# 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

Overall, I compiled a set of 54 predictor variables in total. Not all of these predictors were ultimately used for modeling. Chapter X on feature selection discusses why some of these predictor variables were left out of the modeling process. These predictor variables were chosen both for their documented use in previous studies on wildfire modeling, as well as their availability for the study area of Northern California and the period of 2010-2018.

These data were acquired at the highest available granularity in order to introduce as much variation into the final data set as possible. I included predictors of multiple categories, all of which are relevant to the occurrence of both human-caused wildfire ignitions and naturally occurring wildfires, similar to the study conducted by Oliveira and co-authors (Oliveira *et al.*, 2012). The included categories of predictor variables are environmental data (including both topographic, meteorologic and data concerning land cover), infrastructure data (both the proximity to human-made infrastructure, in addition to binary data concerning the presence of infrastructure in the units of observation), as well as demographic and socio-economic data for the study area.

### 2.3.1 Environmental Predictors

Topographical features such as elevation are important predictors of spatial patterns of fire, as they account for local variations in climate, in addition to exerting influence on ground flammability through their impact on soil and fuel moisture and the vegetational distribution of land cover (Whelan, 1995; Syphard *et al.*, 2008; Oliveira *et al.*, 2012). A digital elevation map of California at a 90m resolution based on satellite imagery has been compiled by the National Aeronautics and Space Administration (NASA) and the National Geospatial-Intelligence Agency (NGA) and is distributed as a raster band data set, where each pixel of the map corresponds to a numeric elevation value (NASA & NGA, 2000). This very high resolution means that the data had to be aggregated to the 4 level of the grid elements of this study. The QGIS software provides tools to process raster data and the zonal statistics toolset could be used to calculate the average elevation value of each intersection of the grid of 4 squares and the elevation raster map.

Furthermore, topographical data on the presence of major bodies of water, such as lakes and rivers, were added to the map. Bodies of water act as natural fire barriers and directly influence soil moisture and vegetation in their vicinity. These data are provided as shapefiles by the California Department of Fish and Wildlife (California Department of Fish & Wildlife, 2015, 2018) These data were used twofold: In a first step, a binary variable was created indicating whether an object of observation is intersected by either a lake or a river. In a second step, the distance of each 4 square’s centroid to the nearest element of both the river and the lakes data set was calculated with the *v.distance* algorithm of the GRASS package for QGIS. This provides an additional indicator to the presence of water bodies, that is numeric and continuous, as opposed to the logical dummy variables created in the first step.

The local vegetation and land cover are regularly cited as being associated with fire occurrences – both natural and caused by humans (Syphard *et al.*, 2008; Martínez, Vega-Garcia and Chuvieco, 2009; Oliveira *et al.*, 2012). Due to the strong local variations in climate, land cover not only indicates the naturally occurring fuel types, but also the various biomes found in Northern California. I hence included categorical data on the land cover and land use of California. The data was compiled by the Department of Geography at the University of California as a single shapefile, depicting the canopy dominant vegetation species for the entire state (Department of Geography UC Berkeley, 2014). The “California Wildlife Habitat Relationships” system provides a detailed classification of tree dominated, shrub dominated, herbaceous dominated, aquatic, developed and non-vegetated habitats, each with their own subcategories. Due to this highly detailed breakdown of the dominant land cover, this data can act as a proxy for the primary fuel type within the 4 squares that act as the units of observation. QGIS was used to determine the most frequent land cover type for each square.

Meteorological factors are known predictors of wildfire occurrence, as they affect fuel accumulation and ground moisture, creating the conditions that may favor or hinder fire ignitions from occurring (Syphard *et al.*, 2008; Vilar *et al.*, 2010; Oliveira *et al.*, 2012). The WorldClim database offers monthly historical temperature and precipitation data at spatial resolution of 2.5 minutes (corresponding to roughly 21) in raster format (SOURCE). Due to the large number of raster layers (monthly interval, eight-year study period, three data sets), these predictors had to be constructed algorithmically using QGIS’ python interface and extracting the mean values of minimum temperature, maximum temperature and mean precipitation for each unit of observation algorithmically. Ultimately these predictor variables were aggregated to the seasonal level, along with the target variable.

### 2.3.2 Infrastructure Predictors

Access to roads has often been described as a driver of economic activity and a proxy for infrastructure development (SOURCE). In the context of fire occurrence, road access and the distance to roads are frequently used predictor variables, since these factors also determine the speed of the response of a given fire containment strategy (Martínez, Vega-Garcia and Chuvieco, 2009; Oliveira *et al.*, 2012). The Californian road system is well documented and provided as a shapefile containing all major roads (MTFCC codes S1100 and S1200) by the U.S. Census Bureau (US Census Bureau, 2015). Similar to how the GIS data on rivers and lakes was processed, this data set was used to both create dummy variables indicating the presence of a major road for each 4 unit and calculate the distance from each unit’s centroid to the nearest major road as well. The same process was repeated for a data set of powerlines, resulting in predictor variables indicating both their presence (binary) as well as the distance from each unit’s centroid to the nearest powerline.

For recreational routes, campgrounds, picnic sites and state parks only their presence was determined, as I do not expect these recreational structures to have any continuous effect if they’re not present – unlike powerlines, where larger distances function well as a proxy for a lack of economic development.

### 2.3.3 Demographic Predictors

County pop-growth

County persons per households

County vacancy rate

Perc democrats / republicans

Perc yes prop 21, 23, 65, 76

### 2.3.4 Socio-economic Predictors