# 2. Data

## 2.1 Study Area

* Northern California
* Wildfires in Northern California
* 4km grid
* figures with Northern California and Grid

## 2.2 Target Variable

The occurrence of wildfire in the observed geospatial units during the study period serves as the target variable for this study. The period between the years 2010 and 2018 constitutes the period of analysis.

Data on the occurrence of wildfire ignitions were obtained from the "Fire Perimeters" data set, compiled, and provided by the Fire and Resource Assessment Program (FRAP), a joint effort of the California Department of Forestry and Fire Protection (CAL FIRE), the United States Forest Service Region 5, the Bureau of Land Management, and the National Park Service of the United States (CAL FIRE, 2021). “Fire Perimeters” is the most complete and frequently updated database on wildfire occurrences in California. This data set is provided as a shapefile and displays the perimeters of all recorded wildfire occurrences in California, along with the exact date of a wildfire’s discovery, as well as its extinguishment, all harmonized in the database. I used a subset of this data set corresponding to the study area and period, including all recorded fire throughout the entire years.

The location accuracy of the recorder wildfire ignitions made this data well suited for spatial analysis. QGIS can access “Fire Perimeters” directly through the ArcGIS REST API, after which it must be projected to a suitable map projection for further processing. For this project I chose to use the “NAD 1983 California (Teale) Albers (Meters)” projection, which is recommended for statewide datasets of California due to its property of having the coordinate system’s origin at the center of the state (Patterson, 2021).

After projection the QGIS spatial analysis join algorithm could be used to register all intersections of a wildfire perimeter and the grid made up of 4 squares, which serve as the units of observation of this study. The resulting table records all dates for which the 4 square elements of the grid have intersected with a fire perimeter. Note that this does not mean that a given 4 square element was completely covered by a wildfire perimeter (and hence was burned completely), merely that at least a single wildfire ignition has taken place and was recorded within the bounds of that specific 4 square.

In order to further process this data, this table had to be transformed. It is not the date of a wildfire ignition that is of interest for this study per se, but the wildfire ignition status of the grid elements during the observed intervals of the study period. To represent this within the data set, the data was transformed so that each sample represented the wildfire ignition status of a 4 square for each month of the study period of 2010 to 2018. This binary variable called *fire*, with the possible values of *fire* and *none*, served as the target variable for all predictive models of this study.

The monthly distribution of wildfire ignition events shows a clear seasonality, as displayed in figure X. The vast majority of recorded wildfires were registered as active during the summer and autumn months. This meant that the monthly data could be aggregated to a seasonal level, reducing the overall number of samples in the data set while preserving as much information on wildfire occurrence in Northern California as possible. Another motivation for this aggregation was the fact that many predictors were not available at the monthly level, making a data set at this level too granular for the variation contained in the predictor variables. In line with the study conducted by Tonini and co-authors, the period from May to October was assigned to the summer season, leaving the period from November to April to the winter season (Tonini *et al.*, 2020).

The number of recorded events (*fire*) and non-events (*none*) have proven to be strongly imbalanced, with non-events making up the vast majority of all samples in the data set. The aggregation to the seasonal level has slightly improved this circumstance, increasing the share of samples reporting an active wildfire from 0.72% to 2.11%. Despite this, the seasonal data set still musters a high imbalance in the target variable’s values. This is common in cases of extreme-events prediction. The implications of this circumstance on the modeling process and different strategies for addressing potential problems are discussed in depth in chapter X on the methodologies employed by this study.

## 2.3 Predictor Variables