# 2. Data

One of the primary challenges of this study was the compilation of a harmonized data set that was sourced from both GIS data (“geographic information system”) for all the spatial variables, as well as tabular data for the socio-economic, demographic and political predictors. All GIS data was processed and joined in QGIS, whereas the tabular data was joined to the harmonized GIS data set in R, a software designed for statistical computing.

## 2.1 Study Area

The study area encompasses the northernmost counties of California, stretching from Sutter County north of Sacramento, up to the Canadian border. This diverse region is made up of 18 counties[[1]](#footnote-1) with a combined area size of roughly 113’380 (USA.com, 2021). This region typically has a milder climate than the more southern counties with higher precipitation and more humid climate. Conifer forests, oak woodland and shrubland dominate the flora, although there is a considerable variation in the dominant vegetation pattern across the area. National forests and rugged mountain ranges dominate the landscape in this northern part of the state, as the foothills of the Sierra Nevada stretch up until the very north of the state.

This area has seen a large number of devastating wildfires in recent years. The three largest wildfires in the history of California have all erupted in the study area over the past three years. The Dixie fire, ignited in the Sierra Nevada in summer of 2021, has evolved into largest single-source fire in the history of the state (Bermel, 2021). It rivals the August Complex wildfire of 2020 in size and has grown to almost twice the size of the Mendocino Complex fire of 2018 (CAL FIRE, 2021c). The deadliest wildfire in the history of California ignited in Butte County in 2018, with 85 recorded deaths (CAL FIRE, 2021b).

Ein Bild, das Karte enthält.

Automatisch generierte BeschreibungUsing the QGIS software I have divided the study area into a grid of squares with an area size of 4, making up the units of observation of this study. I have chosen this area size for the individual units due to some of the environmental predictor data only being available at this resolution, as well as the implications on computational cost when increasing the granularity any further.

Figure 1: Study Area, with all recorded wildfires during the study period of 2010-2018

## 2.2 Target Variable

The occurrence of wildfire in the 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, 2021a). “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 the area, along with harmonized data such as the exact date of a wildfire’s discovery, as well as its extinguishment. I used a subset of this data set corresponding to the study area and period, including all recorded fires throughout each year.

The location accuracy of the recorded 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 was 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*, serves as the target variable for all predictive models estimated for this study.

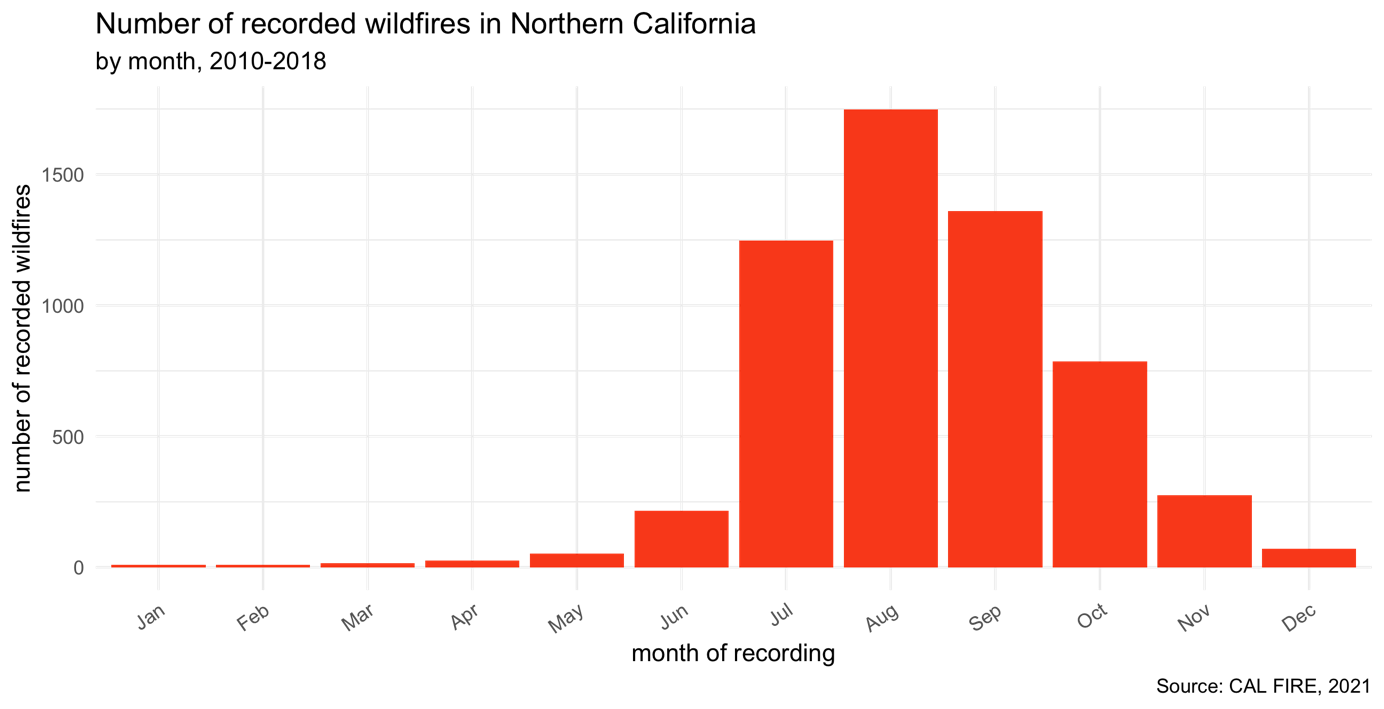


Figure 2: Monthly distribution of wildfire occurrences during the study period

The monthly distribution of wildfire ignition events shows a clear seasonality, as displayed in figure 2. 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 3 on the methods used for data pre-processing and model evaluation.

## 2.3 Predictor Variables

Overall, I compiled a set of 54 predictor variables in total. Not all of these variables were ultimately used for modeling. Chapter 3 on data preprocessing and predictor selection discusses why some of these variables were left out of the modeling process. These predictor variables are chosen both for their documented use in previous studies on wildfire modeling, as well as their availability for the study area of Northern California during the study 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 indicating 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 has to be aggregated to the 4 level of the grid elements of this study. The QGIS software provides tools to process raster data and toolsets to calculate the zonal statistics such as the average elevation value for each of the 4 squares of the grid.

Furthermore, topographical data on the presence of major bodies of water, such as lakes and rivers, are 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 are used twofold: In a first step, a binary variable is 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 is 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 often 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 units of observation. QGIS is 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 are constructed iteratively, using QGIS’ python interface extract the mean values of minimum temperature, maximum temperature and mean precipitation for each unit of observation from all raster layers. 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. Powerlines are a frequently cited cause of wildfire ignitions and have been used to construct predictor variables in similar studies (Oliveira *et al.*, 2012; Texas Wildfire Mitigation Project, 2014; McFall-Johnsen, 2019).

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

Population density is an especially important predictor for human-caused wildfires, as it describes the distribution of potential “causative agents” for fire ignitions (Oliveira *et al.*, 2012). I obtained data on population density from SOURCE, which provide annual raster data at the LEVEL level (SOURCE). I calculated the average population density per 4 unit in QGIS. Additionally, I added the annual population growth at the county level to the data set. This data is supplied by SOURCE and acts as another proxy for economic development. Martínez and co-authors’ study on human-caused forest fires in Spain included a variety of housing data as their predictors (Martínez, Vega-Garcia and Chuvieco, 2009). In a similar fashion, I included annual data on the average vacancy rate at the county-level, as well as the average number of people per household at the county level as predictors.

In order to measure the political leanings of the inhabitants of the study area I included voting district level data on the share of registered Democrats and Republicans. Similarly, California’s direct democratic instruments allow for operationalization of political leanings beyond the mere party spectrum by taking vote shares on referendums into account. Californians are able to propose laws and constitutional amendments by way of so-called “ballot initiatives”, even without the support of the Governor or the Legislature (*Ballot Initiatives*, 2011). Out of all the ballot initiatives that were voted on during the study period I selected a subset of four initiatives that pertained to environmental issues, gauging the environmental sensibilities at the county level:

|  |  |  |
| --- | --- | --- |
| Proposition number | Year | Description |
| 21 | 2010 | Vehicle License Fee Increase, dedicated to state parks and wildlife programs |
| 23 | 2010 | Suspension of GHG- emissions reduction law, until California's unemployment rate decreases to 5.5% |
| 65 | 2016 | Dedication of Revenue from Disposable Bag Sales to Wildlife Conservation Fund |
| 67 | 2016 | Ban on sale of plastic bags. |

Table 1: Californian Ballot Initiatives relativ to environmental issues during the study period

For each of these ballot initiatives the number of Yes-votes at county level were added to the data set.

### 2.3.4 Socio-economic Predictors

Previous studies have found the rate of unemployment to be important predictors for wildfire occurrences (Martínez, Vega-Garcia and Chuvieco, 2009; Oliveira *et al.*, 2012). Oliveira and co-authors mention two specific channels how the rate of unemployment may factor into wildfire risk: Both as a proxy for social conflict, which can cause increases in arson as part of generic vandalism, or arson as a deliberate strategy to increase the local demand for firefighters and thus further a person’s chance of employment (Oliveira *et al.*, 2012). Besides the rate of unemployment, I also added the absolute number of unemployed adults and the monthly growth in unemployment as well. The data on monthly unemployment rates at county level are provided by the California department of employment and development (EDD), which I aggregated to the seasonal level along with the target variable (EDD, 2021b).

The EDD also provides monthly labor data denoting the share of employed workers by industry at county level (EDD, 2021a). These data were also added to the data set and aggregated to the seasonal level. Implicit zeros, which are appear in the data base as missing values, had to be introduced during the data cleaning process.

# 3. Methods

I split the data into subsets of training and testing data. All data from the years 2010-2016 are used for training the models. The data from the years 2017 and 2018 will be used for evaluating the predictions made based on the training data. I chose this approach over the usual random split done in most Machine Learning use cases, since it mirrors the process of forecasting wildfire risk on past data for future seasons.

The training set is used to estimate increasingly complex and sophisticated models, using three algorithms that are appropriate for binary classification cases to model wildfire occurrence: logistic regression (GLM), Random Forest (RF) and xgboost (XGB). For each of these three algorithms models are estimated in three distinct steps:

1. A naïve model estimation strategy without resampling or hyperparameter tuning, using the heavily imbalanced training data set (no up- or downsampling).
2. A more informed model estimation strategy where the training data is resampled using 5-fold cross validation and the imbalance in the training set is addressed both with upsampling and downsampling methods.
3. A modeling strategy focused on maximizing predictive performance by selecting hyperparameters from a grid using grid search, again using resampling and subsampling methods to ensure stable results.

## 3.1 Logistic Regression

Regression models have previously been widely used in cases that model wildfire occurrences, especially logistic regression (Syphard *et al.*, 2008; Catry *et al.*, 2009; Martínez, Vega-Garcia and Chuvieco, 2009; Oliveira *et al.*, 2012). This parametric, linear model is used for estimating the probability of event occurrences and is common for binary classification cases. The logit transformation of the binary target variable is modeled with the set of predictor variables. Since logistic regression uses maximum likelihood, many assumptions of linear regression (which uses ordinary least squares instead) do not apply to logistic regression. Assumptions of logistic regression include independent errors, the absence of multicollinearity among the predictors, linearity in the logit for continuous variables, and a lack of strongly influential outliers (Stoltzfus, 2011).

In the case of this study especially multicollinearity must be addressed during the preprocessing of the data, as data exploration identified many predictors as strongly correlated. Data exploration also revealed that the distributions of the distance predictors are heavily skewed. A testing of multiple methods showed a power transformation to be the most effective in creating a more normal-like distribution for these variables. The pre-processing steps are handled for each model separately, taking the individual features of each algorithm into account. Most of the pre-processing steps are similar, however.

For all models, the dummy variables indicating the presence of infrastructure (such as roads) or topological features (such as lakes) are removed, as long as there is a corresponding distance variable present in the data set. These variable pairs are highly correlated, and the distance variables provide more information and introduce more variance into the data compared to the dummies. These distance variables undergo a power-transformation for the regression models only, as logistic regression might benefit from predictors having a more normal-like distribution (Kuhn and Silge, 2021).

In a next step predictors with zero variance are removed, as are strongly correlated predictors (with a threshold of 0.75), similar to Oliveira and co-authors’ study (Oliveira *et al.*, 2012). Categorical predictors are turned into dummy variables through one-hot encoding.

In order to balance out the classes in the target variable, upsampling is done using the SMOTE algorithm, creating additional synthetic observations. For the downsampled versions of the models the NearMiss 1 algorithm is used, which retains observations of the majority class with the smallest distance to the k-nearest neighbors of the minority class. The training data set for both the upsampled and the downsampled models undergo an additional step of Tomek’s links removal, a procedure that removes majority class observations that are the nearest neighbor of an observation belonging to the minority class. This is intended to improve the classification boundary of the training data and the predictive power of the trained model.

During hyperparameter tuning I introduce both and regularization to logistic regression, resulting in an elastic net model. For this specific case all predictors underwent normalization. Elastic net combines both the Ridge and LASSO penalties to shrink the estimated coefficients. Both the size of these penalties as well as the mixture between both are tunable hyperparameters that I tune using a grid search strategy.

## 3.2 Random Forest

Random Forest is a non-parametric ensemble learning algorithm, comprised of a set of Decision Tree models. These submodels aim to create splits in the training data that minimize the heterogeneity in the resulting subsets, learning the best rules to create these splits in the process (Kuhn and Johnson, 2013). This algorithm is known to suffer from large variance, meaning that the smallest changes to the training data can have a large influence on the overall model fit. Random Forest is built upon the idea of exploiting this property by letting its *weak classifiers* vote on the final ensemble predictions based on their own, highly variant predictions. The singular trees are made weak due to limited access to the training data, as each tree is only trained on a set number of bootstrapped predictor variables (Kuhn and Johnson, 2013). This approach intends to counteract overfitting on the training data as no tree is trained on the entire data set. This strategy uses the bias-variance-tradeoff in Machine Learning to drive down bias at the expense of increased variance, intended to ultimately arrive at more accurate and stable predictions.

Random Forest does not share the assumptions of logistic regression, but fewer correlated predictors might improve the variance important score estimates of tree-based algorithms (Kuhn and Johnson, 2019). Strongly correlated predictors indicating the presence of infrastructure or topological features are hence removed from the training data for Random Forest as well, as are predictors with a correlation coefficient exceeding 0.75 and zero variance predictors.

Unlike logistic regression and xgboost, Random Forest does not require one-hot encoding of categorical variables. Despite this, due to the used software requiring exclusively numeric predictors for subsampling one-hot encoding still has to be used for all Random Forest models except the naïve estimation. Subsampling is conducted in the same way as for logistic regression, with SMOTE and NearMiss 1 being used in combination with Tomek’s links removal. Neither Random Forest nor xgboost require any additional preprocessing, highlighting this inherent strength of these non-parametric, tree-based models.

Random Forest has multiple hyperparameters that can be tuned, such as the minimal number of data points required at each node to qualify for further splitting, the number of bootstrapped predictors at each split, as well as the total number of trees grown in the forest. During hyperparameter tuning the latter is kept constant at 500 trees due to computational restraints, while the two former parameters are tuned using grid-search.

## 3.3 xgboost

xgboost is an implementation of the boosted trees method. Similar to Random Forest it is an ensemble algorithm that relies on using multiple Decision Trees as *weak classifiers* (with a predictive power marginally better than random classification)to ultimately combine into a *strong ensemble classifier* (Kuhn and Johnson, 2013). Despite this similarity in concept boosted trees algorithms do not grow independent trees in parallel, as Random Forest does. Instead the Decision Trees grown by boosted trees are fit sequentially with each tree subsequent tree attempting to minimize the loss of its preceding trees by placing more weights on their misclassified samples, updating the predicted values by adding the previous tree’s predictions to the predicted values of the current tree (Kuhn and Johnson, 2013). This approach has been wildly successful and popular in classification tasks, especially with the xgboost implementation that has taken gradient boosting to the extreme with parallelization, fast optimization and convergence and overall computational efficiency. Like Random Forest xgboost also creates *weak classifiers* by limiting the access of Decision Trees to training data, but unlike Random Forest it does so by subsampling the number of samples accessible during each boosting iteration instead of bootstrapping predictor variables.

Like Random Forest, xgboost does not share the assumptions of logistic regression. Nevertheless, the same selection of variables is used to remove highly-correlated predictors from the training data and ensure reliable variable importance scores. The pre-processing steps for xgboost are the same as for Random Forest, with the exception of one-hot encoding of categorical predictors being uniformly applied all models, due to the requirements of the implementation of xgboost.

xgboost shares the hyperparameters of Random Forest and additionally offers tuning the depth of each tree in the ensemble, the reduction in the loss function for further splits, as well as the sample size of the data used for modeling within each boosting iteration. During hyperparameter tuning all of these are tuned using grid-search, except for the number of trees grown, due to the high computational cost associated with large numbers of trees in an ensemble.

## 3.4 Evaluation

Due to the heavy imbalances in the data set’s target variable special precautions must be taken in order to evaluate the estimated models properly. Accuracy is hence disregarded as a metric for evaluating model quality, as models can achieve very high accuracy scores when predicting imbalanced data if all predictions blindly belong to the majority class. In these cases, the *F-measure*, defined as the harmonic mean between *recall* and *precision*, is a more reliable score due to taking both the true positive rate as well as the reliability of positive predictions into account. The *F-measure* is defined as follows, where *TP* denotes true positives, *FP* denotes false positives and *FN* denotes false negatives.

Additionally, not only the imbalance between the classes in the target variable are to be considered, but also their associated cost. There is a higher cost associated with misclassifying wildfire ignitions as non-events than predicting a high probability for wildfire occurrence when in truth no fire ignition took place. This circumstance can be built into model evaluation by weighting misclassifications in such a way that false negatives are penalized more heavily than false positives. I implement such a metric by customizing the classification cost function with a cost matrix that penalizes false negatives twice as much as false positives. This metric is used to select the best hyperparameter settings and ensure the highest possible predictive power of the finalized models.

Since the models used in this study don’t directly predict classes, but class probabilities *Receiver Operator Characteristic* (ROC) curves can be used to evaluate the ideal probability thresholds separating events from non-events (Kuhn and Johnson, 2013). ROC-curves for all model fits are displayed in the appendix. The *Area under the curve* (AUC) of ROC-curves will not be used for model evaluation, due to it being biased towards the majority class and hence not ideal for the context of this specific study. Nevertheless, ROC-AUC is also listed as a performance metrics for all model fits, despite not being used to select the best models or hyperparameter settings, along with the *F-measure*. Performance of models is also visually represented with confusion matrices, displaying a cross-tabulation of the predicted and observed classes, in the appendix.

1. Included counties: Butte, Colusa, Del Norte, Glenn, Humboldt, Lake, Lassen, Mendocino, Modoc, Nevada, Plumas, Shasta, Sierra, Siskiyou, Sutter, Tehama, Trinity, Yuba [↑](#footnote-ref-1)