California Wildfire Events Prediciton

Jiaying Li 12/16/2019

Introduction

Every year, the California Wildfire is one of the biggest issues that the state is suffering from. It could help to prevent the expansion of wildfire if the event can be predicted in advance. Weather condition is one of the well-acknowledged variable that people believe which correlates with the wildfire events, and our team has received a request to understand how can the weather implement to help understand the wildfire, and as the weather prediction are becoming more accurate in the future, we have made an hypothesis on whether the weather data can make predictions on the wildfire event in California and the performance of the model. The datasets, including data of California weather stations and California recorded wildfire events, are obtained from a research team, and the goal for the project is to apply exploratory and predictive data analysis approach to understand how the weather data can help to predict the California Wildfires. Spatial analysis, data visualization, linear modeling and machine learning are implemented in this study.

Dataset

Section 1. Wildfire Dataset

The wildfire dataset is obtained from the fire department with 40 variables and 189551 observations, which includes recorded wildfire events from 1992 to 2015. Variables included the reporting code, agency, burn time, date and cause of the fire event The dataset received contained many variables with missing values, so we conducted data cleaning processes to ensure that the further analysis can be performed smoothly. We keep the wildfire happened after (including) 2000, as the other accessible datasets for data analysis only have observations after 2000, which reduced the total amount of observations to 121535. The number of observations for data analysis will be discussed in detail in each section based on the analysis need. Table 1 described the core variables that are used in this project.

Table 1. Description of the variables for the cleaned Wildfire Dataset

Discovery date	Date on which the fire was discovered or confirmed to exist.
Burn time	Burn time of the wildfire event
Fire cause description	Description of the (statistical) cause of the fire.
Latitude	Latitude (NAD83) for point location of the fire (decimal degrees).
Longitude	Longitude (NAD83) for point location of the fire (decimal degrees).
Fire size	Estimate of acres within the final perimeter of the fire
Fire size class	Code for fire size based on the number of acres within the final fire perimeter expenditures (A=greater than 0 but less than or equal to 0.25 acres, B=0.26-9.9 acres, C=10.0-99.9 acres, D=100-299 acres, E=300 to 999 acres, F=1000 to 4999 acres, and G=5000+ acres).
FIPS name	County name from the FIPS publication 6-4 for representation of counties and

	equivalent entities.

Section 2. Weather Station Datasets

To analyze the correlation of weather condition and the wildfire occurrence, 62 weather stations' data are retrieved from ASOS online system (https://mesonet.agron.iastate.edu/request/download.phtml), with 29 variables included in the initial dataset. Some of the variables are re-sampled by the data cleansing team, which turned the sub-hourly readings into max, min and average value for the same variables in each observation with interval of one hour. The dataset is retrieved for the dates between 2000 to 2015; however, many datasets are incomplete with many missing observations. Some of the weather stations only have 2-3 years of weather data, and these will not be considered comprehensive enough to support our analysis process. Further selections of weather stations are described in the data-preprocessing section based on the need of the analysis. Table 2 describes the core hourly recorded variables that we were interested in joining with the wildfire event in the analysis section. Geometry data with longitude and latitude of weather stations are stored in a separated dataset with cleaned information of the 62 stations.

Table 2. Description of the cleaned weather station dataset

Temperature	Air Temperature in Fahrenheit, typically @ 2 meters
Dew point temperature	Dew Point Temperature in Fahrenheit, typically @ 2 meters
Relative humidity	Relative Humidity in %
Wind speed	Wind speed
Wind direction	Wind Direction in degrees from north
Pressure Altimeter	Pressure altimeter in inches
Precipitation	One hour precipitation for the period from the observation time to the time of the previous hourly precipitation reset.
Timestamp	Time of the recorded observation with hours and dates

The detail of data preprocessing before the Predictive Data Analysis are all included in the Exploratory Data Analysis in in the following section.

Section 3. Population Density Dataset

The California population density dataset is also provided by the team for data analysis, and the dataset is cleaned with only the county name and population density included. Not further correction is needed. 58 counties' population density values are included.

Analysis

Section 1. Exploratory Data Analysis (EDA) – Complete Dataset for Wildfires

The total number of fire observations is large, and we first plotted Figure 1 for the fire events in 2010 to see whether there is a spatial pattern for the occurrence of fire. We see that there are certain regions, such as the inner California and the south California has tremendous red spots, indicating the fires events are having high density in these regions. The 62 weather stations are plotted in the figure as well, and many weather stations are close to each other and located at the center of the wildfire clusters. These patterns indicate that our data analysis approach can be conducted by matching fire event and the weather condition that retrieved from close weather stations according to the date stamps.

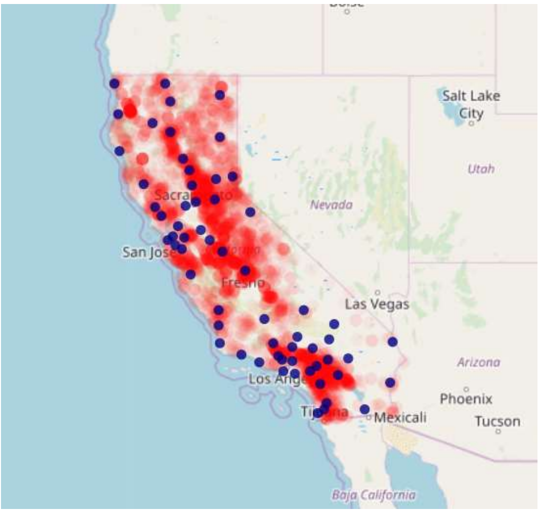

Figure 1. Plot of Wildfires in 2010. Clear patterns can be seen from the figure, spatial clustering is transparent

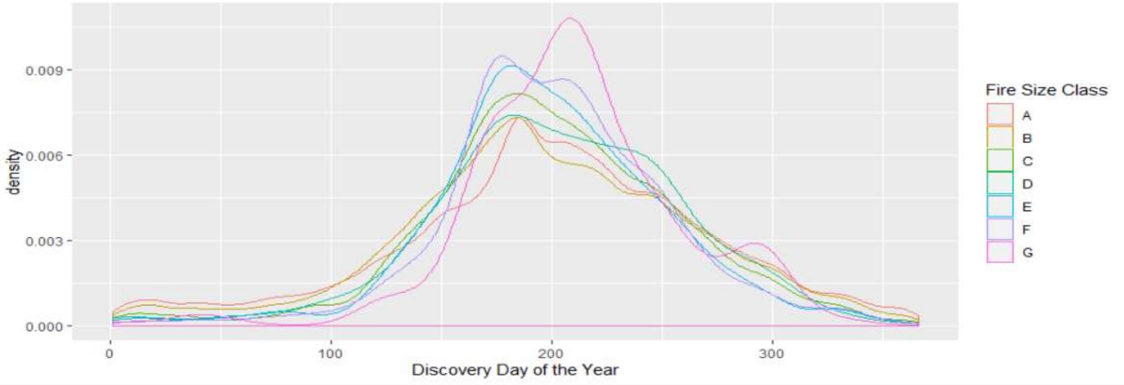

Figure 2. Density of fire frequency by fire size class; mostly clustered during the summer of the year, but class G shows special patters with a small peak in October.

Since one of our goal is to find a method to predict wildfire event with severe impact, we studied on the density of the wildfire frequency in a year based on the fire size class, as shown in Figure 2. The density curve is close to a bell shape, but there are some discrepancies between each curve, and the curve for class G shows interesting small peak around 300 of the discovery days of the year.

The average burn time and frequency of the wildfires are calculated and visualized in Figure 3a and 3b. It seems that for the place with higher number of frequency of wildfire, the average burn time is low, which indicates that many small but non-significant fire happened in these regions.

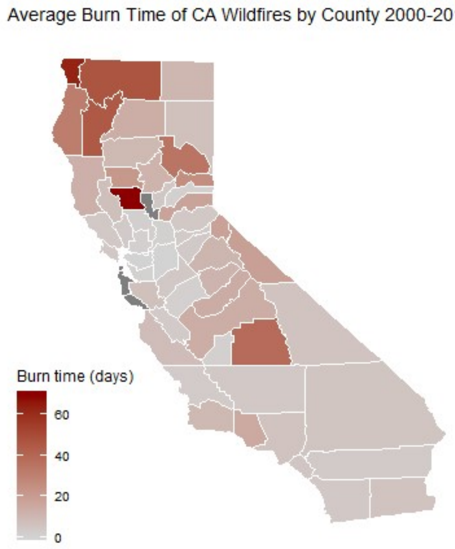

Figure 3a. Visualize the average of burn time of California Wildfires between 2000 and 2015. Two county does not have burn time data, and North Cal have higher average burn time compares to South Cal, with only one outstanding county showing average burning time higher than 40 days

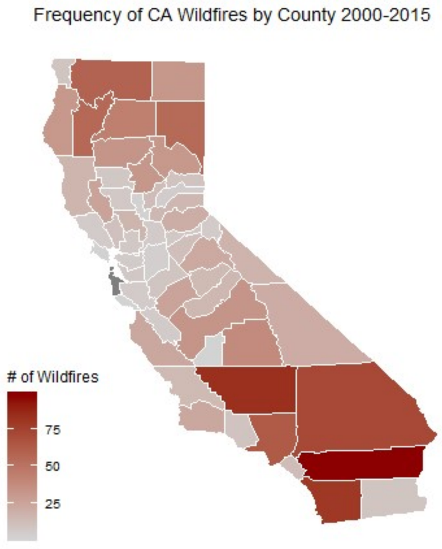

Figure 3b. Frequency of wildfires by county in California between 2000 and 2015. One county in the Bay Area does not have wildfire, and significant amount of wildfire events occurred in South California

Since the dataset is too large for us to make further analysis, we describe our approach regarding how we narrow down the dataset and the study.

Section 2. Data Preprocessing - Prior to Exploratory Data Analysis in detail and Predictive Data Analysis with Machine Learning

Stage 1. Weather Station Selection

As we think that the dataset of wildfire is too large with many unimportant small fires, and not all fires should be worried by the emergency or fire department as most of them barely influence the environment, we will narrow down the dataset in the following analysis on fire size class including D, E, F and G. The total number of observations at this point reduced to 2685, which is more reasonable for local CPU to conduct the predictive data analysis.

As we want to find a way to classify whether there will be a wildfire under certain weather condition or not, the wildfire event need to match the weather condition that can best represent the actual condition on that day. Since many weather stations are clustered in regions with similar weather pattern, we will select some of the weather stations to match its data with the wildfire on that day. In addition, wildfires with fire size class over D show a spatial clustering pattern as well, which can be seen from Figure 4a.

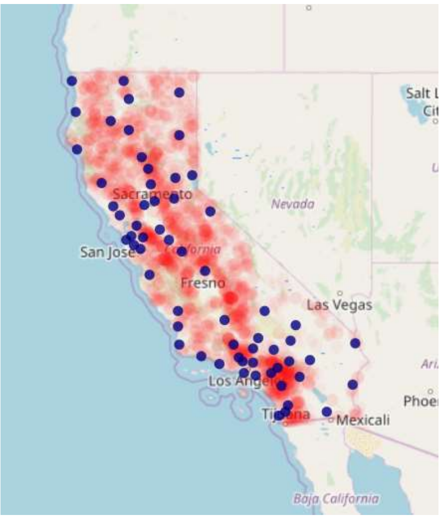

Figure 4a. Clustered wildfire events with fire size class level D, E, F and G and 62 weather stations locations. Several clusters can be seen, and each cluster have many weather stations in the region where the climate might be similar.

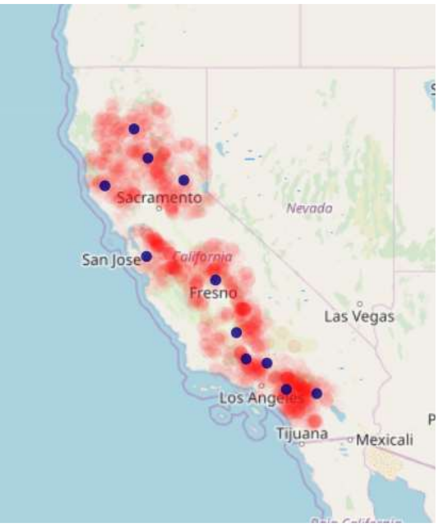

Figure 4b. Selected 11 weather stations and their coverage of wildfires in 100 km. The weather stations covered significant number of wildfires comparing Figure 4a.

11 weather stations are selected at this stage based on visualizing the regions with the most concentrated amount of wildfire events between fire size class D to fire size class G. The weather stations selected are: SDB, PMD, RAL, PSP, BFL, CIC, RDD, UKI, BLU, SJC, and FAT. The full name of weather station can be found in Table A.1 in the Appendix. Instead of selecting the weather stations only by looking into the location and coverage of wildfires, these weather stations should also have comprehensive and almost full coverage of hourly data between 2010 and 2015, which are 140256 hours in total. This constraint in selection ensures that all fire events can match to a weather condition successfully.

In order to match the weather condition (with hourly observations) with the wildfire events (with daily observations), the data for each selected weather station are processed from hourly data to daily data by calculating the average value of each variable in a day, with all the not available data points omitted during the calculation. The total number of observations in each weather station has been reduced to 5844, and only averaged hourly data are converted into daily average data, which will be used in Predictive Data Analysis.

Stage 2. Distance Calculation and Filtering Process of Observations for Analysis

The distance between each of the wildfire events over class C and the 11 weather stations selected has been calculated, and the shortest distance has been chosen for the weather data to be assigned to. To ensure that the data of the weather station can represent the actual condition on that day, only wildfire within 100km are kept for the next step of data analysis. The total number of wildfires events kept at this point is 1848. Longitude and latitude for both weather station dataset and wildfire dataset are used, and distances are calculated based on projection to EPSG 4326 (long-lat). Figure 4b are the wildfire events that have distance less than 100 km, where the distance is calculated by the method discussed here.

Stage 3. Finalizing dataset as input

Due to the characteristic of the machine learning model, the prediction cannot be well performed if the ratio of the observations in two different classes for classification are extremely biased, as what we see from our dataset. Therefore, data preprocessing is conducted to fix this problem. The amount of observations with the exist of fire event is maintained to be the same value in the final dataset prepared for machine learning, while the dataset with the no-fire existed observations are reduced to the same amount of the dataset with wildfire. The ratio is 1:1 for the 0 (no fire) and 1 (with fire) datasets, and the observations for the 0 (no fire) labels are randomly selected from the overall observations from the 11 weather stations between 2000 - 2015. Therefore, the total number of observations for machine learning is 1848*2 = 3796, and by omitting the observations with variable values as not available, the final number of observations becomes 3432.

In addition, the dataset has been separated by ratio of 8:2 for the train and the test dataset, and by 9:1 within the train dataset to fit and validation sets. The fit and validation datasets will be used for initial model selection, and the train and test datasets will be used for final performance check of the machine learning outcome in wildfire prediction.

Section 3. Confirmatory Data Analysis (CDA)

Before we move on to study the correlation of weather condition and the occurrence of wildfire, we would like to justify our thoughts of the correlation between burn time and wildfire frequency as we mentioned in Section 1. As we also have the population data for each county, we would like to see if there is a correlation between them as well. Figure 5 shows the plot of these variables which are matched by county name.

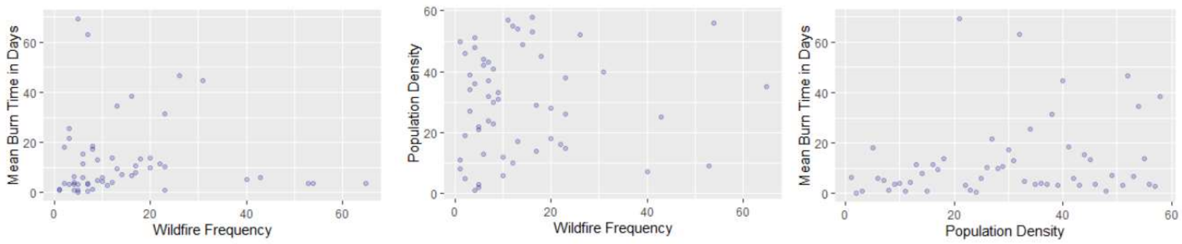

Figure 5. Plot of different combination of variables. Very limited correlation can be seen from the figure.

Instead of simple visualization, we performed a confirmatory data analysis by creating a linear model with burn time as the output and population density and the wildfire frequency as the input of the model. The significance level for population density is relatively high, with p-value less than 5%, but the R-squared is only 0.11 for the model. Therefore, we believe that there are small correlation between the variables, but it is not strong enough with high significant level.

Section 4. Exploratory Data Analysis – Selected Wildfires and Weather Stations

Before the Predictive Data Analysis, we explored the dataset again and found some interesting features. All dataset used in this section are preprocessed, where the total of observations for fire event is 1848, and the total of observations for no-fire event recorded is 1848 as well.

Exploratory Data Analysis on Wildfire Dataset

We analyzed the distribution of wildfires in D to G fire size class in each month and causes of the events based on the filtered data, with a range of 100km from the selected weather stations. The total number of wildfires analyzed is 1848. The distribution of wildfire event is in bell shapes as shown in Figure 6a. In addition, we see that causes of wildfires are close to evenly distributed through the year, with more lightning caused fires between 06-10 compared to the rest of the year. Equipment Use, lightning and arson are top causes of wildfire excepting the unidentified and less cleared sources.

The frequency of causes of the wildfires by fire size class is plotted as well, which is shown in Figure 6b. The frequency of G class fire caused by lightning is highest amount all the categories, and the second highest frequency of G class fire falls into the category of undefined miscellaneous. If we look at the second severity level of fire size, F Class, it also has its highest frequency in the category that belongs to cause of lightning.

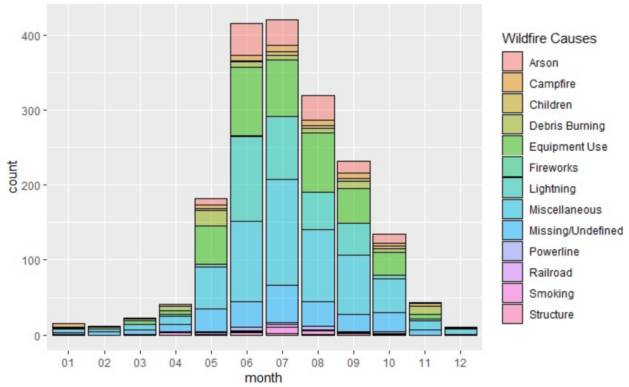

Figure 6a. Distribution of wildfire events on months between 2000 and 2015 with the separation on causes of wildfire. Wildfires are concentrated during the mid of the year and left skewed with higher frequency by the end of the year compared to the start of the year.

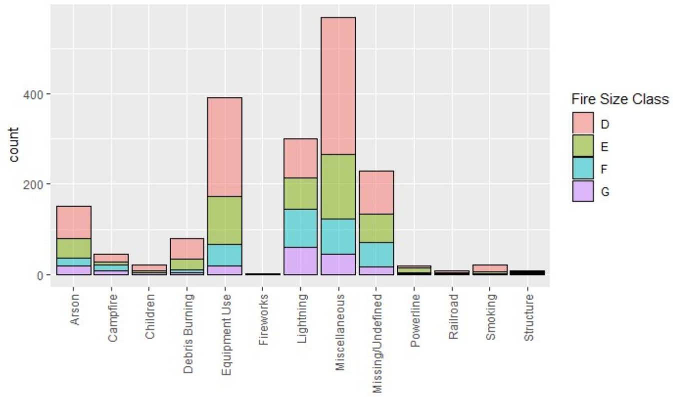

Figure 6b. Frequency of Causes of Fire– Wildfire data filtered by retaining fire size class over D. Since the Miscellaneous and Missing/Undefined are with limited information on identifying the reason of wildfire, we will focus on the causes with the top three frequency excluding these two categories

In the Predictive Data Analysis section, we will look more into the accuracy of prediction of the wildfire events by causes, which will help to prevent future fire with the specific cause.

EDA on Weather by the Appearance of Wildfire Events (X0/0 = no wildfire, X1/1 = wildfire)

To investigate further on whether the weather data can be applied to train and predict the wildfire by using machine learning models, we analyzed the distribution of weather variables by the date with fire and the date without fire.

Figure 7a and 7b shows the range of data under different categories. The dataset analyzed in these three figures are based on 3432 data points with 1:1 ratio of dates with wildfire (X1) and date without wildfire (X0). Two most representative variables with significant difference in the averaged value of weather condition between X1 and X0 are selected, which are temperature and pressure altitude. We see that the average value for each category have very clear descrepancies, so we move on to visualize if there is a clustering pattern for the weather condition under the X1 and X0 categories.

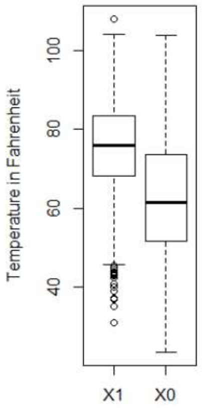

Figure 7a. Distribution of temperature by X1 and X0, X1 has higher temperature average compares to X0.

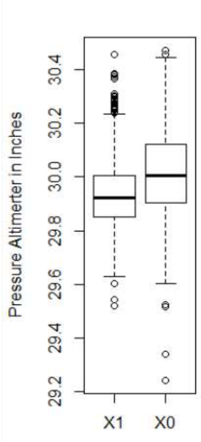

Figure 7b. Distribution of Pressure Altimeter by X1 and X0, X0 has higher average value compares to X1

Several representative figures that shows the displacement of clusters grouped by observations with wildfire event discovered (X1) and without wildfire events discovered (X0) are presented in the set of figures as below:

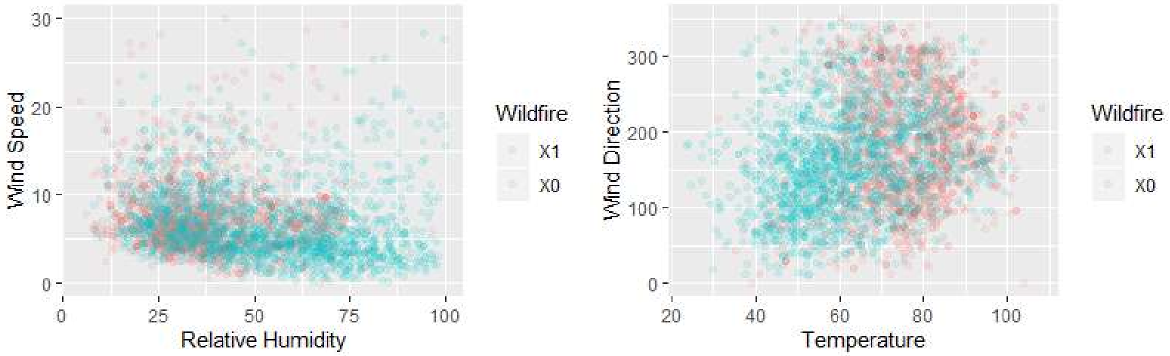

Figure 8 (a-b) Clustering of Obsevations by the Appearance of Wildfire; (a) plot with relative humidity and the wind speed; (b) plot with temperature and the wind direction; red(X1=with wildfire) and green(X0 = no wildfire) points are clustered differently on each of the two dimensions. Units listed in Table 2.

In addition, linear correlation between variables of weather data is observed during the EDA process as well, so we will look at the importance of the variables used in the prediction in the next section. Several representative plots are selected and presented in the figure below:

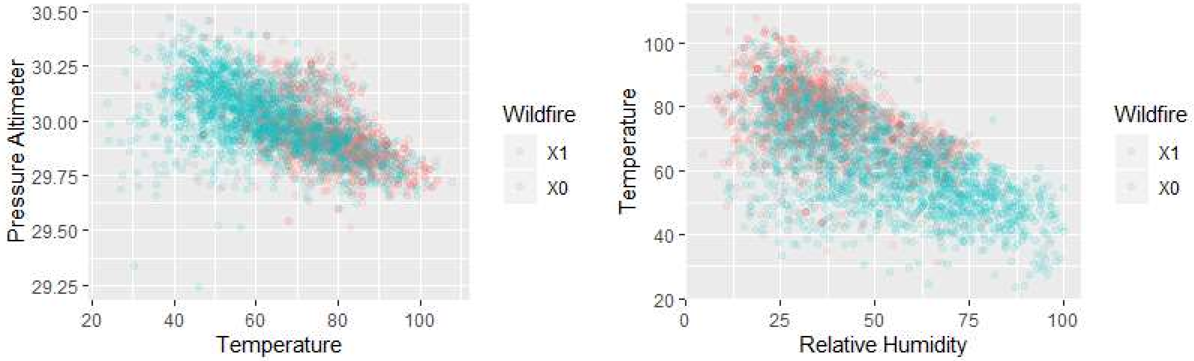

Figure 9 (a-b) Clustering of Obsevations with Linear Correlation Between Variables by the Appearance of Wildfire; (a) plot with relative humidity and the wind speed; (b) plot with temperature and the wind direction; red(X1=with wildfire) and green(X0 = no wildfire) points are clustered differently on each of the two dimensions, and also linear correlation between different vairbales can be seen from the figure. Units listed in Table 2.

Section 5. Predictive Data Analysis (PDA)

In this section, we will study the capability of classifying whether the there will be a wildfire event on a specific date based on the weather data. The data on the same day is used for the prediction as we assume that the weather prediction is accurate. This PDA section will be separated into two parts. One part is using the dataset for 11 weather stations in predicting the California's wildfire. Another part is using the subset of data for three randomly selected weather station in testing the performance of prediction on a specific geographical region, as the local geographical condition for each station could be significantly varied. The three weather stations selected are: RAL, FAT, UKI. The test on classification performance will be built on the final model chosen from the first part of the PDA.

PDA with Dataset of Wildfire Class D, E, F, G

We first conducted PDA for the whole dataset with fire size class over C. Four models are selected to test the capability of classifying wildfire events based on the weather data on the same date. Two consistent metrics to compare the performance of model are used in the following training processes, which is ROC and accuracy. The train test, validation and fit set's sizes are following the logic that we described in the previous sections, and the total observations of events used in the following analysis is 3432 with 7 variables, including: daily average temperature (F), daily dew point average temperature (F), daily average relative humidity (%), daily average wind speed, daily average pressure altimeter and daily average wind direction. Models will be trained with fit set and tested by validation set for model selection process, and the final model will be trained in training set and tested on testing set for final conclusion.

Model 1. Naïve Model

The naïve model is based on the probability of the occurrence of 1 in the total dataset. The prediction on validation dataset is shown in Table 4, where the predictions are based only on the probability of the occurrence of the event.

Table 4. Confusion Matrix for Naïve Model (on validation set)

		Actual	
		Wildfire Exist (1)	No Wildfire Exist (0)
Prediction	Wildfire Exist (1)	69	72
	No Wildfire Exist (0)	59	75

Without implementing the machine learning model, the prediction on wildfire can reach 52.36%, however, significantly large amount of positive (1) classes are incorrectly predicted as negative (0) classes.

Model 2. General Linear Model (GLM)

The GLM used linear regression model. Table 5 shows the result in predicting the validation set based on training with the fit dataset. The ROC for the fitting process is 0.7505, and the accuracy reaches 66.9% for the validation set.

Parameters used for tuning model:

- method = "glm",
- trControl = trainControl(method="cv",number=5, savePredictions = TRUE, classProbs=TRUE,summaryFunction = twoClassSummary),
- metric="ROC"

Table 5. Confusion Matrix for GLM Model (on validation set)

		Actual	
		Wildfire Exist (1)	No Wildfire Exist (0)
Prediction	Wildfire Exist (1)	85	48
	No Wildfire Exist (0)	43	99

Table 6 shows the coefficient in the tuned model, which is the result of the fitting process. Air temperature has the highest significant level in the model, which indicates that it is a core factor in classifying the occurrence of wildfire event compared to other variables.

Table 6. Coefficient in the GLM model (on validation set)

	Coefficient	Significant Level
Air Temperature	-0.0434	**
Dew point temperature	-0.026	
Relative Humidity	0.023	*
Wind Speed	-0.036	*
Precipitation	14.488	*
Pressure Altimeter	-0.396	
Wind Direction	-0.001	

Model 3. Random Forest

Random Forest model has been trained by the fitting set with the following parameters:

Parameters used for tuning model:

- method = "parRF",
- metric="ROC",
- trControl = trainControl(classProbs=TRUE,summaryFunction = twoClassSummary)

Table 7. Confusion Matrix for Random Forest Model (on validation set)

		Actual	
		Wildfire Exist (1)	No Wildfire Exist (0)
Prediction	Wildfire Exist (1)	106	61
	No Wildfire Exist (0)	22	86

The performance of random forest is interesting, as it shows high accuracy in predicting the positive class and low accuracy in predicting the negative class. The overall accuracy is 69.82%, with ROC equals to 0.7915

Model 4. Multivariate Adaptive Regression Splines (MARS)

MARS implemented the following parameters in the model:

Parameters used for tuning model:

- method = "earth",
- tuneGrid = expand.grid(.degree=1:5,.nprune=2:20)
- metric="ROC",

- trControl = trainControl(classProbs=TRUE, summaryFunction = twoClassSummary)

Table 8. Confusion Matrix for MARS Model (on validation set)

		Actual	
		Wildfire Exist (1)	No Wildfire Exist (0)
Prediction	Wildfire Exist (1)	101	52
	No Wildfire Exist (0)	27	95

The best performance of ROC is observed when the nprune is 15 and the degree is 2. The importance level of variables is shown in Figure A.1, with only pressure altimeter not significant in the modeling process. The result shown in Table 8 is similar to the result from parRF model shown in Table 7, as less false negative predictions for this model can be seen from the matrix. The ROC is 0.768, and the accuracy of prediction on the validation set is 71.27%.

Comparison of Four Models

In order to determine the final model to be used on train and test set, we compared the performances of models in Table 9. The ROC for Random Forest and MARS are very close, and their accuracy on prediction are similar as well. We decided to use MARS as the final model for prediction due to its better performance on dealing with the true negative predictions and its higher accuracy.

Table 9. Compare all models based on the accuracy on prediction and ROC on the training model

	Accuracy	ROC
Naïve Model	52.36%	NA
GLM	66.9%	0.7505
Random Forest	69.82%	0.7915
MARS	71.27%	0.768

Final Model Selected

From the previous comparison, we determined that the MARS model should be the best model to train the whole wildfire dataset and made the prediction. The importance level of variables is plotted in Figure A. 2. Table 10 shows the accuracy of the prediction on each class is in the test set. The accuracy reaches 71.72%, and the ROC is 0.769.

Table 10. Confusion Matric for MARS Model (compared on test set)

		Actual	
		Wildfire Exist (1)	No Wildfire Exist (0)
Prediction	Wildfire Exist (1)	270	108
	No Wildfire Exist (0)	86	222

The accuracy of prediction on the occurrence of fire by fire size class and fire causes are shown in Figure 10a and Figure 10b. Class G are bet predicted by the model, and small fires with class D has the lowest accuracy rate, but the over all accuracy rate are consistent for all fire classes. This indicates that wildfires with G class have values in input variables that are more different compared to other fire classes. However, this is very different if we look at the accuracy of prediction by the group of cause of the fire. Wildfire with the causes of fireworks, railroad and powerline can be well predicted with 100% of accuracy, but the prediction of wildfire caused by children and smoking can be hardly predicted. This might be the reason that human behavior is less predictable comparing the system false.

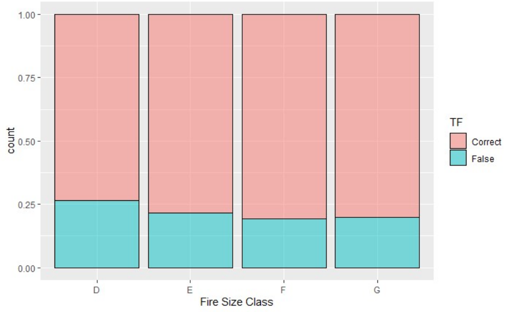

Figure 10a. Correct and False Prediction on Test Set by Fire Size Class. All accuracies are close, with G class having the highest accurancy.

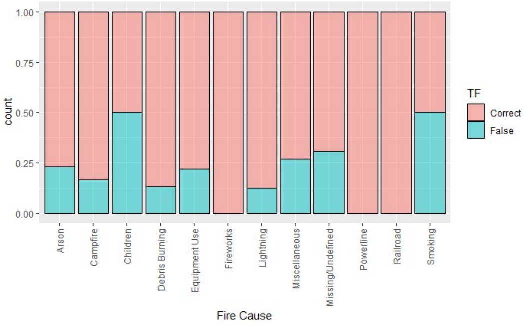

Figure 10b. Correct and False Prediction on Test Set by Fire Cause. Significanlty various between different fire causes. Wildfire with human cause are less predictable comapres to wildfires that are addressed to system cause.

PDA with Subset of Wildfire Events by Regions

To test whether the model can have better prediction on specific regions, the subset of wildfire data from that close to the selected weather stations with a distance of 100km are extracted based on the location of the weather station assigned to the observation, and three of them are selected for the following training and prediction process: RAL, FAT UKI. The test of performance of the model is based on the ROC and Accuracy, with confusion matrix help reporting the result of the prediction. Dataset for training and testing are separated with 8:2 ratio, and the subset of the negative observations (no wildfire) are down sampled with random selection as what we did in the previous training process.

PDA with Dataset of Wildfire in <100km Weather Station RAL

- Tuning Parameters: Consistent with the Final Model for the training of whole dataset with all type of causes
- Size of the dataset: 404
- Training Variables: Consistent with variables in the whole dataset training

Table 11. Confusion Matric for MARS Model for <100km RAL Subset (test set prediction)

		Actual	
		Wildfire Exist (1)	No Wildfire Exist (0)

Prediction	Wildfire Exist (1)	28	8
	No Wildfire Exist (0)	6	39

ROC at nprune = 4 and degree = 1 reached the highest value, 0.789, and the accuracy on the test set is 82.72%

Temperature, wind speed and dew point temperature are significant variables in training the model, figure of the importance are shown in Figure A.3.

PDA with Dataset of Wildfire in <100km Weather Station FAT

- Tuning Parameters: Consistent with the Final Model for the training of whole dataset with all type of causes
- Size of the dataset: 516
- Training Variables: Consistent with variables in the whole dataset training

Table 12. Confusion Matric for MARS Model for Equipment Use Subset (test set prediction)

		Actual	
		Wildfire Exist (1)	No Wildfire Exist (0)
Prediction	Wildfire Exist (1)	43	13
	No Wildfire Exist (0)	13	34

ROC is 0.799 when nprune = 3 and degree = 1, and the accuracy on the test set is 74.76%.

Only the temperature and wind speed are significant in the model. Relatively high false positive are detected. The model might reach better performance if the training size could be larger. The level of importance for variables in the training model can be referred to Figure A.4.

PDA with Dataset of Wildfire in <100km Weather Station UKI

- Tuning Parameters: Consistent with the Final Model for the training of whole dataset with all type of causes
- Size of the Equipment Use dataset: 314
- Training Variables: Consistent with variables in the whole dataset training

Table 13. Confusion Matric for MARS Model for Arson Subset (test set prediction)

		Actual	
		Wildfire Exist (1)	No Wildfire Exist (0)
Prediction	Wildfire Exist (1)	30	12
	No Wildfire Exist (0)	8	13

ROC for this model is highest when nprune = 8 and degree = 1, and the value for ROC is 0.829. The accuracy of model on test set is 68.25%

Relative humidity, dew point temperature, wind speed and wind direction are the significant factors that correlate to the appearance of wildfire in this region. The level of importance for variables in the training model can be referred to Figure A.5. This model is trained on very small dataset.

Summary

All the performance of models in this section show high prediction of false positive results, but the overall accuracy is still promising. The accuracy for wildfire prediction close to RAL shows 82~% of high accuracy. We conclude that the model can be used to train and predict the event of wildfires by using regional dataset, and some of the models have great performance. However, the local distribution of land resources and tree coverages might influence the training model as well, so more variables are needed to be considered if we want to improve the training accuracy for certain regions with more complicate distribution of land resources. Furthermore, transfer learning can be considered to predict the regional wildfire event when the dataset for one location is small to conduct machine learning.

Conclusion

The wildfire events in California shows significant spatial clustering pattern, and there does exist high correlation with the weather condition. After assigning fire events that are in the distance of 100 km to the selected 11 weather stations with high frequency of high-level fire size class in the region, we observed from the EDA that on the date of the wildfire events being discovered, their weather data on that date does slightly clustered differently compared with the day without fire events. Further PDA process discovered MARS as a promising model which can accurately predict the wildfire event with 70~% accuracy. However, many negative cases, where there are no wildfires happened, are incorrectly predicted as a positive case. If more data can be included in the analysis, we might have a better result in accuracy. In addition, we found out different type of causes of fire can have different accuracy on prediction and causes with more systematical errors can be better predicted compared to fires caused by human action. Moreover, for models that are trained by local datasets, the accuracy can hit 80~% in predicting regional wildfire events, which can be explained by the better consistency of the regional environment for observations close to each station, while the weather station that have less wildfires around been observed have very weak performance, which indicates that transfer learning can be implemented in the future. In addition, adding more variables in the training model, such as the forest coverage, might also have improvement on the accuracy. CDA has been conducted to study the correlation between average burn time of wildfire, population density and the frequency of wildfire appearance in each county, and very little correlation has been found.

Appendix

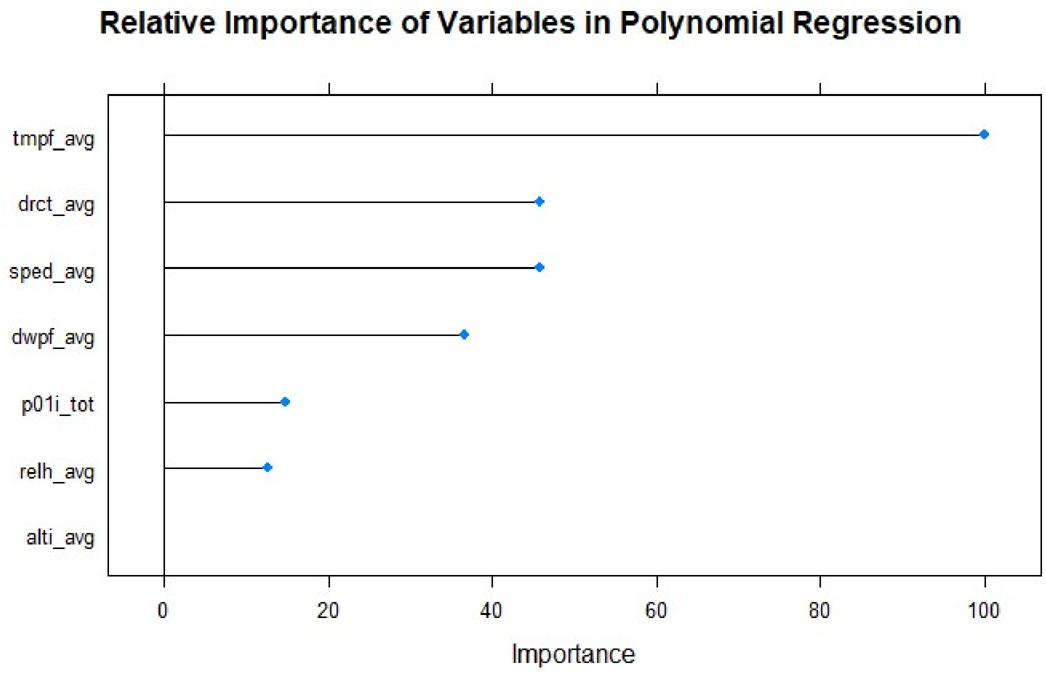

Figure A. 1. Importance level of variables for MARs model with fit set

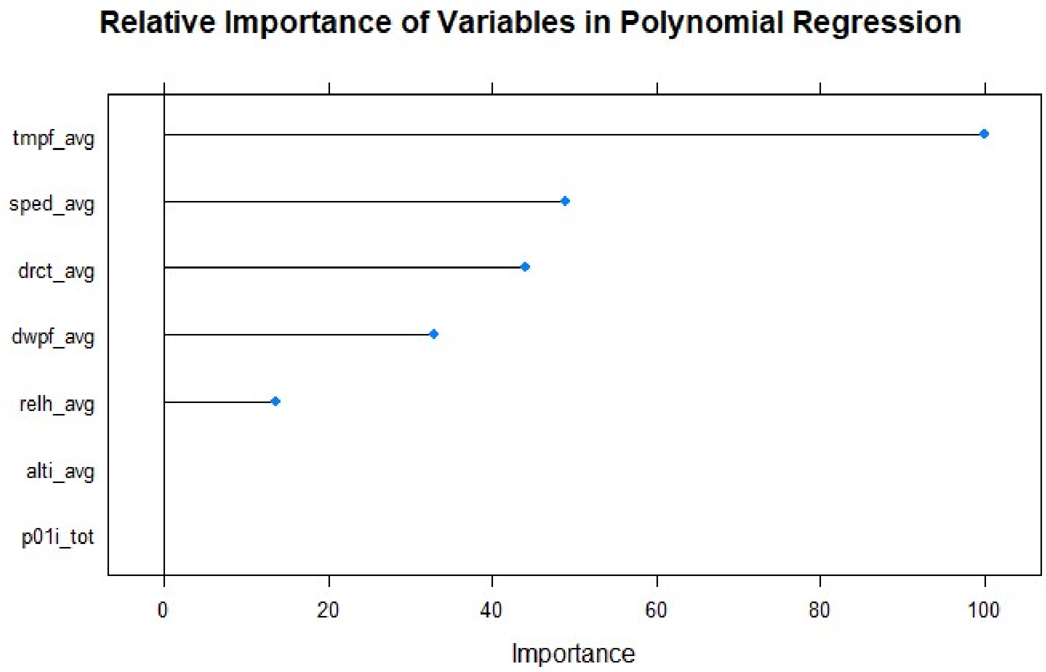

Figure A. 2. Importance level of variables for MARs model with train set (Full Dataset)

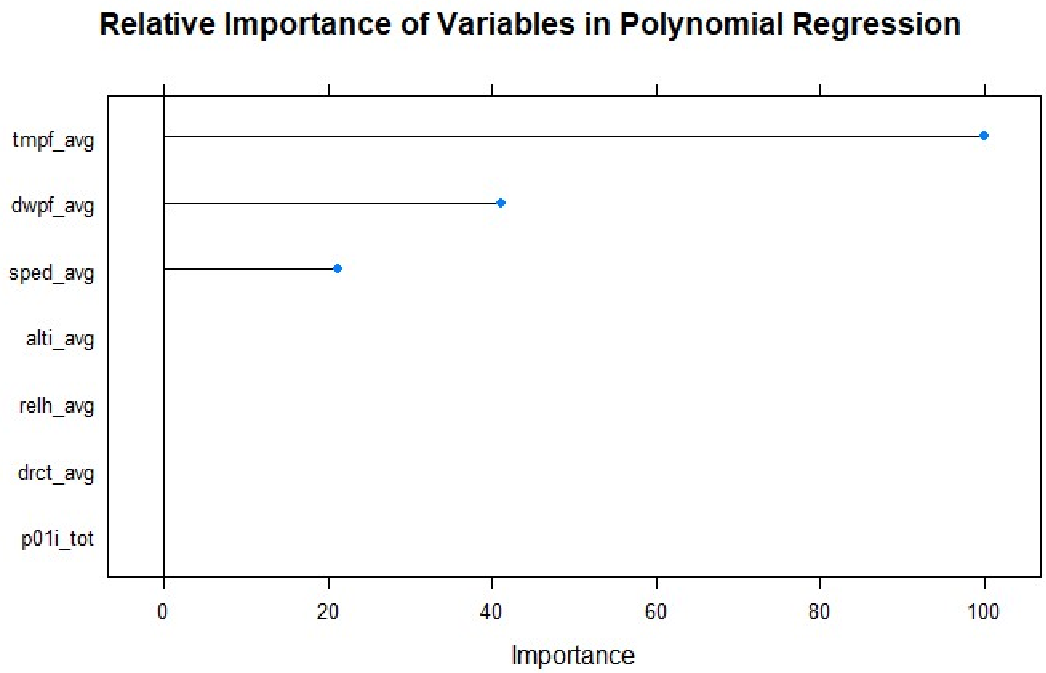

Figure A. 3. Importance level of variables for MARS model with regional analysis: <100km RAL dataset

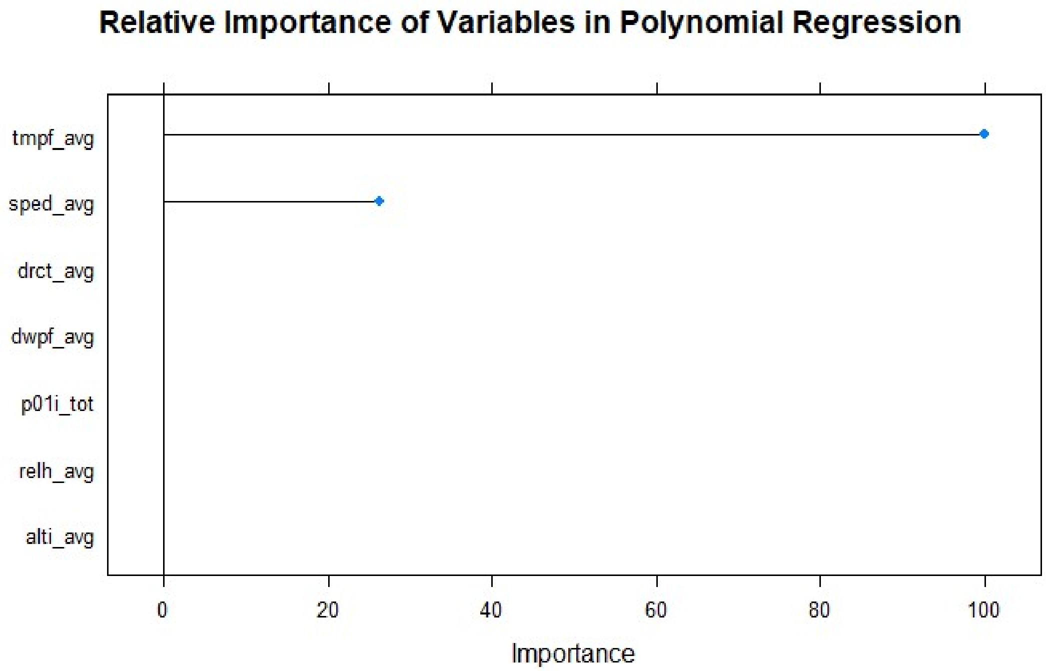

Figure A. 4. Importance level of variables for MARS model with regional analysis: <100km FAT dataset

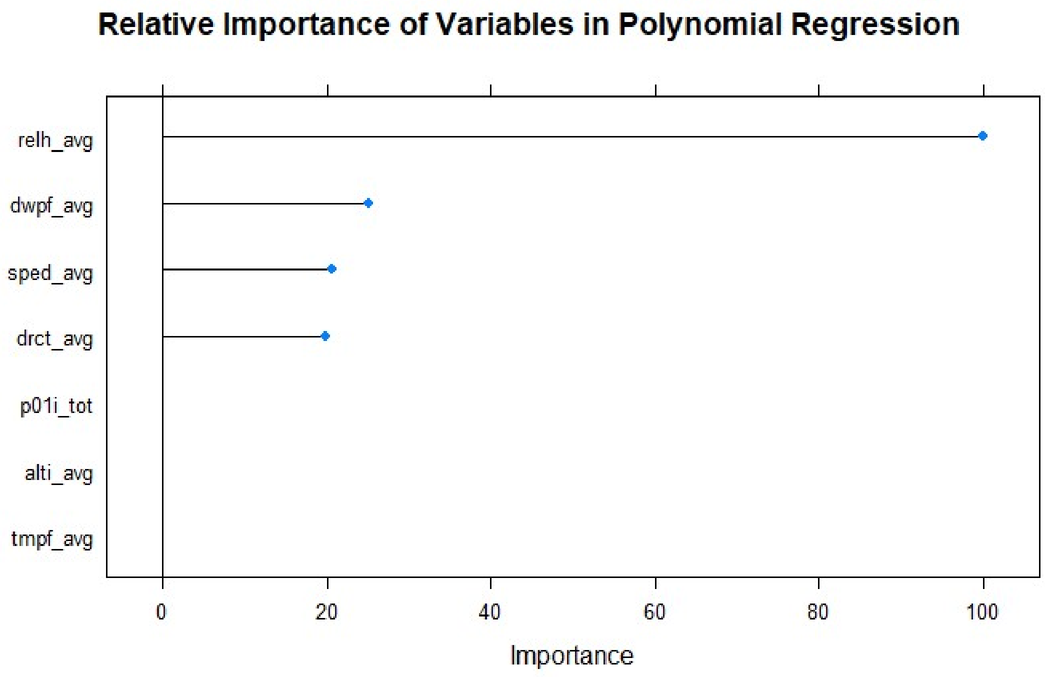

Figure A. 5. Importance level of variables for MARS model with regional analysis: <100km UKI dataset

Table A.1 Full Name of Weather Stations and Locations (long and lat)

Abbreviations	Name of Station	Latitude	Longitude
SDB	SANDBURG (AUT)	34.74338	-118.725
PMD	PALMDALE PRODUCTION	34.62939	-118.085
RAL	RIVERSIDE MUNICIPAL	33.95189	-117.445
PSP	PALM SPRINGS RGNL	33.82967	-116.507
BFL	BAKERSFIELD/MEADOWS	35.4344	-119.054
CIC	CHICO MUNICIPAL	39.79539	-121.858
RDD	REDDING MUNICIPAL	40.509	-122.293
UKI	Ukiah	39.1258	-123.201
BLU	BLUE CANYON (AMOS)	39.27497	-120.71
SJC	SAN JOSE INTL A	37.3594	-121.924
FAT	FRESNO AIR TERMINAL	36.78	-119.719