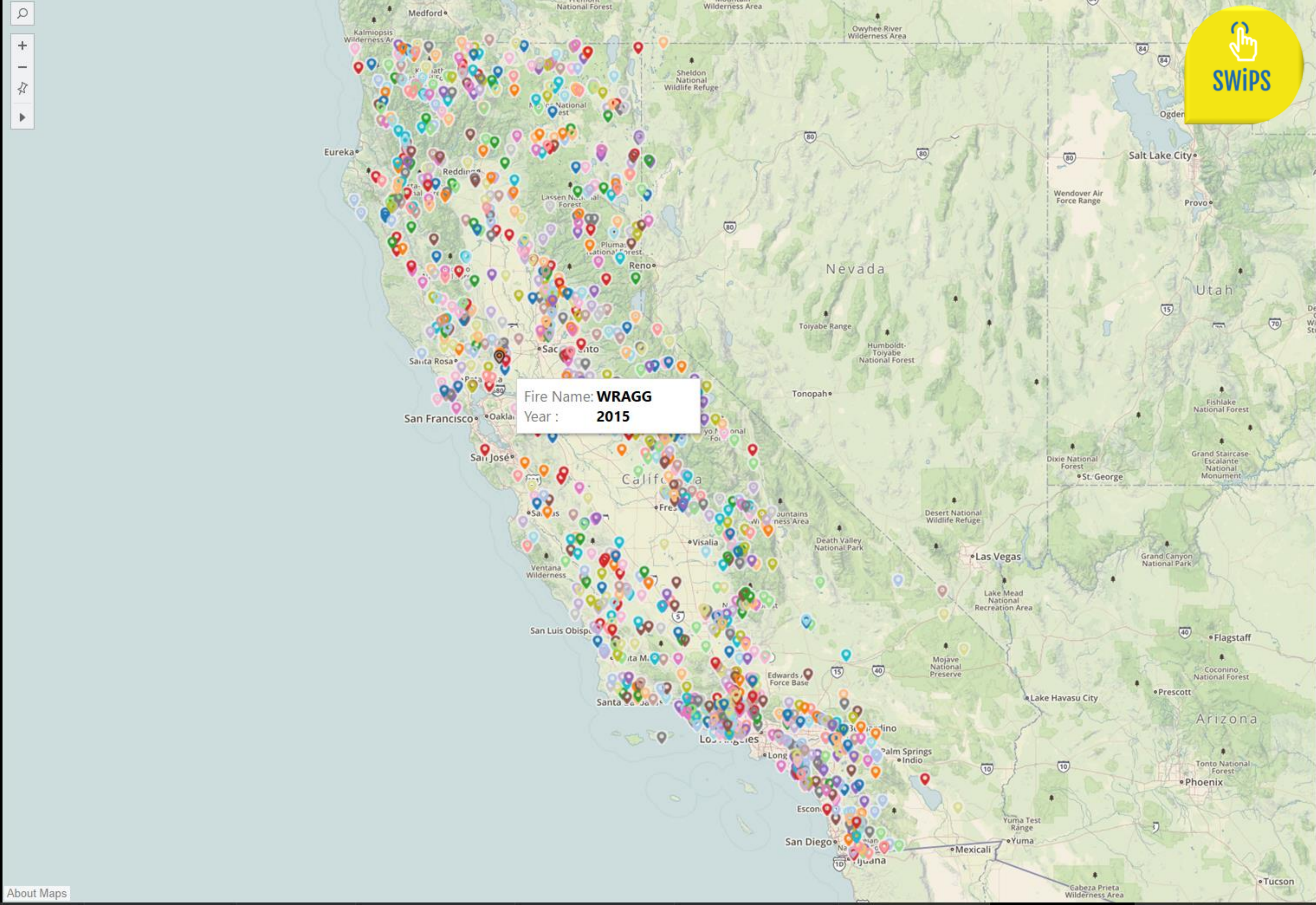


Column Name	Description	Range	Unit
Year	Year of Fire occurred.	(2016, 2018)	year
State	State where the fire occurred	‘CA’	NA
Fire_Name	Name of the fire	NA	NA
Alarm_Date	Alarm date of the Fire.	(01/01/2016 ,12/31/2018)	NA
Cont_Date	Containment date for fire.	(01/01/2016 ,12/31/2018)	NA
Cause	Reason for the fire.	(1,19)	NA
Report_Ac	Area Consumed in the fire	(0, 25)	acre
GIS_Acres	Area Calculated by GIS	(8.266294, 26.002495)	acre
C_Method	Collection of data method coding	(1, 1)	NA
Objective	Suppression or resource benefit	(1,1)	NA
Fire_Num	Number assigned to the fire	(00001890 1716, 00000825)	NA
Shape_Length	Length of the area burnt	(9.093210, 445282.444798)	meter
Shape_Area	Area burnt	(6.130331,1.660030e +09)	Square meter
Geometry	Shape of the area burnt	Within study area	degrees

Wildfires in California
(Years 2014 to 2018)

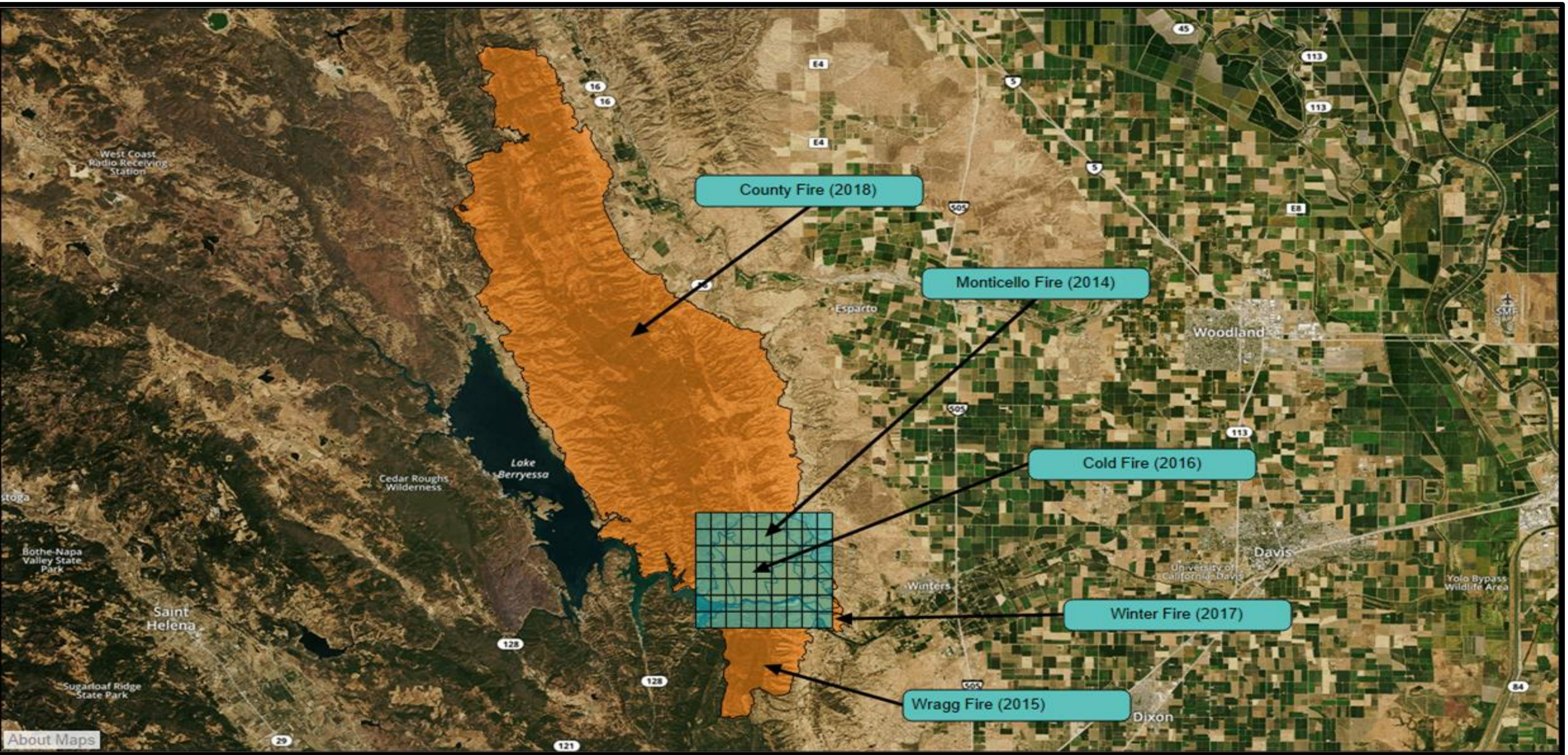
Select the View type

- General View
- Satellite View
- Street View

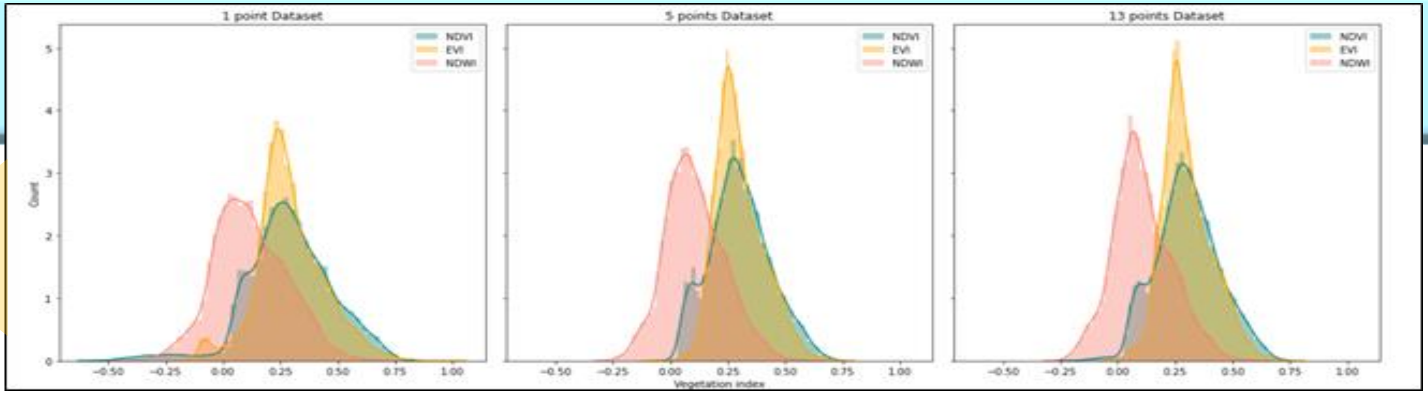


- ZERMATT
- ZENON
- ZEBRA
- ZAMORA
- YUCCA
- YOUNG
- YOSEMITE CREEK
- YORK
- YELLOW
- WRIGHT
- WRAGG
- WOOLSEY
- WOODLOT
- WOODCHOPPER 2
- WOODCHOPPER 1
- WOLFE
- WINTON
- WINTER
- WINERY
- WINDMILLS
- WINDMILL
- WIND
- WINCHESTER 1-47
- WILLMS
- WILLIE

Fire History for our Study Area



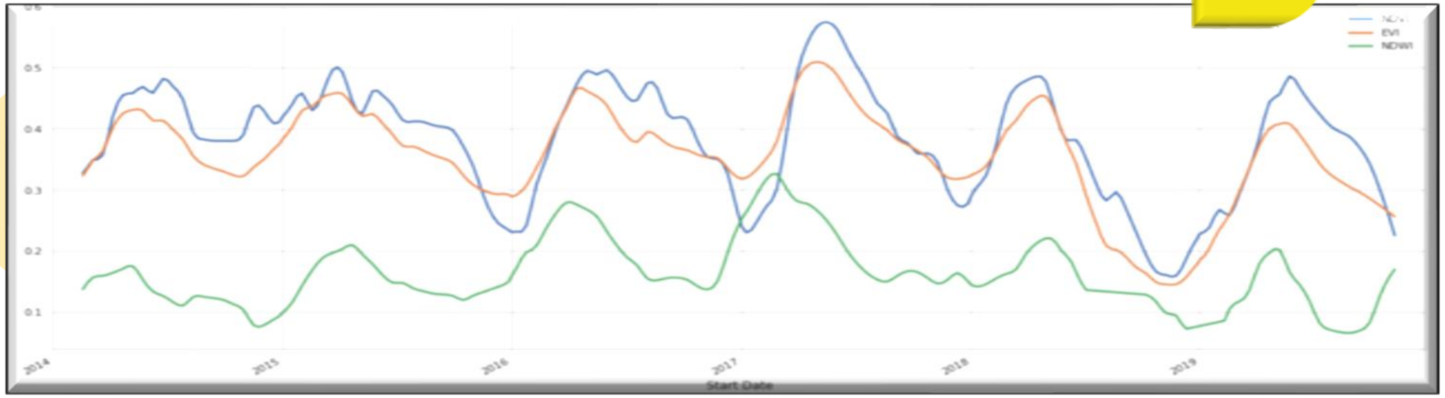
Vegetation Data

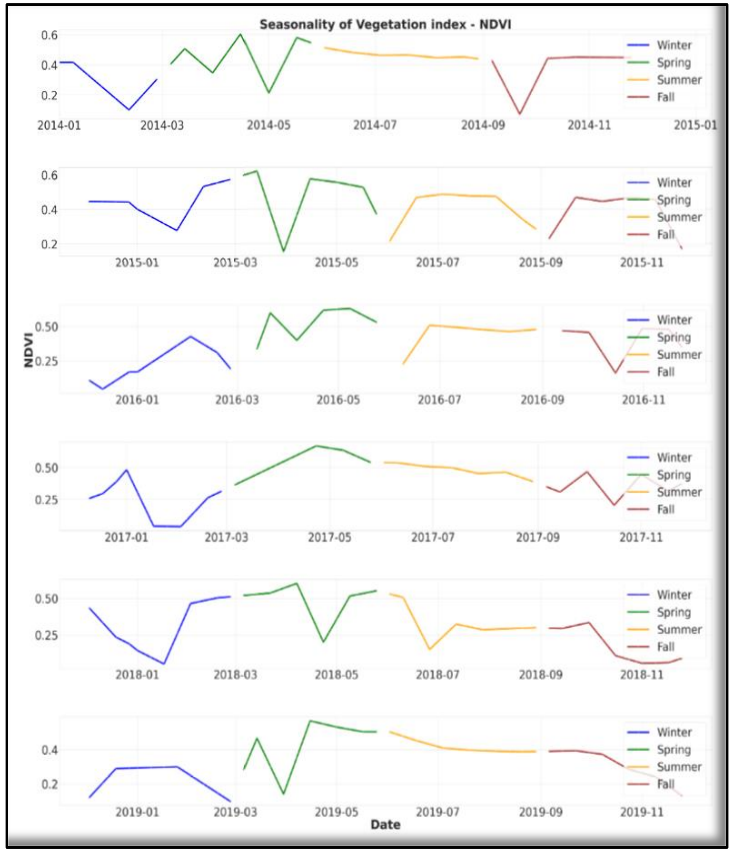


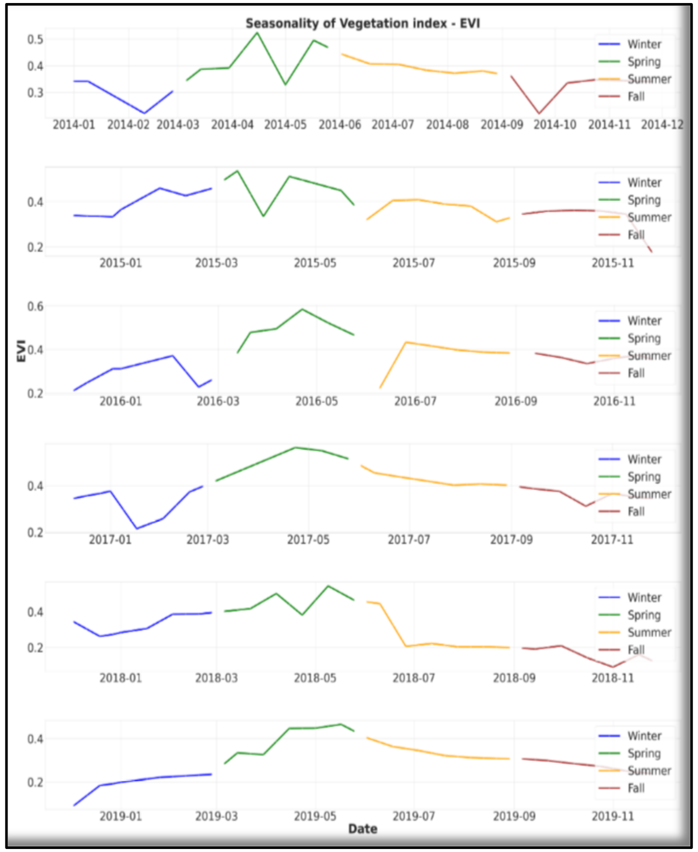
- Landsat 8 remote-sensing satellite data was extracted from Google Earth Engine (GEE) using Python API.
- Custom-made python functions were used for calculating indices from bands as well as fetching aggregate values.
- Aggregates were consistent and reliable as they were pre-processed and devoid of atmospheric hindrances.
- **Size** - 16 MB (final dataset)
- **Dimension** - 226044 rows x 8 columns
- **Frequency** - 16 days (Landsat 8)
- **Spatial Accuracy** - 15 to 30 metres
- **Date Range -** 1st January,2014 to 3rd January, 2020

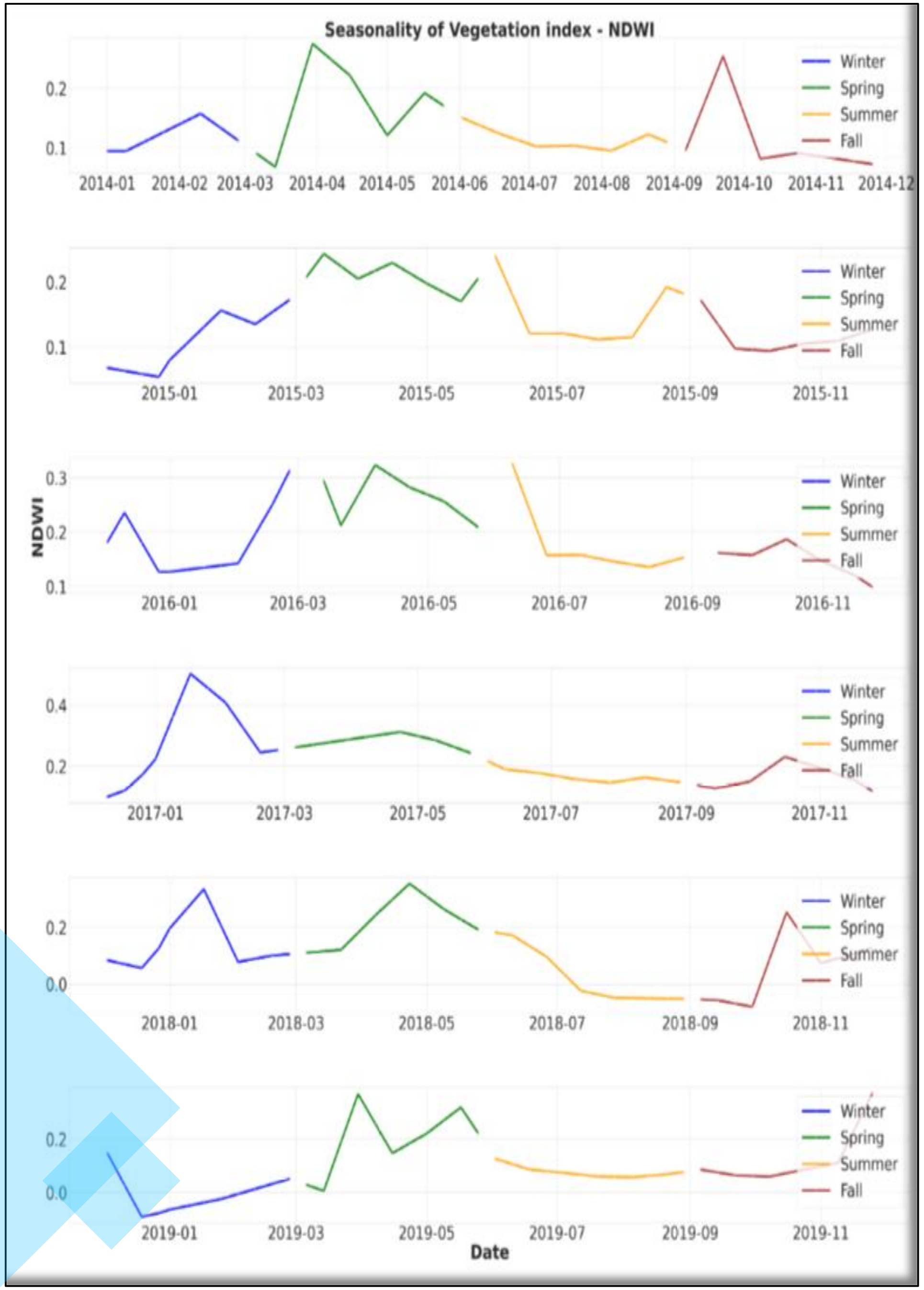
Fields	Data Type	Description	Range	Unit
id	Integer	Unique id for the inner grid	1 to 63	NA
Co-ordinates – Corners and Centroid	Float	Separate latitude and longitude of Left, right, top, bottom and centroid latitude and longitude of each grid	((38.4992, -122.1241), (38.5692, -122.0341))	Degree
Polygon/Geometry	Polygon	Grid coordinates of the inner grid	((38.4992, -122.1241), (38.5692, -122.0341))	Degree
start_date	Datetime	Start date for fetching the 8-day composite indices	(01/01/2015, 12/27/2019)	NA
end_date	Datetime	7 days added to the start date provides the end date for 8-day composite indices	(01/08/2015, 1/3/2020)	NA
NDVI	Float	Calculated from red and infrared bands.	-1 to +1	NA
EVI	Float	Calculated from visible and near-infrared bands.	-1 to +1	NA
NDWI	Float	Calculated from short-wave infrared and near infrared bands	-1 to +1	NA
13 co-ordinates for corners, diagonals, midpoints of edges and corners	Float	13 columns corresponding to the 13 points sampled from each inner grid. The points are Centroid, 4 diagonal points, 4 midpoints of edges and corner points.	((38.4992, -122.1241), 38.5692, -122.0341))	Degree

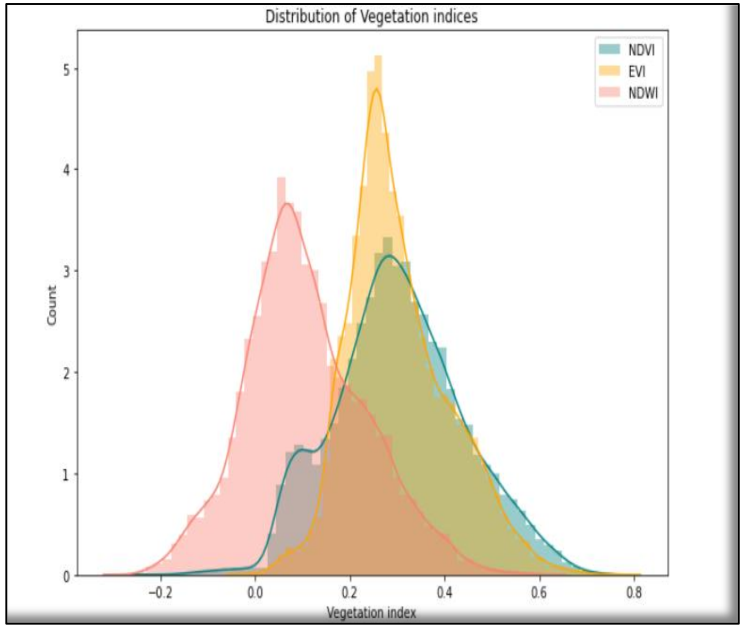
- **Vegetation indices** are measures that map the density of green vegetation around the globe. Below are the chosen relevant indices.
- **Normalized Difference Vegetation Index (NDVI)** - Standard measure of healthy vegetation
- **Enhanced Vegetation Index (EVI) -** Enhanced version of NDVI
- **Normalized Difference Water Index (NDWI)** - Water content in Vegetation
- **Data sampling** - For each grid, a 1-point (Centroid only), 5-point (Centroid and corners) and 13-points dataset (Centroid, Corners, edges of midpoints and diagonals) were generated. The 13-point was chosen due to its completeness after a thorough statistical analysis.

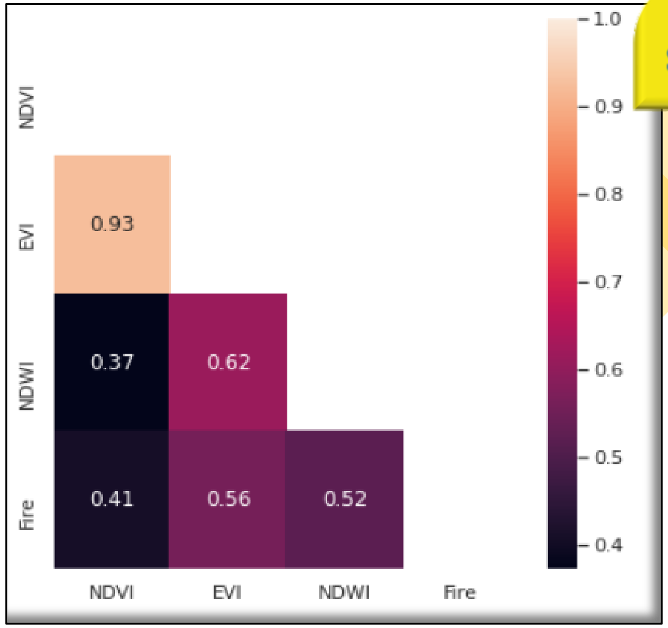


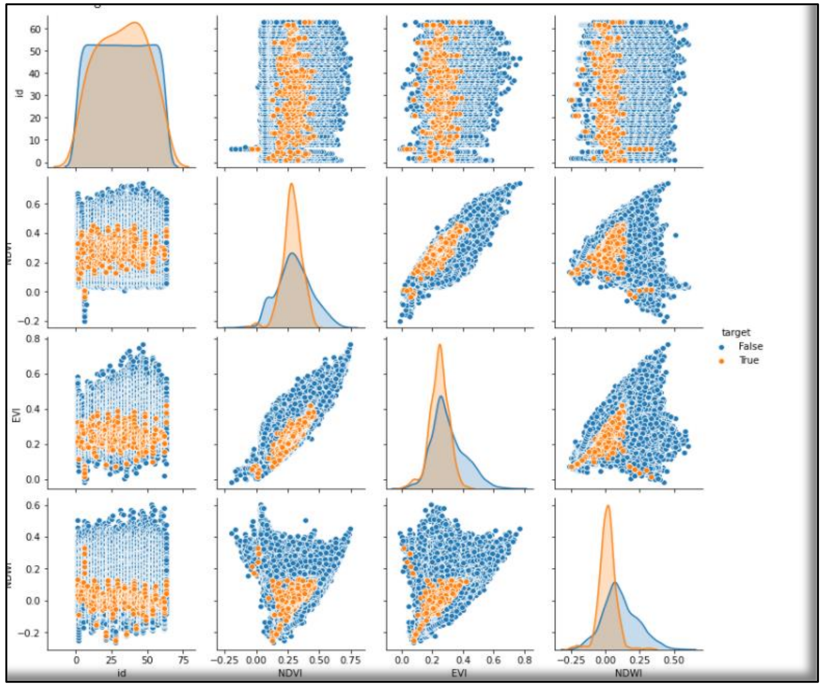


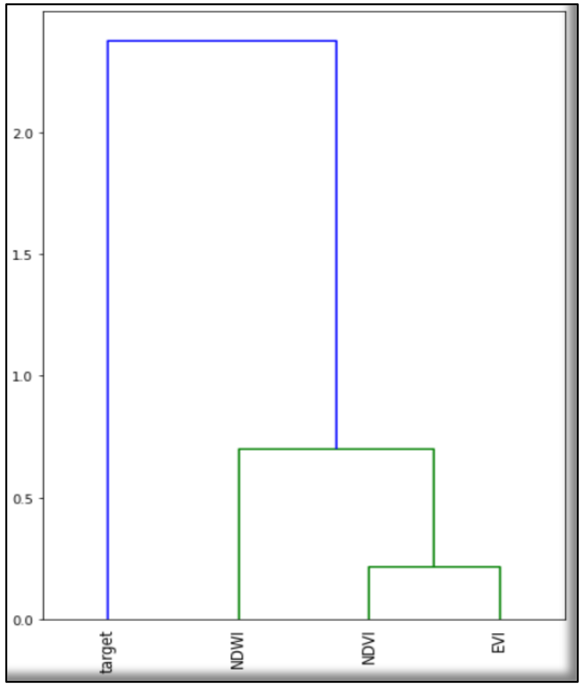



Seasonality of Vegetation index - NDWI
Winter
Spring
Summer
Fall
NDWI
Date

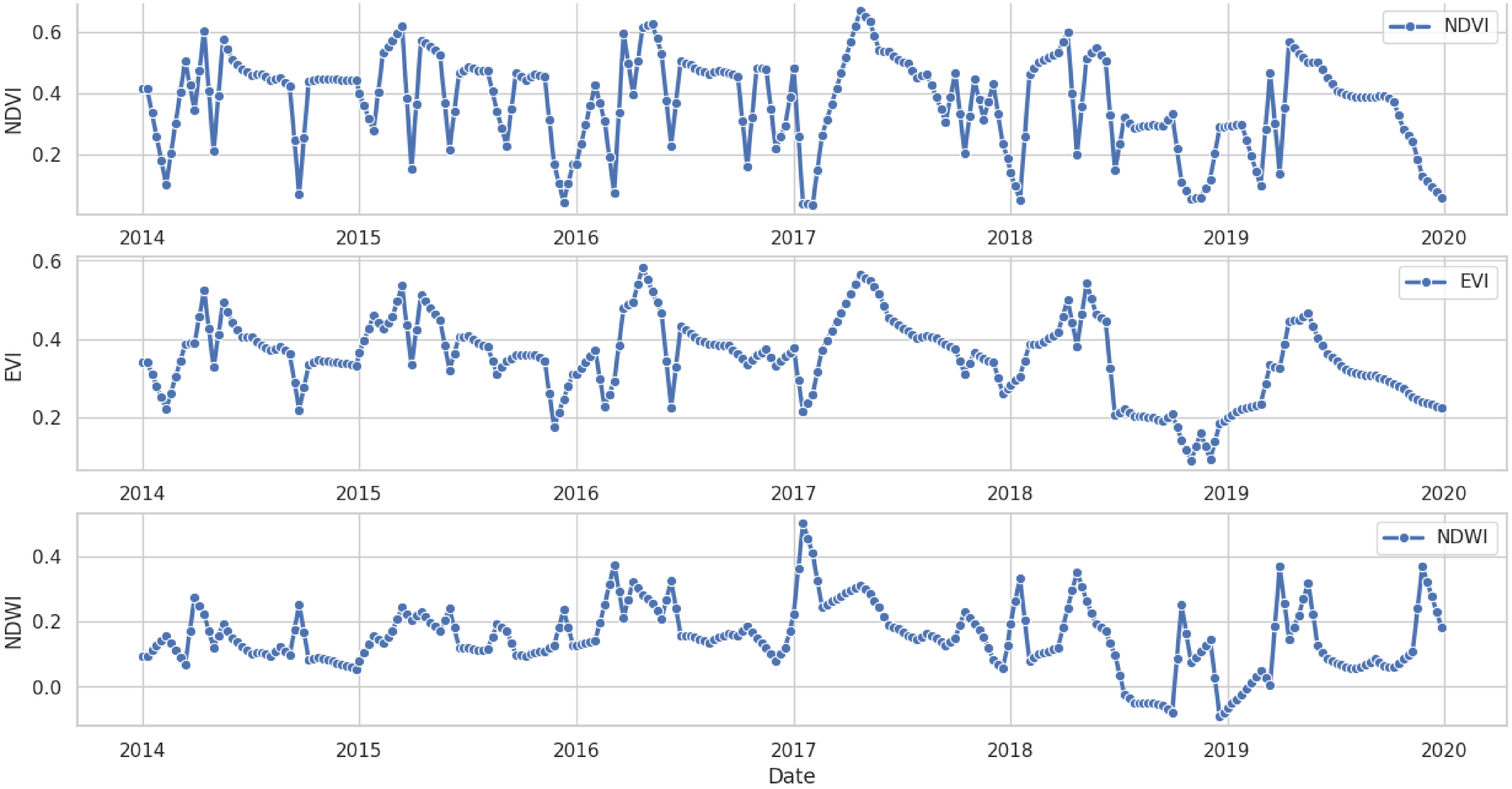

Distribution of Vegetation indices
NDVI
EVI
NDWI
Count
Vegetation index

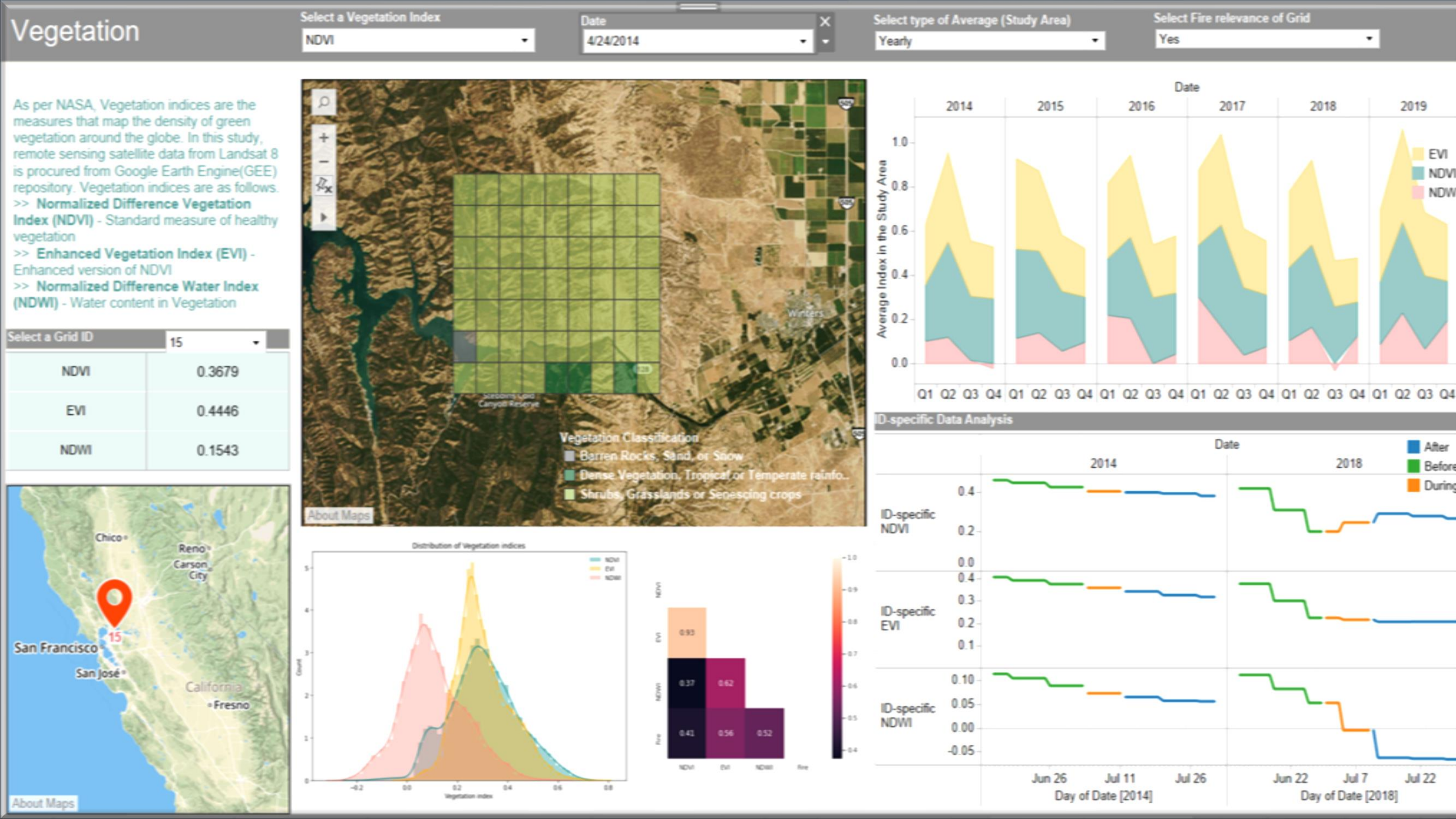

NDVI
EVI
NDWI
Fire
0.93
0.37
0.62
0.41
0.56
0.52


target
False
True
id
NDVI
EVI
NDWI

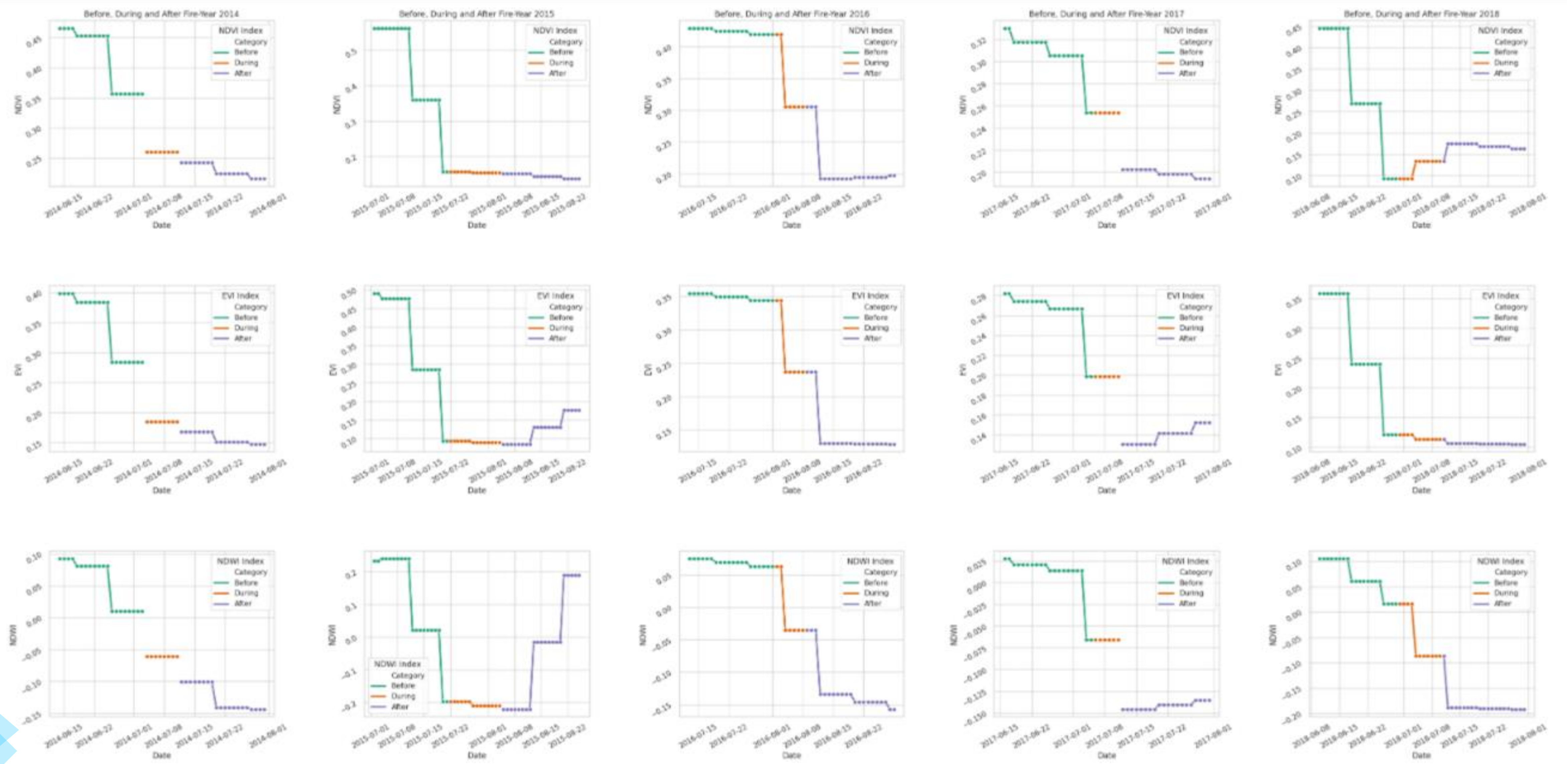

target
NDWI
NDVI
EVI

Landsat 8 Vegetation - time-series data (dashboards in coming slides)

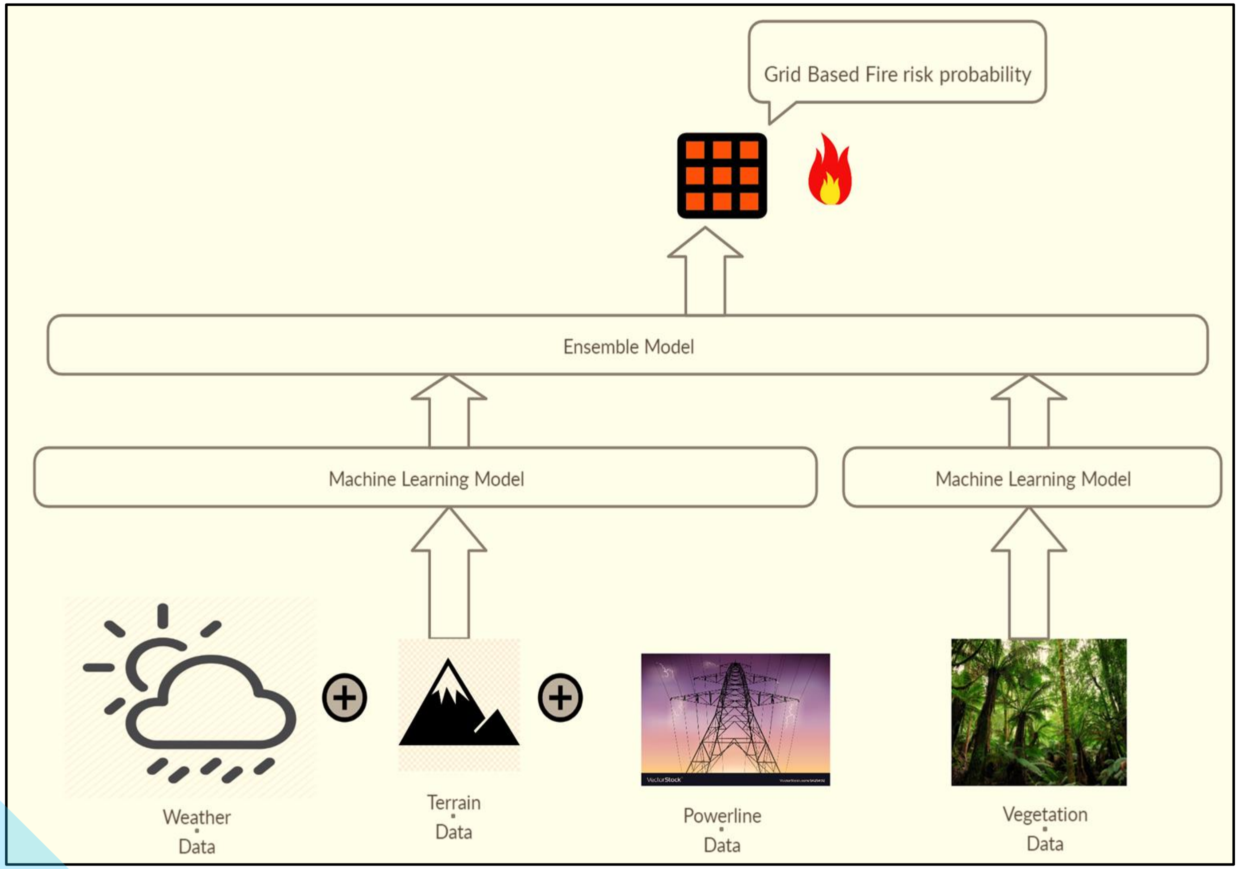



Vegetation
Select a Vegetation Index
NDVI
Date
4/24/2014
Select type of Average (Study Area)
Yearly
Select Fire relevance of Grid
Yes
As per NASA, Vegetation indices are the measures that map the density of green vegetation around the globe. In this study, remote sensing satellite data from Landsat 8 is procured from Google Earth Engine(GEE) repository. Vegetation indices are as follows.
>> Normalized Difference Vegetation Index (NDVI) - Standard measure of healthy vegetation
>> Enhanced Vegetation Index (EVI) - Enhanced version of NDVI
>> Normalized Difference Water Index (NDWI) - Water content in Vegetation
Select a Grid ID
15
NDVI 0.3679
EVI 0.4446
NDWI 0.1543
Chico
Reno
Carson City
San Francisco
San José
California
Fresno
15
About Maps
Vegetation Classification
Barren Rocks, Sand, or Snow
Dense Vegetation, Tropical or Temperate rainfo...
Shrubs, Grasslands or Senescing crops
Winters
About Maps
Date
2014
2015
2016
2017
2018
2019
Average Index in the Study Area
EVI
Q1 Q2 Q3 Q4
ID-specific Data Analysis
Date
2014
2018
After
During
ID-specific NDVI
ID-specific EVI
ID-specific NDWI
Jun 26
Jul 11
Jul 26
Day of Date [2014]
Jun 22
Jul 7
Jul 22
Day of Date [2018]
Distribution of Vegetation indices
NDVI
EVI
NDWI
Count
Vegetation index
0.93
0.37
0.62
0.41
0.56
0.52
NDVI
EVI
NDWI
Fire

Grid-wise pre-fire, during and post-fire analysis for Vegetation (2014 - 2018)

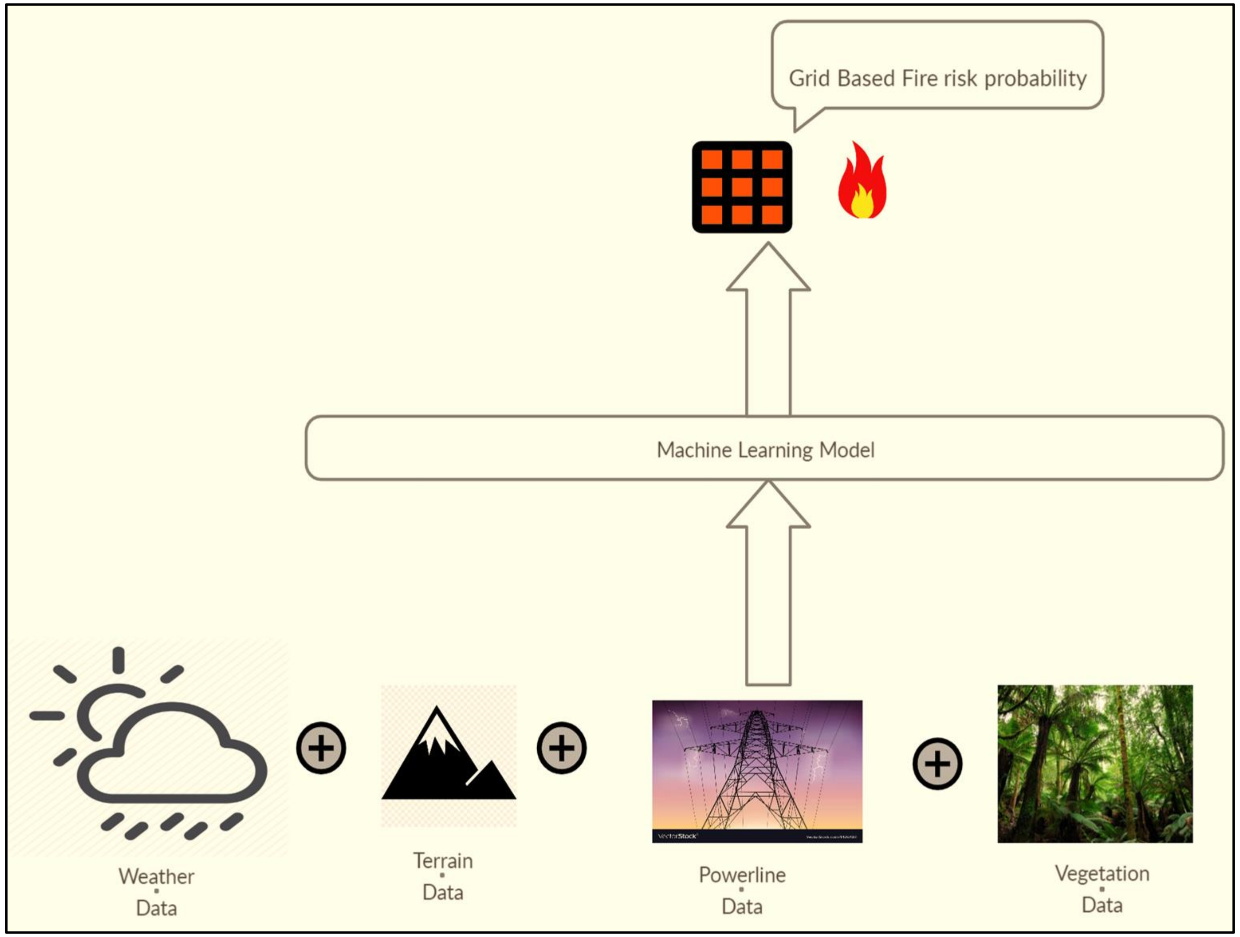


Proposed Solutions



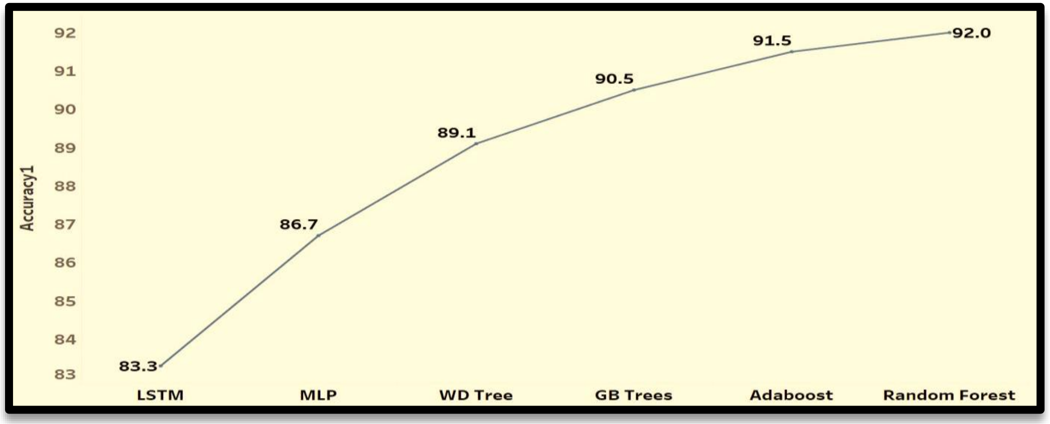
Ensemble Model

Combined Model

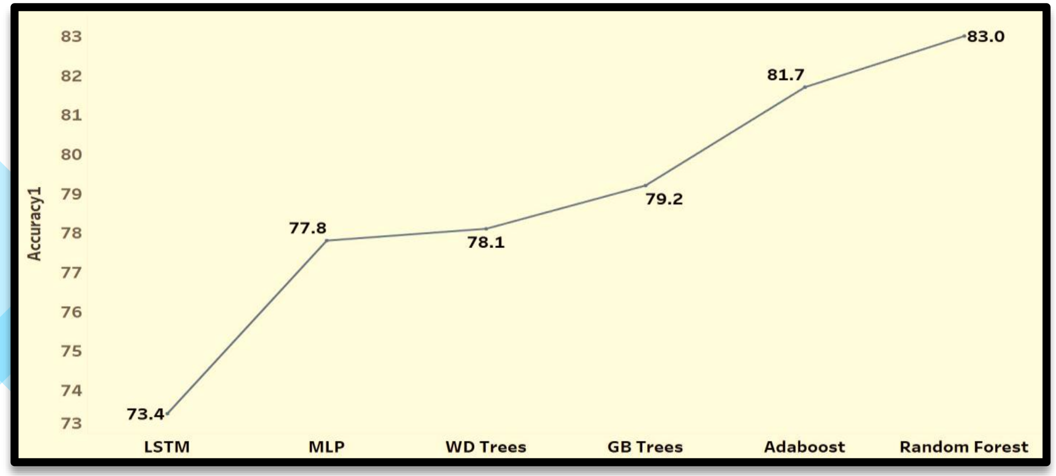


Accuracy Comparison of the Machine learning algorithms

Combined Models



Ensemble Models



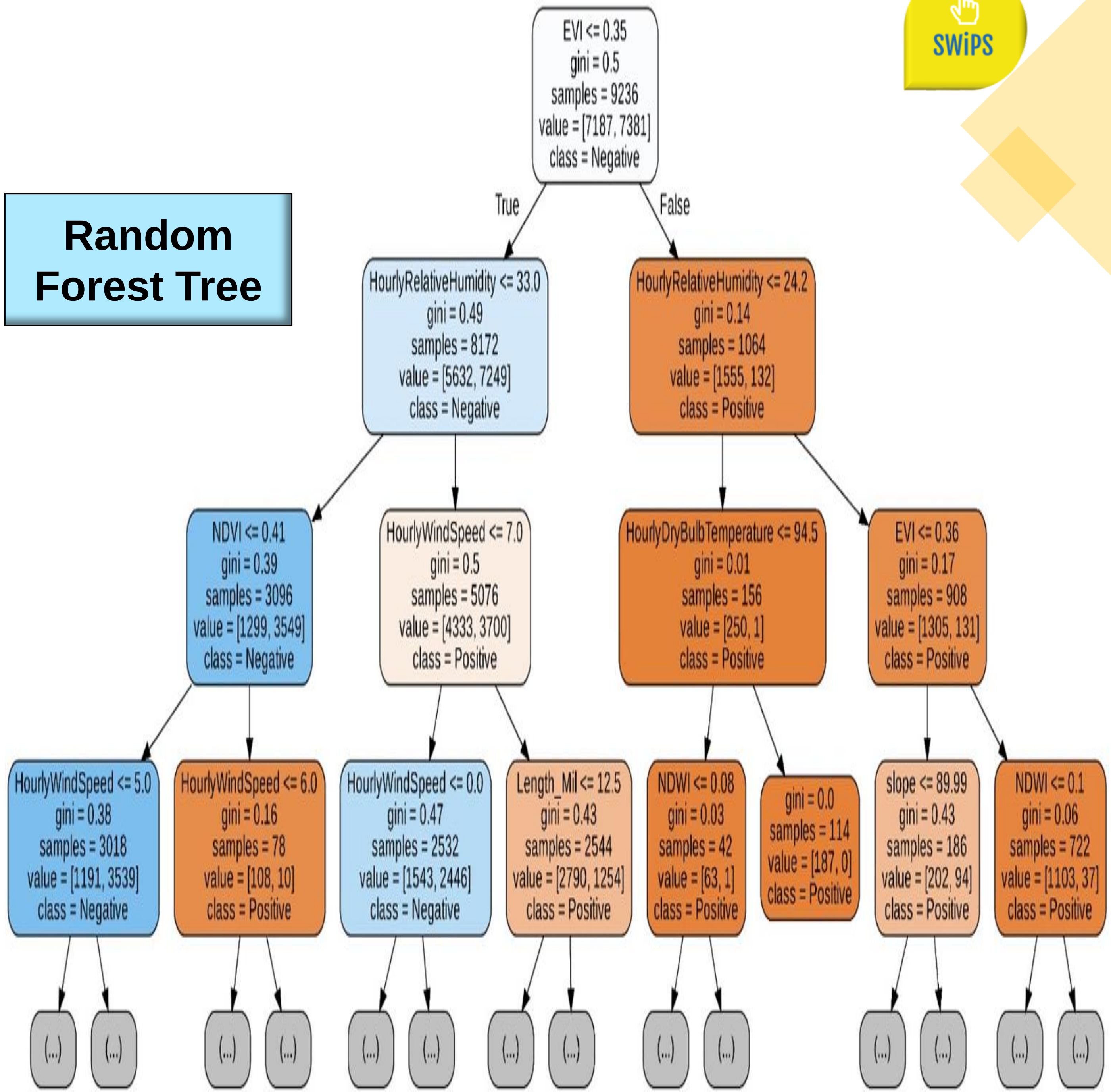
Criteria for selecting Machine Learning (ML) Algorithms for our Model:

- Accuracy of the model.
- Minimize False Negatives.
- Interpretability of the model.
- Algorithms with least assumptions about the data.
- Scalability of the model.
- Infrastructural resources at hand.

Model Experiments

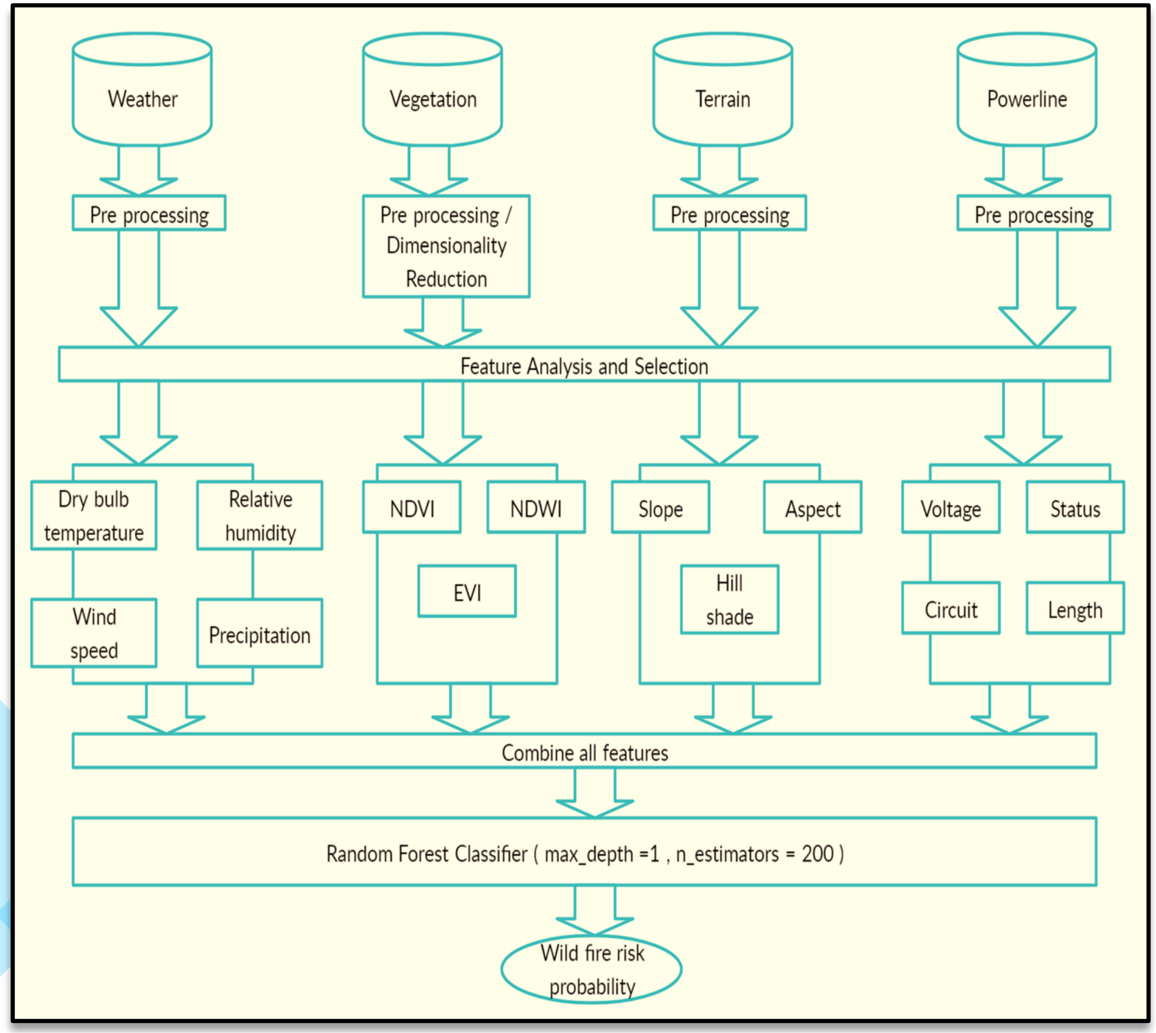
Model	Accuracy	Hyper Parameters used
Random Forest	92	n_estimators = 200
Adaboost	91.5	n_estimators= 50, learning_rate = 1
Gradient Boosting trees	90.5	loss = deviance, n_estimators= 100
Weighted Decision Trees	89.1	criterion = gini, splitter = best
MLP	86.1	activation = Softmax, solver = adam
LSTM	91.6	Dropout = 0.2, activation = Softmax

Random Forest Tree

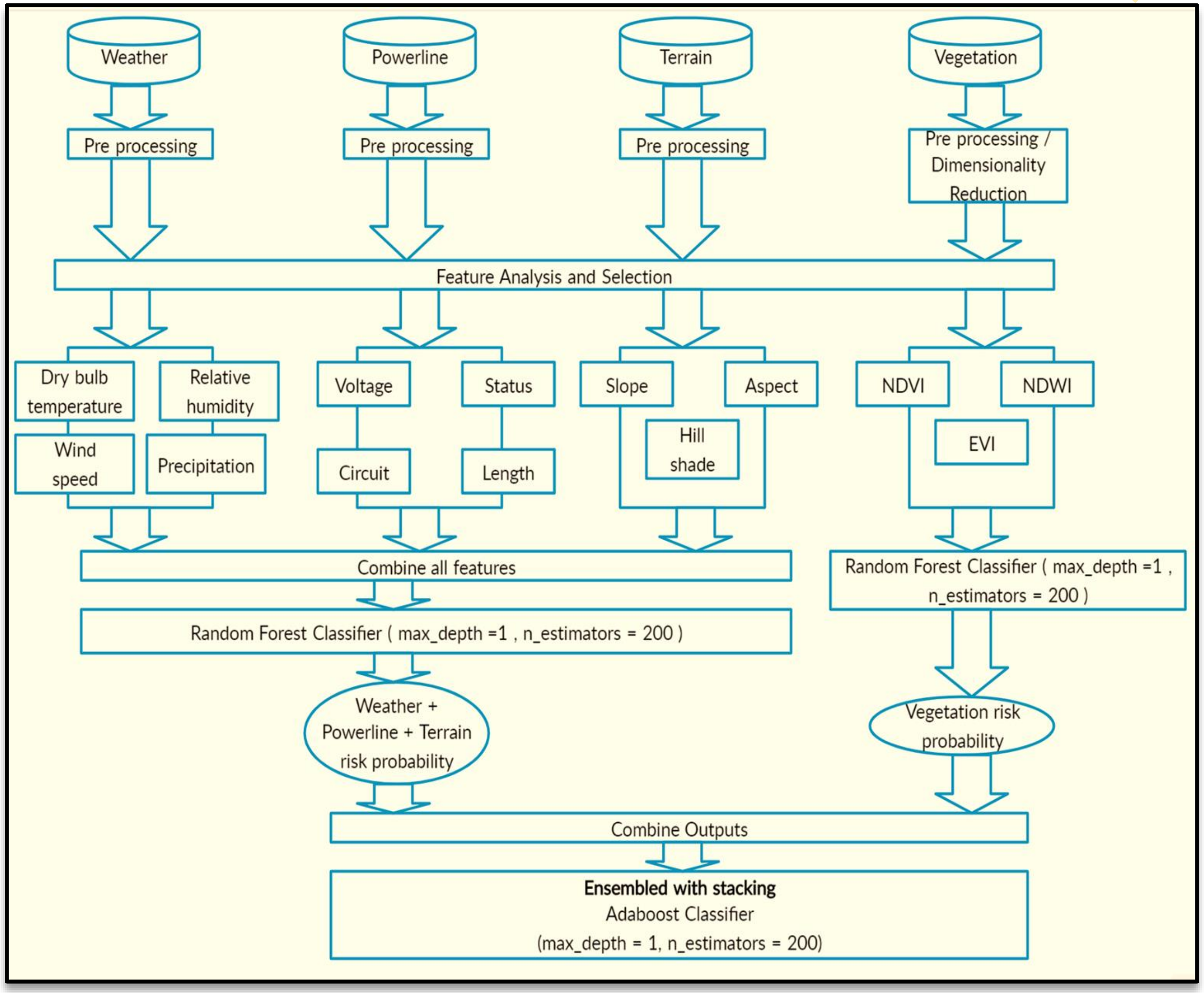


Model Architecture

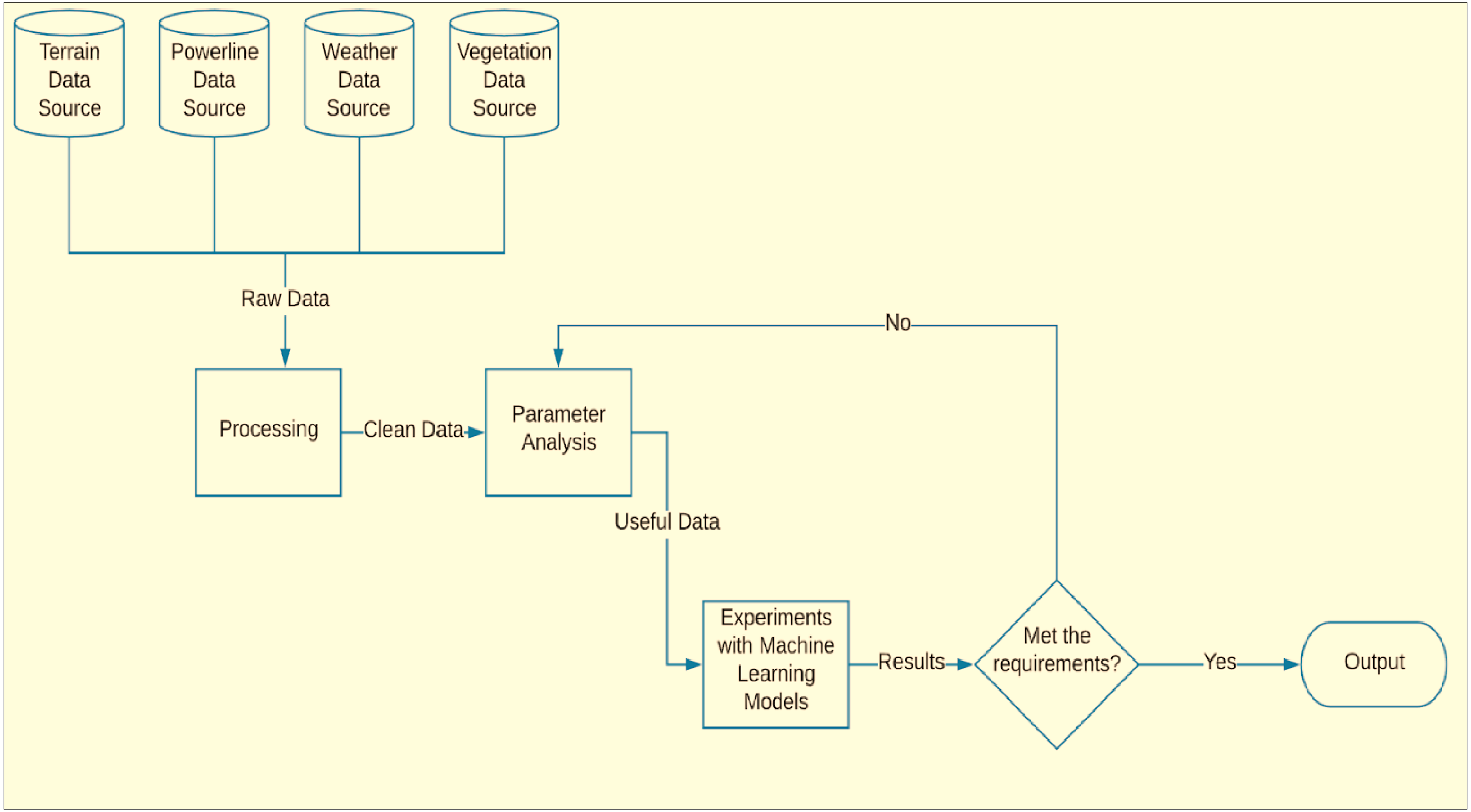
Final Combined Model



Final Ensemble Model



Evaluation strategy



Threshold tuning

- **Receiver Operating Characteristic (ROC Curve)** plots the False positive rates versus the True positive rates predicted by the model.
- Area under the curve is a true mark of model efficiency and its ability to classify.
- Using this ROC Curve, best threshold value for classification was determined.
- Tuning the threshold values reduced the False Negatives and improved accuracy of model.

Improved Model results with the best Threshold

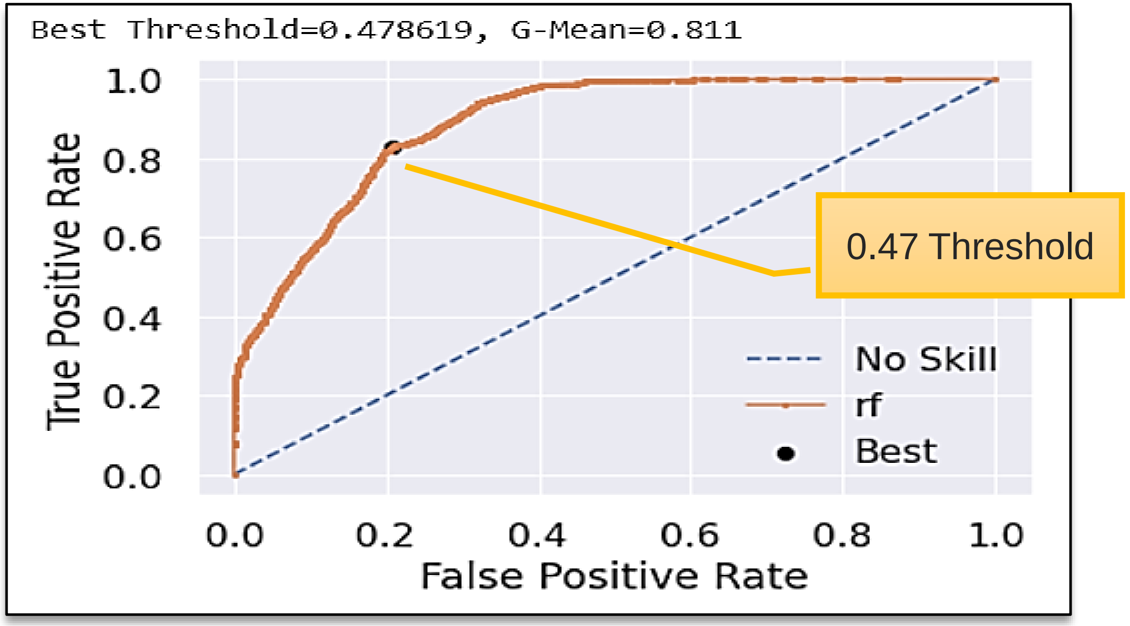
```
0.9223959870168191
Confusion Matrix :
[[1497  198]
 [  65 1629]]
Accuracy Score : 0.9223959870168191
Report :
              precision    recall  f1-score   support

           0       0.96      0.88      0.92      1695
           1       0.89      0.96      0.93      1694

    accuracy                           0.92      3389
   macro avg       0.93      0.92      0.92      3389
weighted avg       0.93      0.92      0.92      3389
```

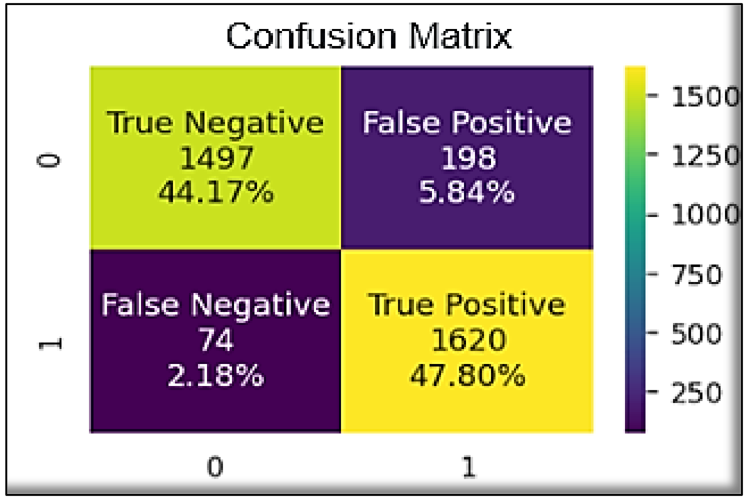
92% Accuracy

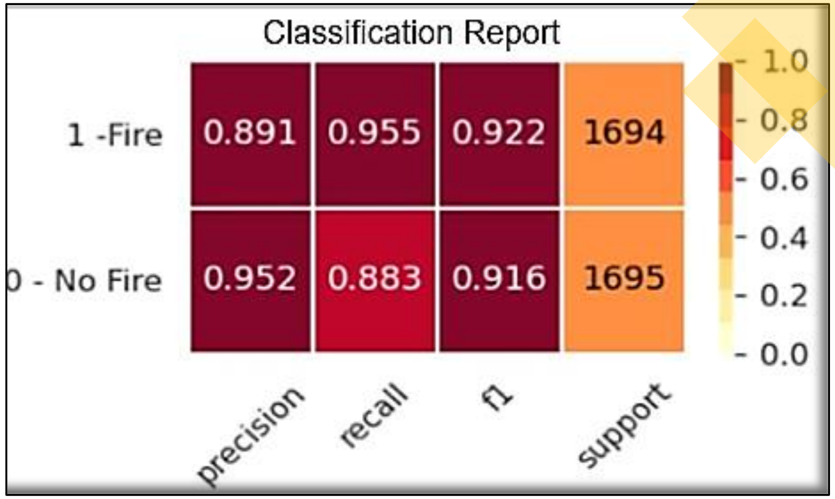
ROC Curve with the best Threshold value

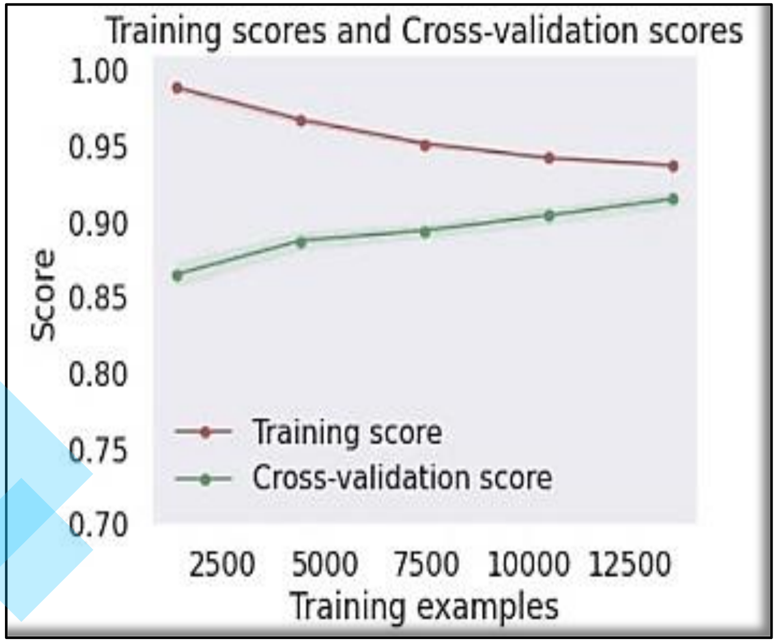


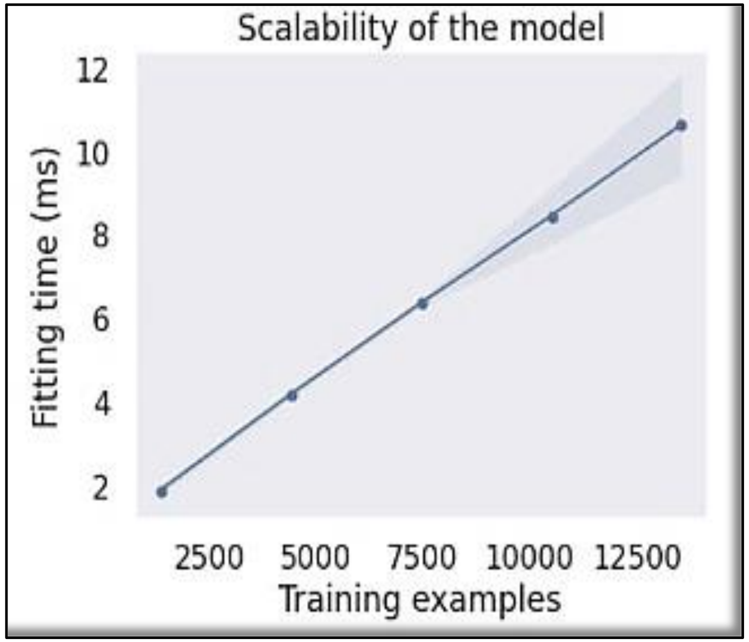
Evaluation Metrics for Random Forest (Combined Model)

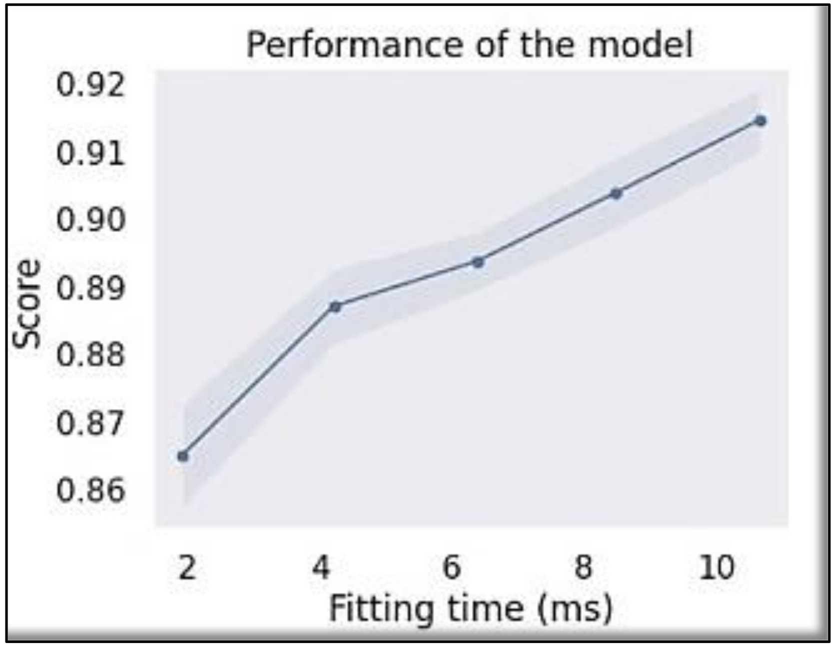
- **Confusion matrix** is a tabular representation of the number of true and false predictions.
- **Classification Report** is based on he confusion matrix and shows the percentage for Precision, recall and F1 Score for all classes.
- **Learning curves** are helpful in evaluating the feasibility, scalability and performance of the model.







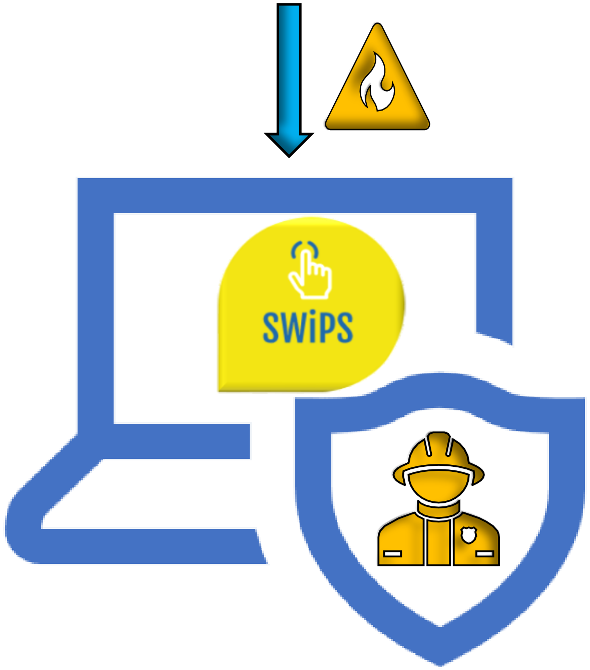




Conclusion

- A robust data-driven fire prediction system named ‘**Spartan Wild Fire Risk Prediction system (SWiPS)’** was created by modeling the precursor fire conditions with a combination strategy that mimics real-time wildfire dynamics.

- It comprises a powerful backend algorithm connected to a front-end interface using Flask that displays the data analyses and results of our projects, along with an add-on option for expert users to determine the fire probability at a specified location.
- **Datasets used:** Vegetation, Weather, Fire history, Terrain and Powerlines.
- **Drawbacks:** Requires high quality data which are not readily available,confined to our study area until further trained.
- **Challenges:** Complexity of parameters, lack of accessible and affordable data acquisition, pre-processing and modeling conundrums.



Lessons Learned

- High resolution consistent data is expensive, unless government sponsored.
- Ready availability of Satellite data in cloud catalogs.
- Usage of Python Application programming interface (API) and plugins for data extraction.
- Importance of feasibility check and null value imputation.
- Visualization for isolation of anomalies.
- Geographic and spatial data maps using QGIS and Tableau.
- Sensor data quality assurance procedure for mechanical failure, data transformation failure and collection failure.
- Hyperparameter tuning, data subsets and experimentation and myriad techniques for evaluation of your prediction results.
- Impressive capability of Folium, Tableau (Desktop and Public) and Flask for interactive maps, dashboards and user interface generation.

Recommendations

- Ideal successor to this undertaking is Fire behavior and spread studies
- Study and formulate the Fuel and fire dynamics, with the help of experts to integrate newer accurate indices in our modelling system.
- Inclusion of additional metrics, parameters and algorithms.
- Access commercial data sources with exceptional spatial and temporal resolution if funding is available.
- Consider parameters such as human activities, lightening, debris, campfire, smoking when data is published by concerned authorities.
- Increase the area and wider the scope to include management and alarm system.
- Data quality assurance standards and missing value imputation techniques can be juxtaposed and compared before modelling.
- Integrate real-time data collection to the User interface for a robust data extraction mechanism.

DEMO


Thank you!
SWiPS
Special thanks to Dr. Jerry Gao, Dr. Lee Chang, NASA, Cal Fire, Everbright technologies, among others.

Demo Screenshots

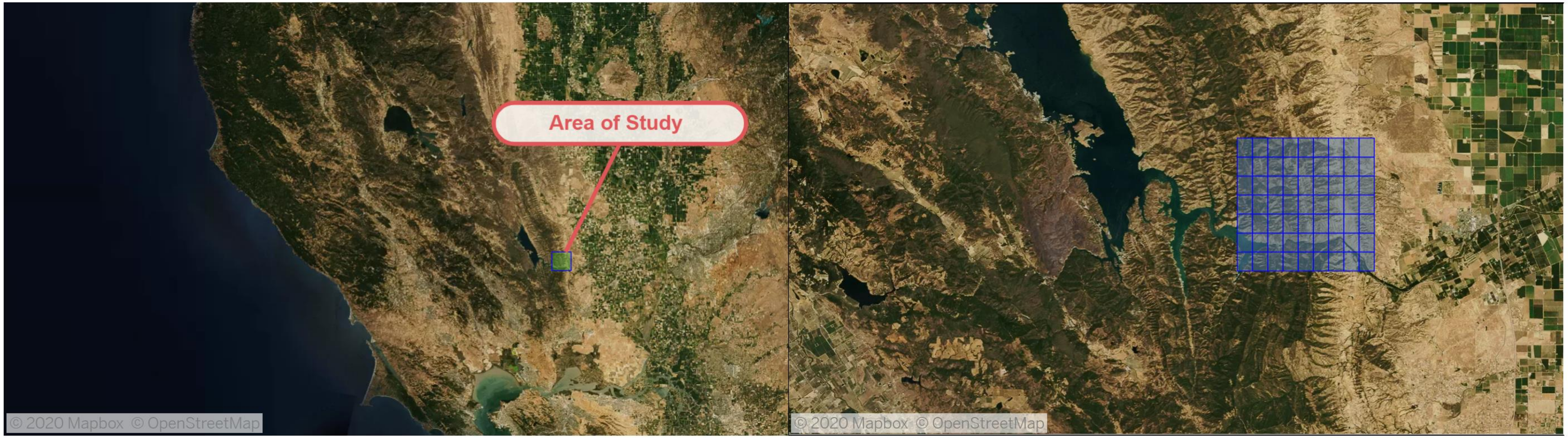
SWiPS

SPARTAN WILDFIRE RISK PREDICTION SYSTEM (SWIPS)

Our Study Area

Area of Study and the Grids

Select a View Satellite View ▾



Bounding box is the chosen fire-prone area near Winters, California, enveloped between Napa, Davis and Sacramento.

Gridded Area of Study has 7 rows and 9 columns, totalling 63 square grids of dimesion 1km * 1km. Here we are seeing the Satellite

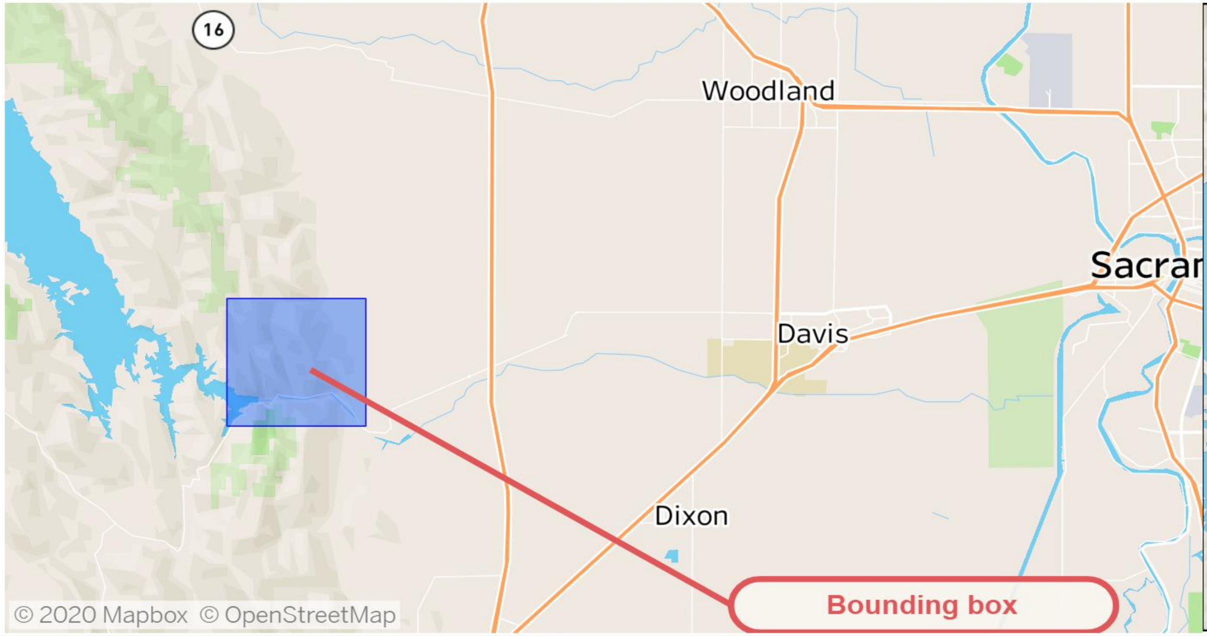
Our Study Area

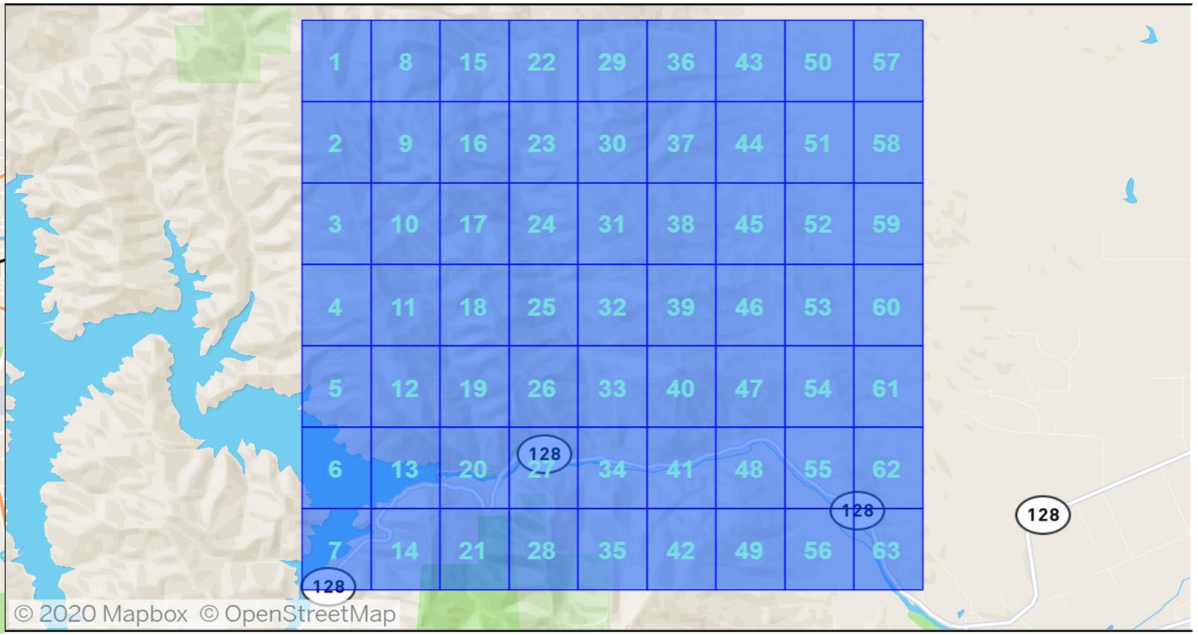
Area of Study and the Grids

Select a View: Street View

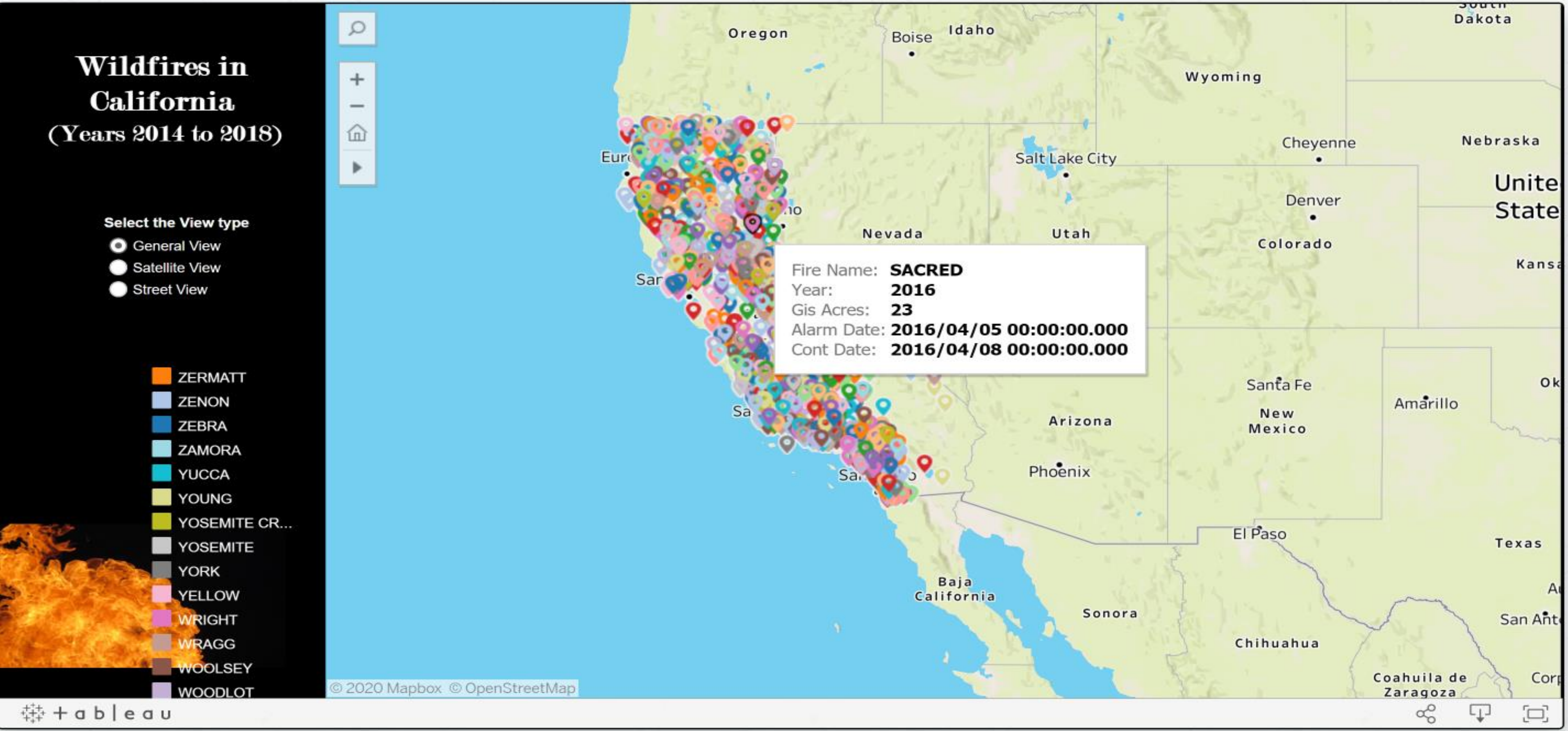
Bounding box is the chosen fire-prone area near Winters, California, enveloped between Napa, Davis and Sacramento.

Gridded Area of Study has 7 rows and 9 columns, totalling 63 square grids of dimesion 1km * 1km. Here we are seeing the Street View of the numbered Grids.





Wildfires in California

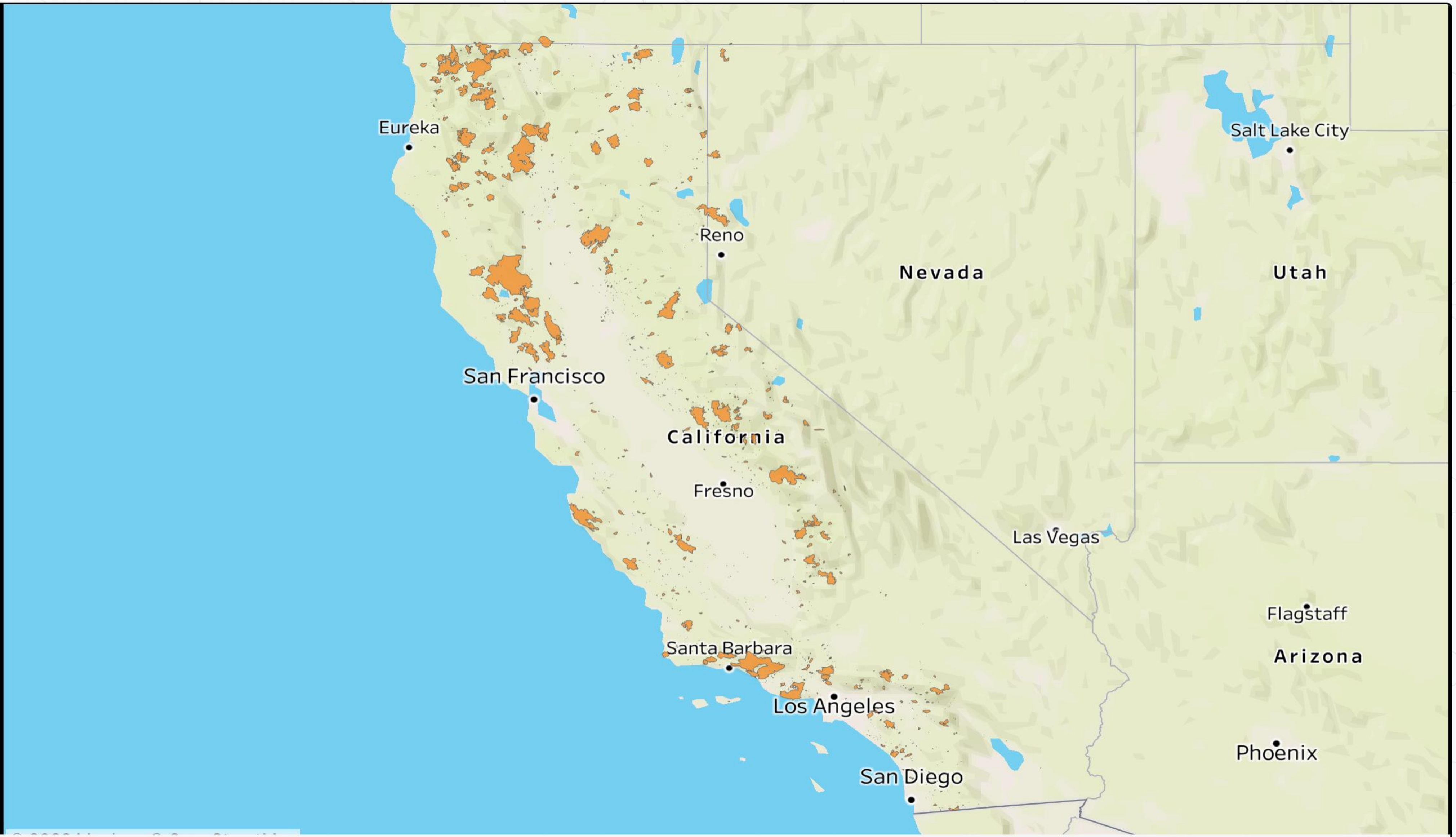


Wildfires in California

Wildfires in California
(Years 2014 to 2018)

Select the View type

- General View
- Satellite View
- Street View



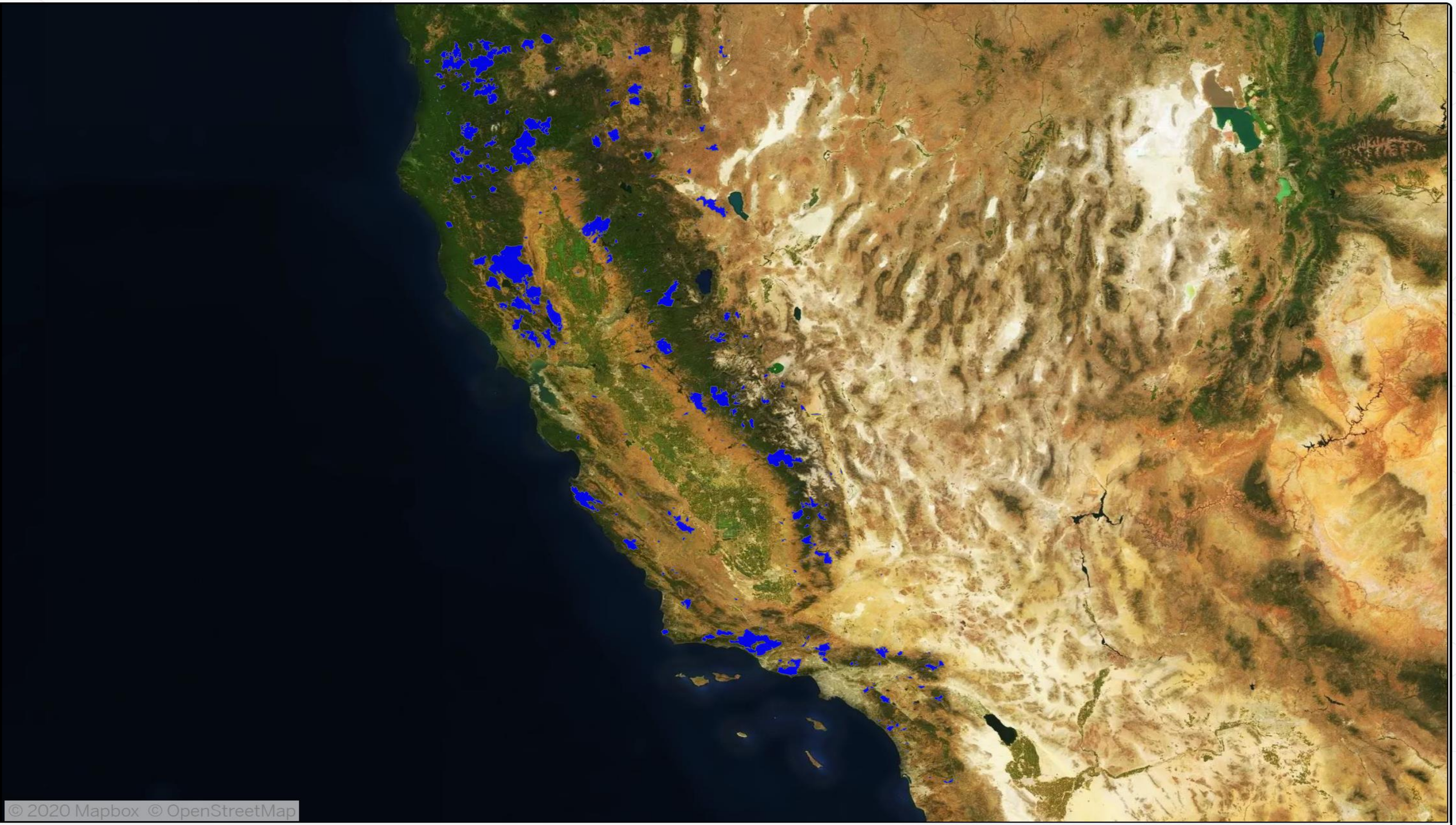
tableau

Wildfires in California

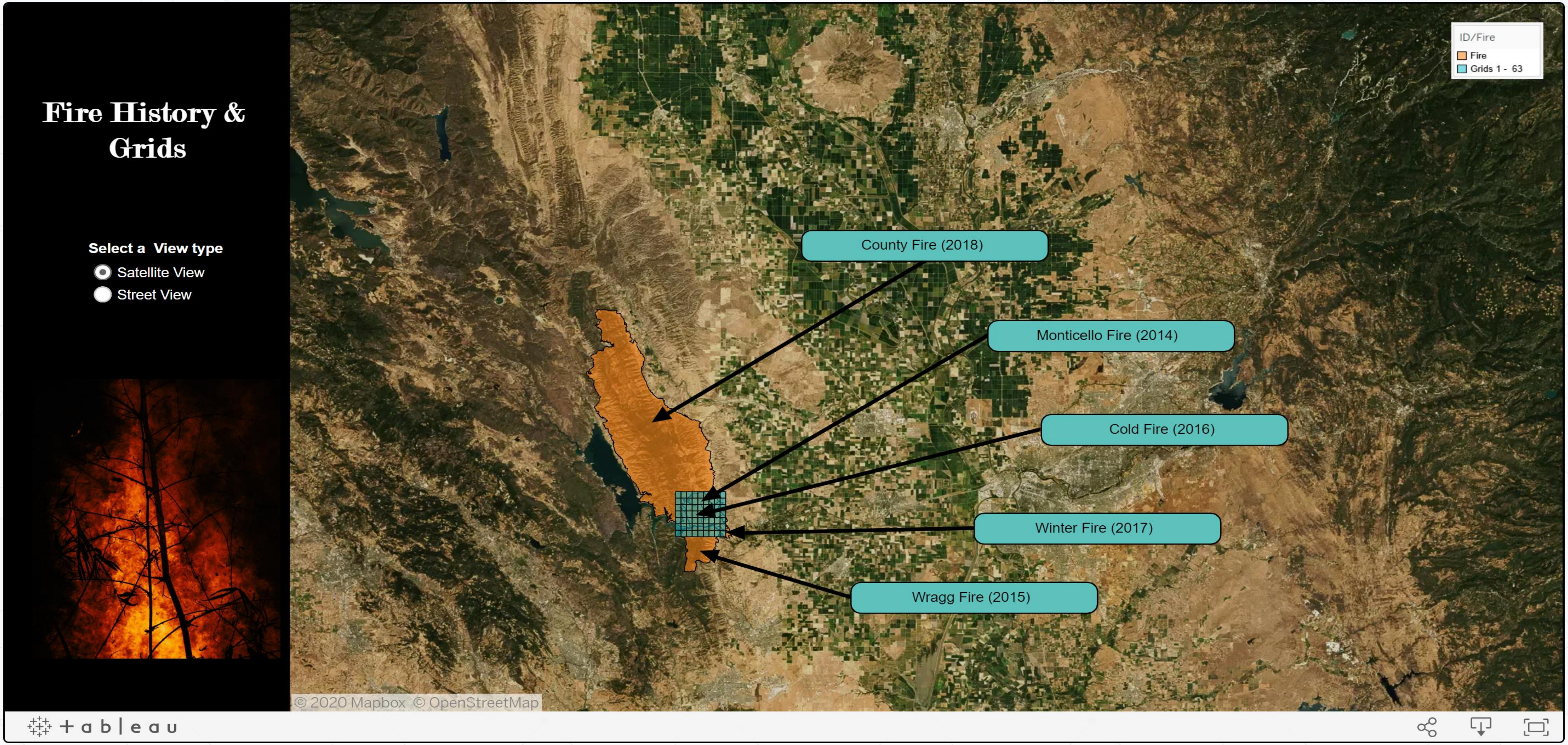
Wildfires in California
(Years 2014 to 2018)

Select the View type

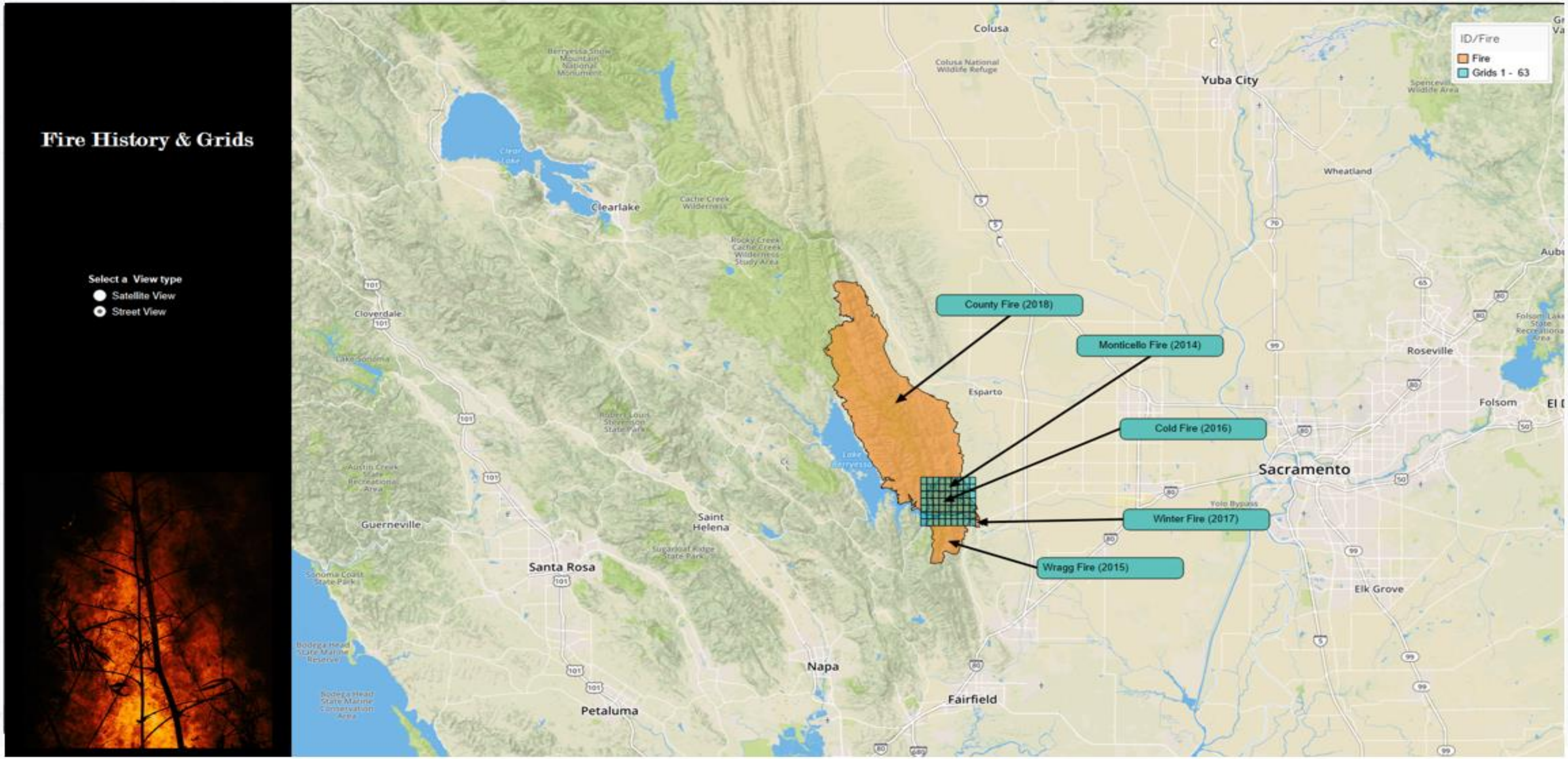
- General View
- Satellite View
- Street View



Fire History & Grids in the Study area



Fire History & Grids in the Study area

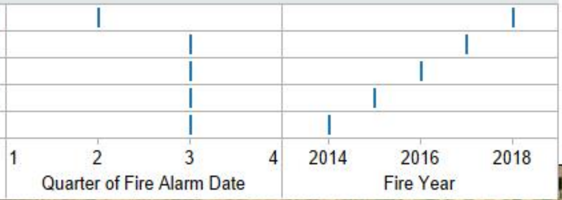


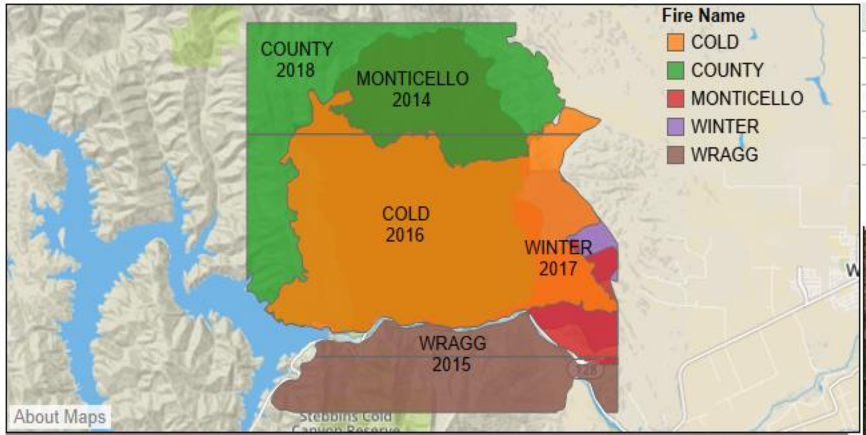
Fire history analysis

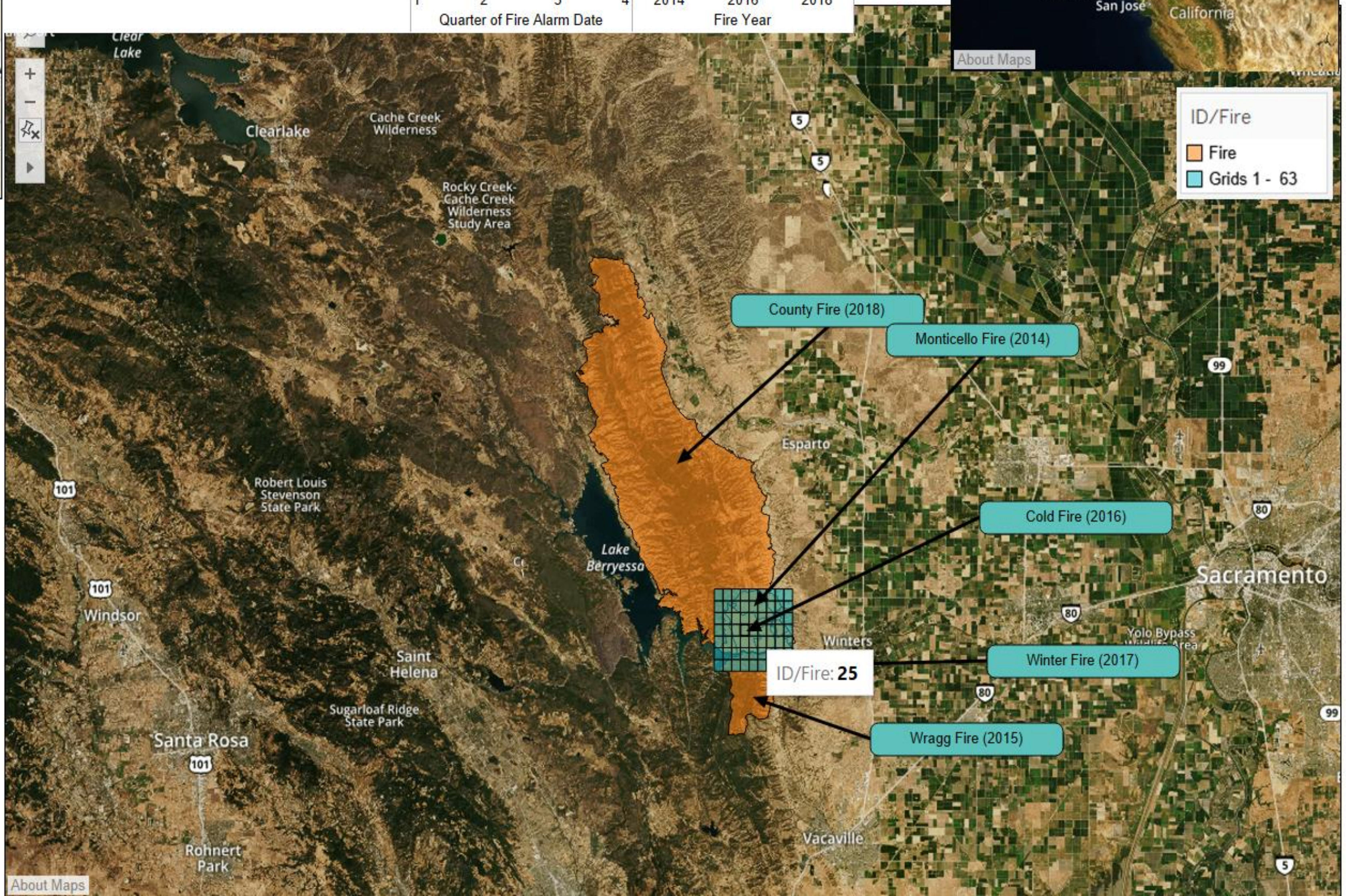
Fire Name

(All)

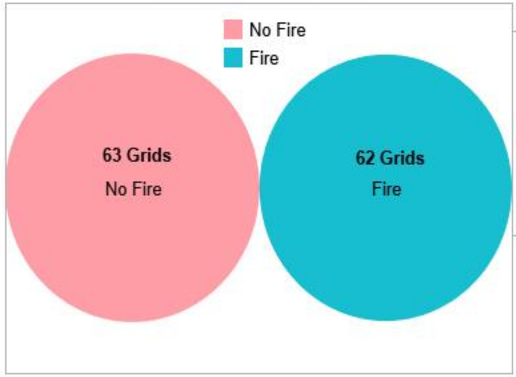
Fire Name	State	Gis Acr..	Start ..	Contained ..
COUNTY	CA	89,831.15	2018-06-30	2018-07-11
WINTER	CA	2,366.23	2017-07-06	2017-07-12
COLD	CA	5,730.07	2016-08-02	2016-08-09
WRAGG	CA	8,049.33	2015-07-22	2015-08-05
MONTICELLO	CA	6,599.39	2014-07-04	2014-07-12

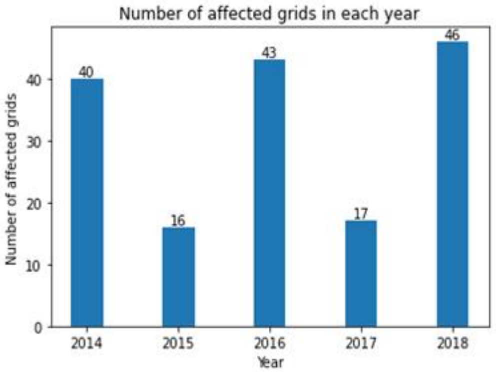




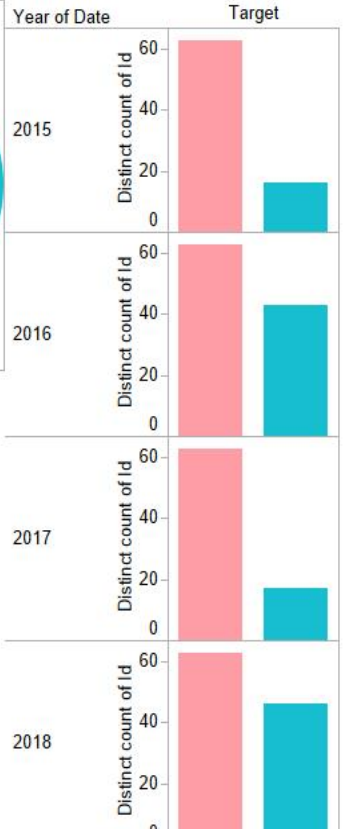


Grid-wise Statistics



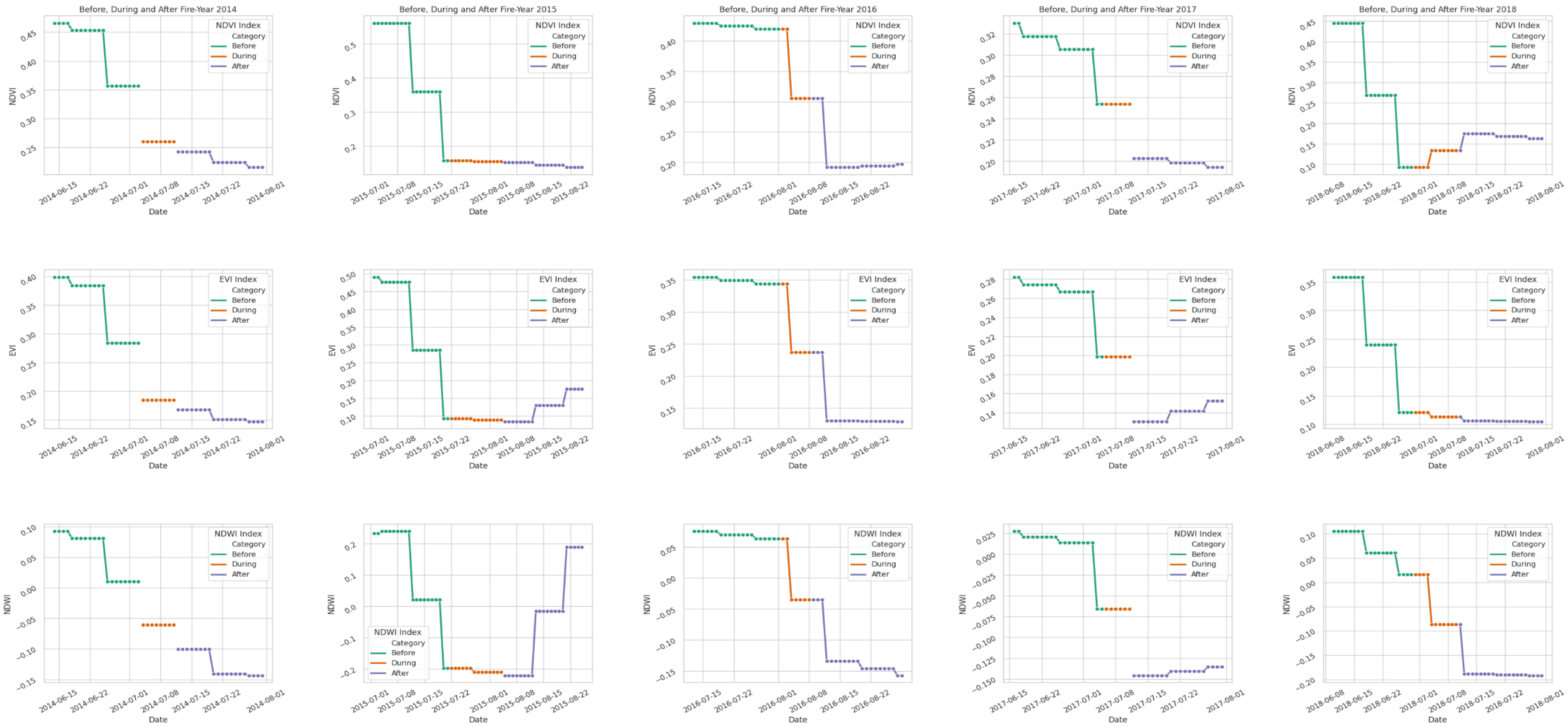


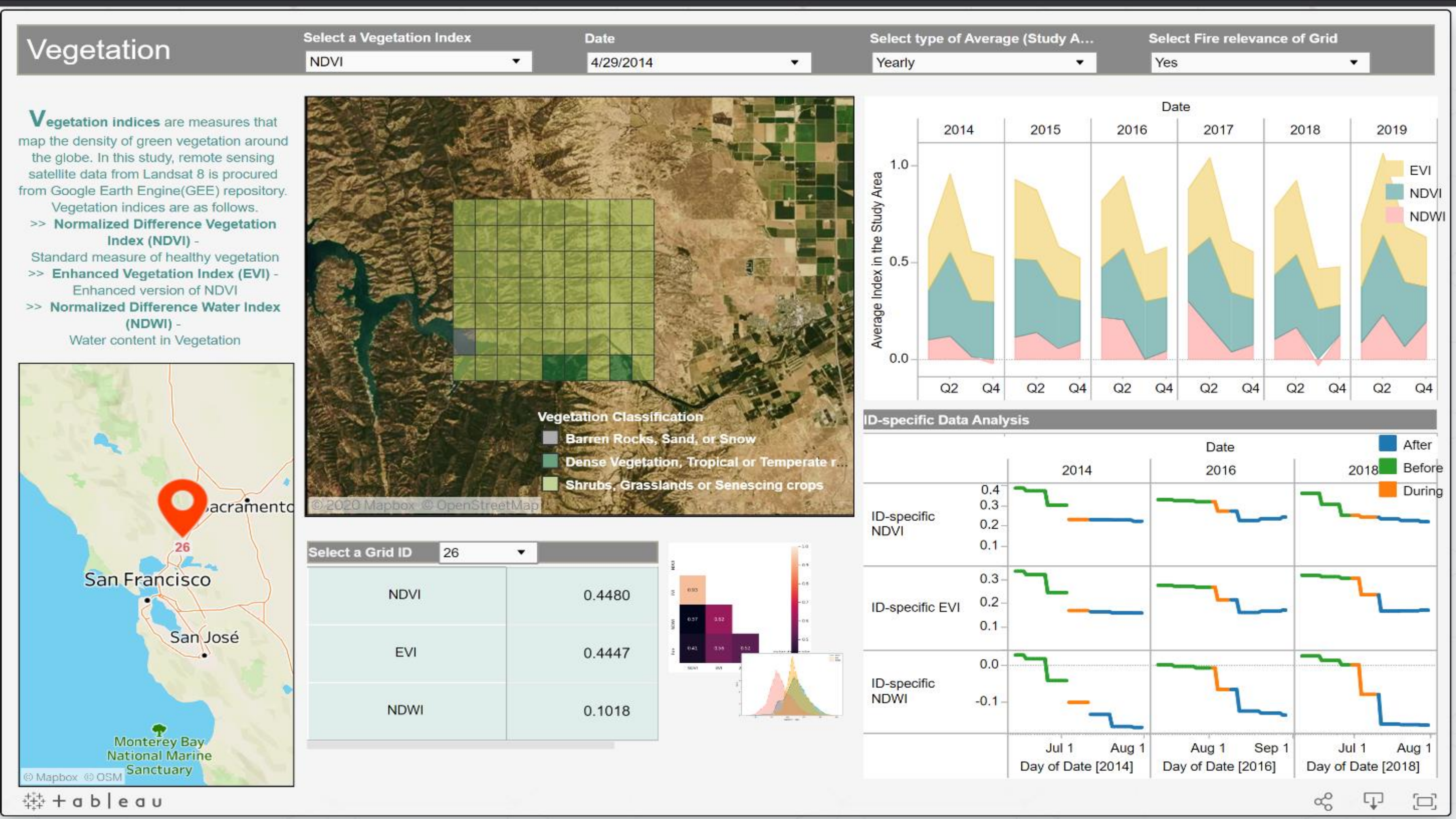
Yearly Statistics

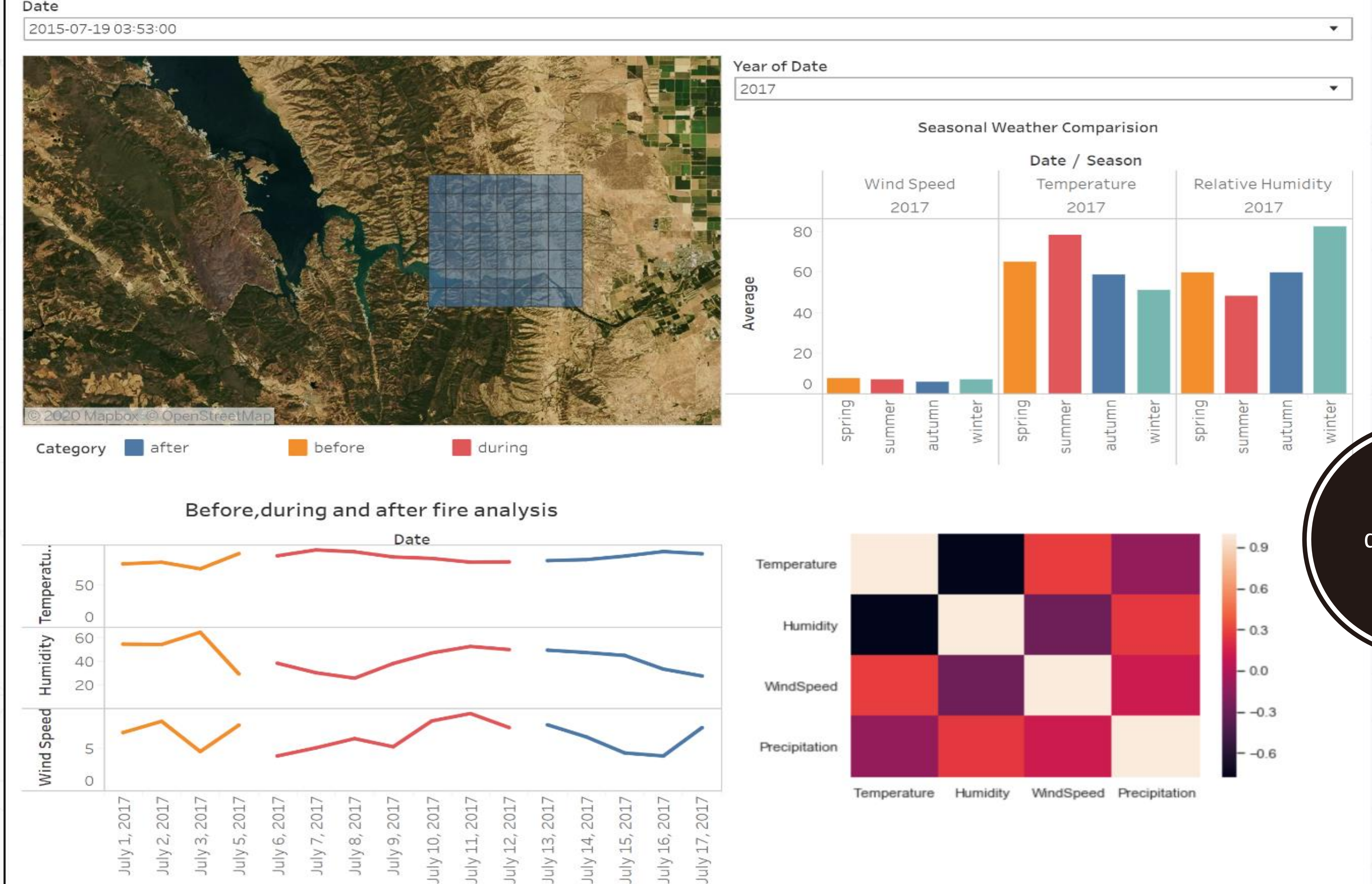


Yearly pre-fire, during and post-fire vegetation indices analysis

Grid-wise pre-fire, during and post-fire analysis for Vegetation data (2014 to 2018)




Vegetation
Select a Vegetation Index
NDVI
Date
4/29/2014
Select type of Average (Study A...
Yearly
Select Fire relevance of Grid
Yes
Vegetation indices are measures that map the density of green vegetation around the globe. In this study, remote sensing satellite data from Landsat 8 is procured from Google Earth Engine(GEE) repository. Vegetation indices are as follows.
>> Normalized Difference Vegetation Index (NDVI) - Standard measure of healthy vegetation
>> Enhanced Vegetation Index (EVI) - Enhanced version of NDVI
>> Normalized Difference Water Index (NDWI) - Water content in Vegetation
Vegetation Classification
Barren Rocks, Sand, or Snow
Dense Vegetation, Tropical or Temperate r...
Shrubs, Grasslands or Senescing crops
© 2020 Mapbox © OpenStreetMap
Sacramento
26
San Francisco
San José
Monterey Bay National Marine Sanctuary
© Mapbox © OSM
Select a Grid ID
26
NDVI 0.4480
EVI 0.4447
NDWI 0.1018
Date
2014
2015
2016
2017
2018
2019
Average Index in the Study Area
1.0
0.5
0.0
EVI
NDVI
NDWI
Q2
Q4
ID-specific Data Analysis
After
Before
During
ID-specific NDVI
ID-specific EVI
ID-specific NDWI
Day of Date [2014]
Day of Date [2016]
Day of Date [2018]
Jul 1
Aug 1
Sep 1
tableau



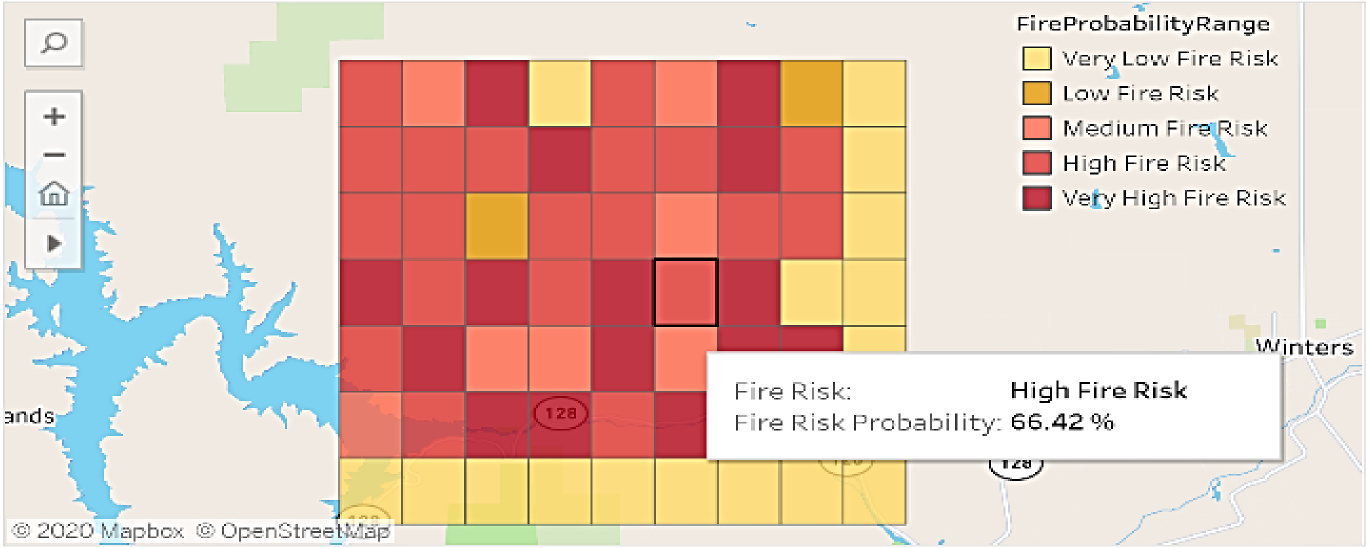
Weather dashboard here

Date

6/30/2018 1:53:00 PM

Fire Risk Prediction on 6/30/2018 1:53:00 PM

Fire Risk



Powerlines Data for Grid - All

Circuit -	Single
Status -	Operational
Voltage (KV) -	115
Length (Miles) -	25

Weather Data for Grid - All

Relative Humidity (%) -	14.0
Temperature (°F) -	103.0
Wind Speed (km/hr) -	14.0
Precipitation (mm) -	0.0

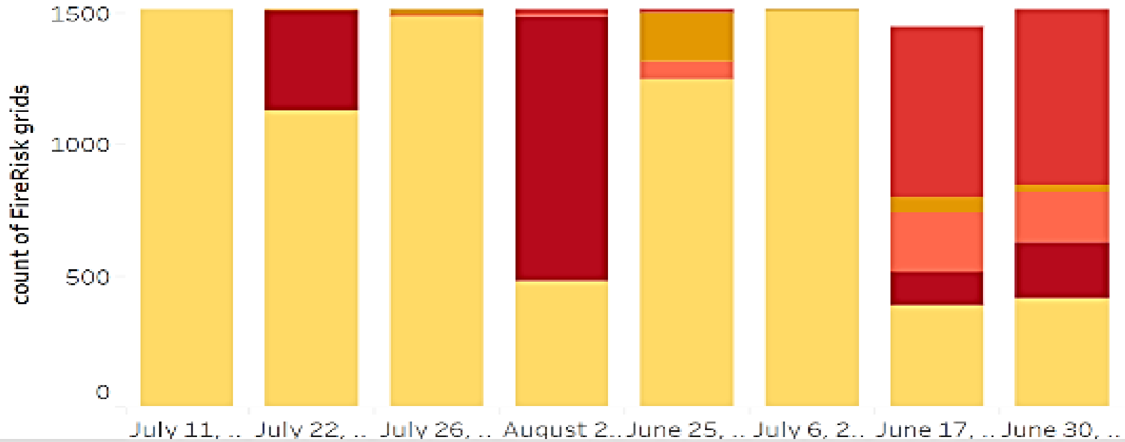
Terrain Data for Grid - All

Hillshade direction -	North
Slope direction -	Moderate Slope
Aspect -	South West

History of Fires

Fires - 46

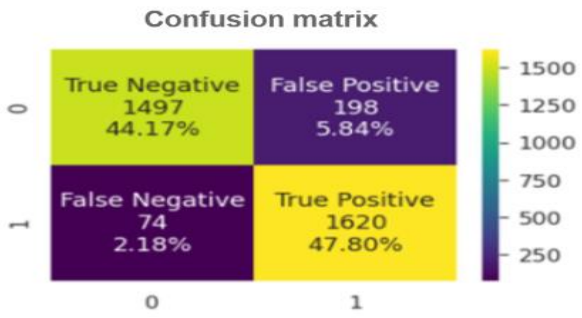
Fire Risk Analysis for grid - All



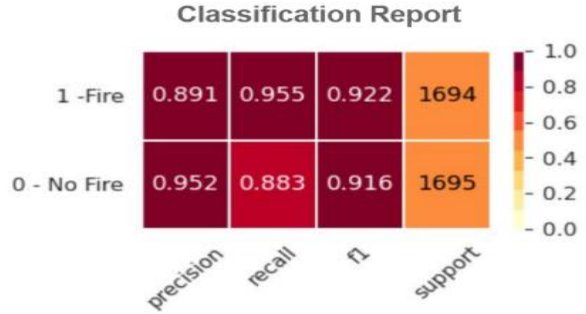
Vegetation Data for Grid - All

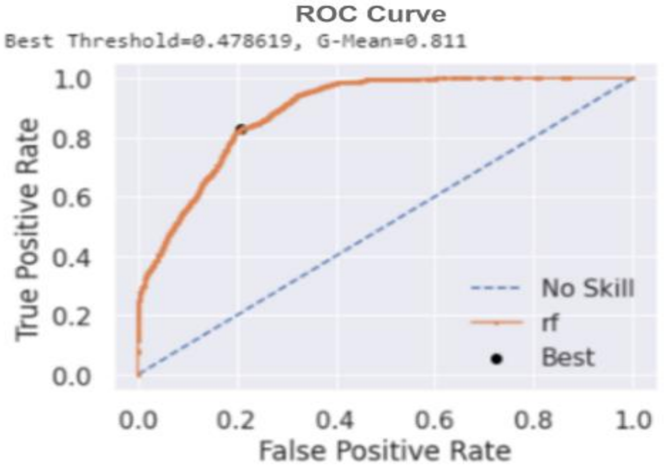
Vegetation type(NDVI) -	Shrubs and grasslands
Veg Water Content(NDWI) -	Medium (Moist) water content

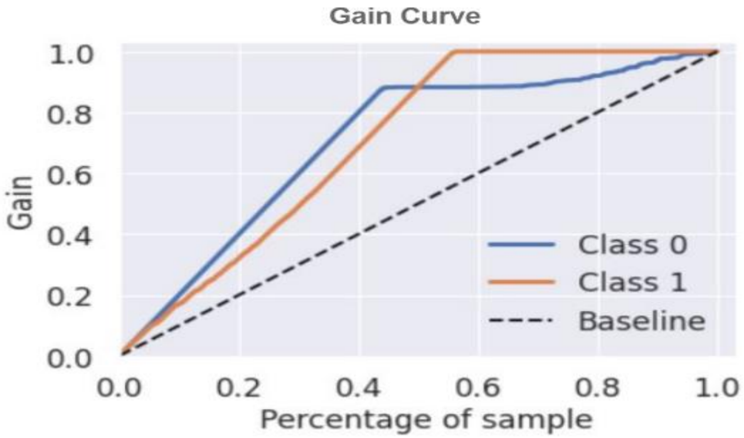
Final Classification Model (Random Forest) Evaluation Metrics

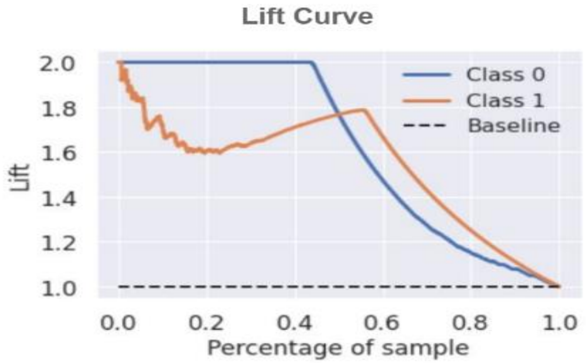


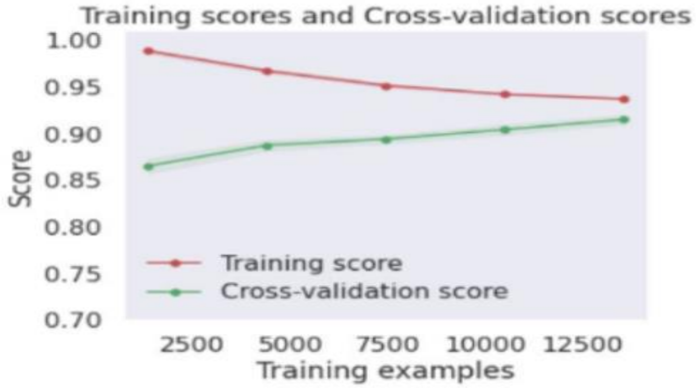
Confusion matrix is a tabular representation of the performance of classification models by analyzing the model based on the number of true and false predictions.
Classification Report is based on confusion matrix numbers and shows the percentage for Precision, recall and F1 Score for all classes.
ROC Curve plots False positive rates vs True positive rates. Area under the curve is a true mark of model efficiency and its ability to classify.
Gain curve evaluates the performance of the best model against a model highly random in feature selection.
Lift curve evaluates the likelihood of target or prediction probability of the model against a random data used as a baseline model.
Learning curves include 3 graphs which display the comparison between training and cross validation scores, scalability and performance of model. They are helpful in evaluating the feasibility of the model.

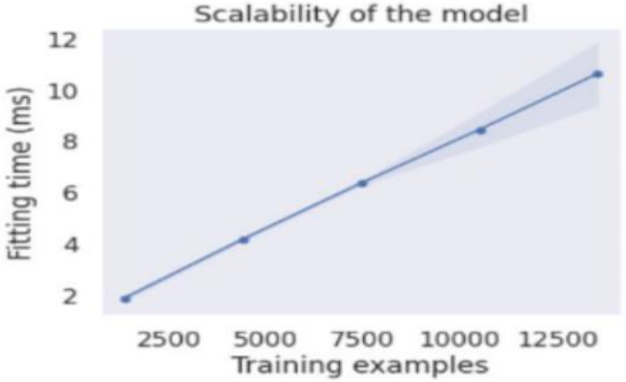


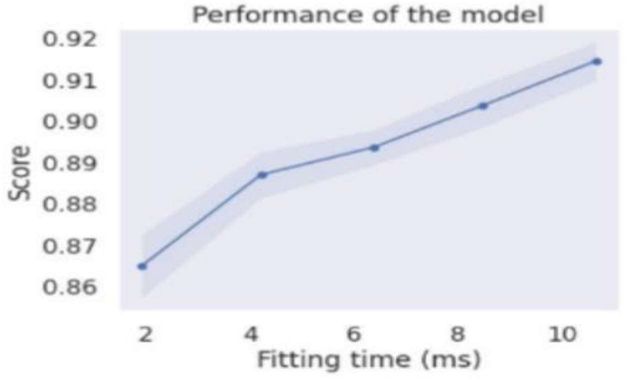













Thank you
again!!