STA2201 Project: Wildfire in California

Yang Zixuan

Abstract

This project aims to study the wildfire in California during the year of 2013 to 2020. The wildfire incidents are manually classified into two categories, the major incidents and non-major incidents. We want to develop a statistical model to classify the wildfires into these two categories, based on the information including location and time that they occurred. The methods to be used are based on logistic generalized additive model (GAM). In particular, different smoothing structures on the (longitude, latitude) pairs are considered, and there are three candidate models to be compared.The logistic GAM with smoothing terms on (longitude, latitude) and year did a reasonable job on classifying major and non-major wildfires. However due to the limitation of the available data, no further covariates are taken into consideration. For further studies, it worth considering some more predictors related to the specific wildfire. It is clear that models with only spatial and temporal covariates are not suffice for fire prediction. In order to get a better model for fire prediction, we need to include more ecosystem-specific covariates into the model. In addition, it also worth to collect more year's data, as the current dataset has only 7 years of records.

Introduction

California is one of the places having the deadliest and most destructive wildfire seasons. It would be intersting to study wildfire prediction because this enables the fire managment to faster respond to protential fire incidents and allocate any rescourses more efficiently. There are many existing literatures discussing the wildfire prediction problem, like this one Woolford et al. (2021) that aims on predicting the occurance of wildfire incidences in Ontario, Canada. This project will focus on predicting the severity of California wildfires based on certain covariates. While there are already existing methods on the topic of wildfire prediction using machine learning techniques like neural networks or random forest, this project will only focus on statistial methods.

Data

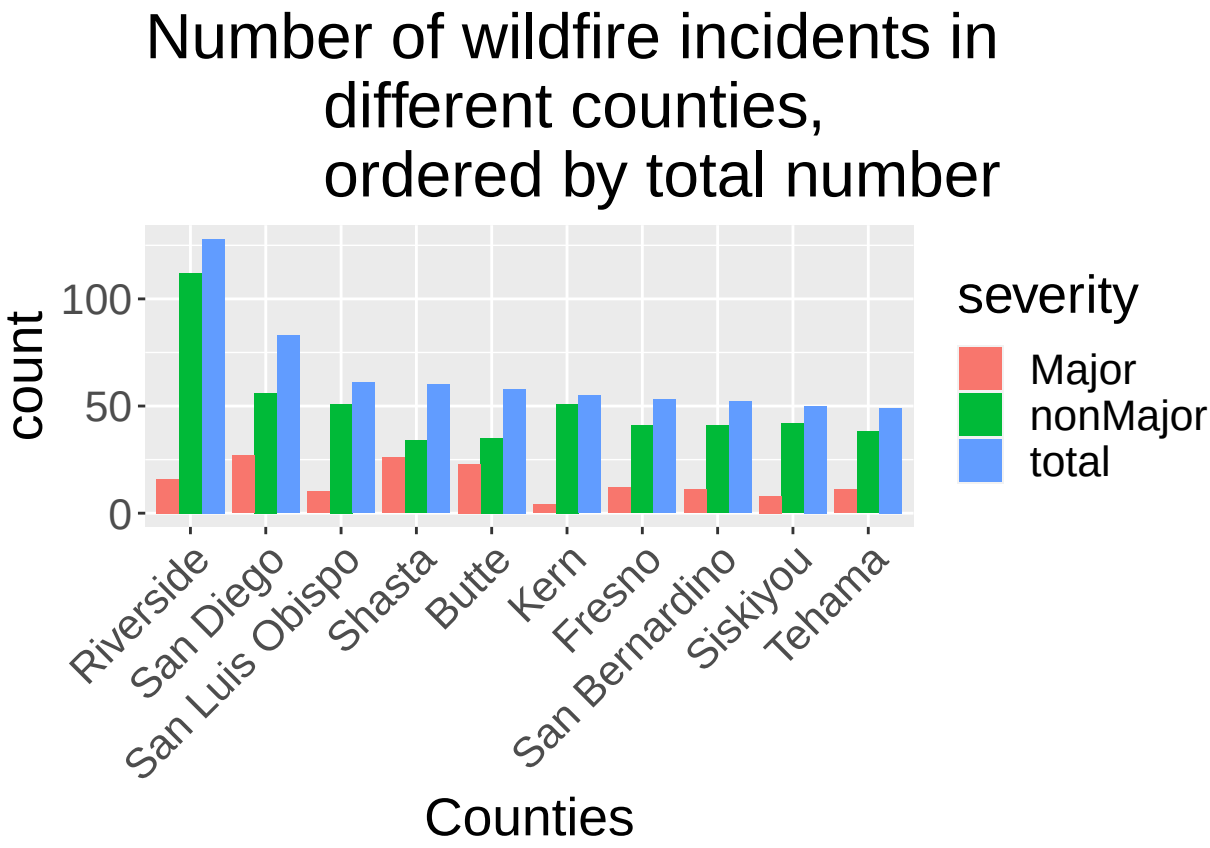
The dataset being considered for this project is about the recorded wildfires occurred in California from 2013 to 2020. This dataset contains the location where wildfires have occurred including the County name, latitude and longitude values and also details on when the wildfire has started. Here is an overview of the data:

Table 1: Overview of the dataset

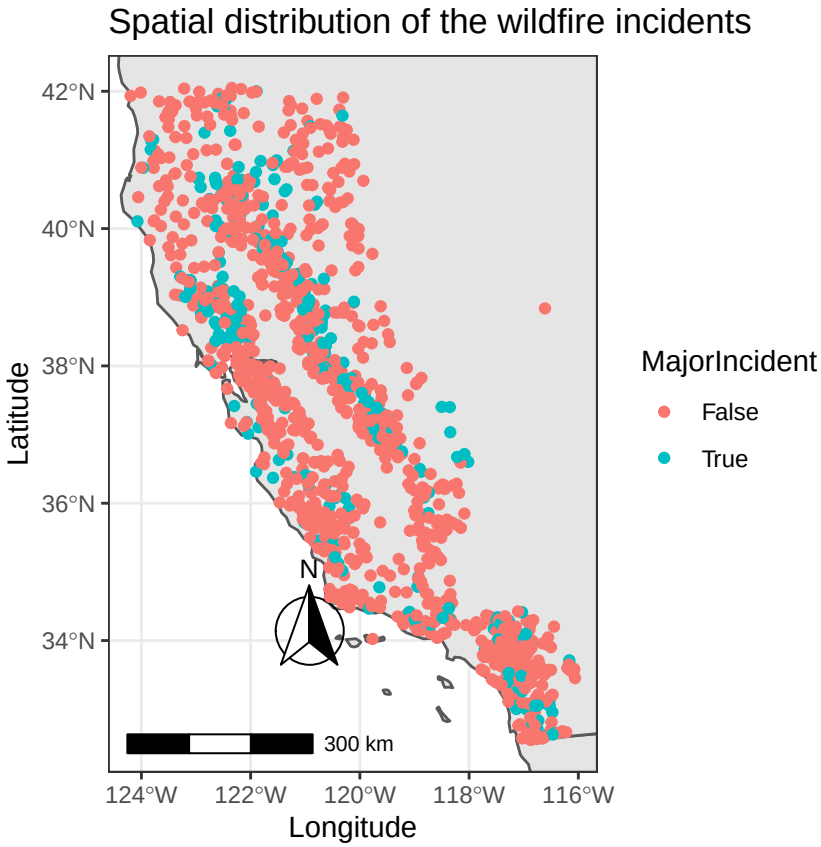
variable	description	type	valid
AcresBurned	Acres of land affected by wildfires	int	100%
Active	If the fire is active or contained?	boolean	100%
AdminUnit	-	char	100%
AirTankers	Resources assigned	int	2%
ArchiveYear	Year the data was archived	int	100%
CalFireIncident	Is the incident treated as a CalFire incident?	boolean	100%
CanonicalUrl	-	char	100%
ConditionStatement	-	char	17%
ControlStatement	-	char	7%
Counties	County name	char	100%

variable	description	type	valid
CountyIds	County id	int	100%
CrewsInvolved	-	int	10%
Dozers	Resources assigned	int	8%
Engines	Resources assigned	int	12%
Extinguished	Extinguished date	date	96%
Fatalities	Fatality count	int	1%
Featured	-	boolean	100%
Final	-	boolean	100%
FuelType	-	char	1%
Helicopters	Resources assigned	int	5%
Injuries	Count of injured personnel	int	7%
Latitude	Latitude of the Wildfire incident	double	100%
Location	Description of the Location	char	100%
Longitude	Longitude of the Wildfire incident	double	100%
MajorIncident	If it is considered a major incident or not?	boolean	100%
Name	Name of the Wildfire	char	100%
PercentContained	What percent of the fire is contained?	real	100%
PersonnelInvolved	-	int	12%
Public	-	boolean	100%
SearchDescription	-	char	99%
SearchKeywords	-	char	88%
Started	Fire start date	char	100%
Status	Status of the fire	boolean	100%
StructuresDamaged	Count of structures damaged	int	4%
StructuresDestroyed	Count of structures destroyed	int	11%
StructuresEvacuated	-	-	0%
StructuresThreatened	Count of structures threatened	int	2%
UniqueId	-	char	100%
Updated	Last updated date	char	100%
WaterTenders	Resources assigned	int	9%

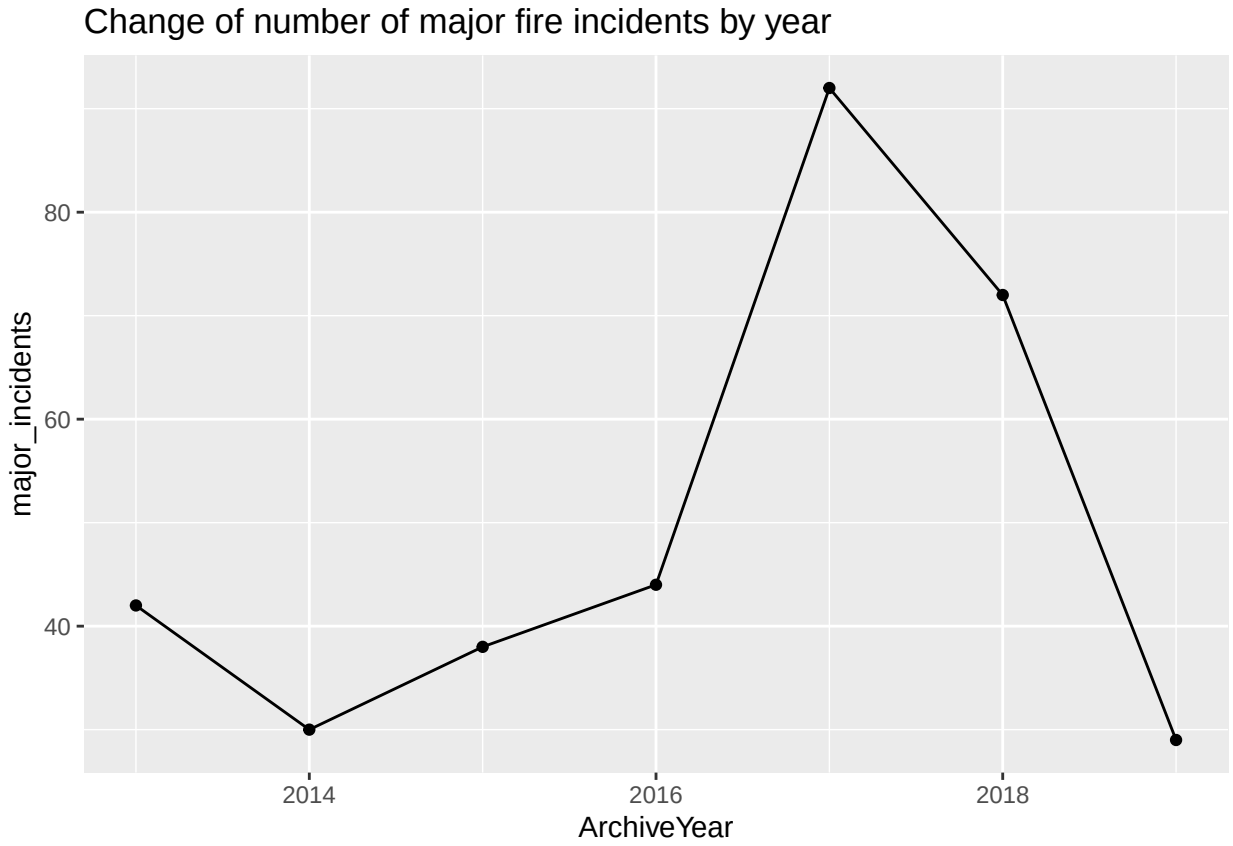
There are several observations where the latitude or longitude are out of range, so will only consider the subset of this dataset where the latitude and longtitude values are valid. There are 58 counties and 6 years (2013 to 2019) in total after filtering out the invalid records, and below is a plot of the number of wildfire incidents in different counties by severity, showing only counties having top 10 total number of incidents:



The severity of fires happened in different counties are quite different. It is obvious that some counties have more major fire incidents than others: for example, most fire incidents in Riverside are classified as non-major despite the large number of total fire incidents, while in Butte there are more proportion of major incidents among the total. In addition, below is a map showing the locations where the wildfire incidents happened, and the change in total number of major fire incidents by time:



The map shows that all the wildfire incidents are concentrated on the western coast, except for one non-major incident to the east most (county Riverside).



Methods

In Woolford et al. (2021) the authors used a logistic generalized additive model (GAM) for fire occurance prediction. In the paper will also use logistic GAM for fire severity prediction. The smoothing term will be put on the pairs $(\text{longtitude}_i, \text{latitude}_i)$, in the form of $f(\text{longtitude}_i, \text{latitude}_i)$ where f is a bivariate smooth function. To simplify the problem, will first consider an additive form of the smooth function in the following section: $f_1(\text{longtitude}) + f_2(\text{latitude})$, and will use the B-spline basis for the basis of f_1, f_2 functions. To be more specific, the models being used will be of the following general form:

$$
\begin{aligned}
& y_i \mid \pi_i \sim \text{Bern}\left(\pi_i\right) \\
& \pi_i = \text{logit}^{-1}\left(X_i\alpha + \sum_{k=1}^{K_1} B_1^{k,t}\beta_1^{k,s} + \sum_{k=1}^{K_2} B_2^{k,t}\beta_2^{k,s}\right) \\
& \beta_1^{1,s}, \beta_2^{1,s} \sim N(0,1) \\
& \beta_1^{2,s} \sim N(\beta_1^{1,s}, \sigma_1^2) \\
& \beta_2^{2,s} \sim N(\beta_2^{1,s}, \sigma_1^2) \\
& \Delta^2\beta_1^{k,s} \sim N(0, \sigma_1^2) \text{ for } k \ = \ 3, .., K_1 \\
& \Delta^2\beta_2^{k,s} \sim N(0, \sigma_2^2) \text{ for } k \ = \ 3, .., K_2 \\
& \sigma_1, \sigma_2 \sim N^+(0,1)
\end{aligned}
$$

where y_i represents whether the ith fire incident is a major incident or not; B_1, B_2 are the matrices of P-splines for longtitude and latitude; s refers to the county membership, t refers to longtitude or latitude, and k refers to the knot position; X_i is the model matrix of some covariates. In addition, in the following sections several models will be fitted based on the selection of additional covariates, and they are going to be

compared based on performance. The candedate covariates include acres burned and the year in which the fire incident is archived. All the models are to be fitted in `R` and `stan` using Bayesian framework, and will put on weakly informative priors on the variance terms.

Model A

The first model to be fitted has only longtitude and latitude as covariates:

$$\begin{aligned} y_i \mid \pi_i &\sim \text{Bern}\left(\pi_i\right) \\ \pi_i &= \text{logit}^{-1}\left(\sum_{k=1}^{K_1} B_1^{k,t}\beta_1^{k,s} + \sum_{k=1}^{K_2} B_2^{k,t}\beta_2^{k,s}\right) \\ \beta_1^{k,s}, \beta_2^{k,s} &\sim rw(2) \end{aligned}$$

Model B

Model B has an additional temporal smoothing term for the year in which the wildfire is archived:

$$\begin{aligned} y_i \mid \pi_i &\sim \text{Bern}\left(\pi_i\right) \\ \pi_i &= \text{logit}^{-1}\left(\sum_{k=1}^{K_1} B_1^{k,t}\beta_1^{k,s} + \sum_{k=1}^{K_2} B_2^{k,t}\beta_2^{k,s} + \sum_{k=1}^{K_1} B_3^{k,t}\beta_3^{k,s}\right) \\ \beta_1^{k,s}, \beta_2^{k,s}, \beta_3^{k,s} &\sim rw(2) \end{aligned}$$

Model C

Previous two models use a naive smoothing structure on $(longitude_i, latitude_i)$ which assumes the additive structure of the bivariate smoothing function. This doesn't take into consideration the correlation between longitude and latitude, which is indeed important and shouldn't be ignored. It is more reasonable to use a more complex bivariate smoothing function. In Greco, Ventrucci, and Castelli (2018), the authors suggest the following smoothing structure:

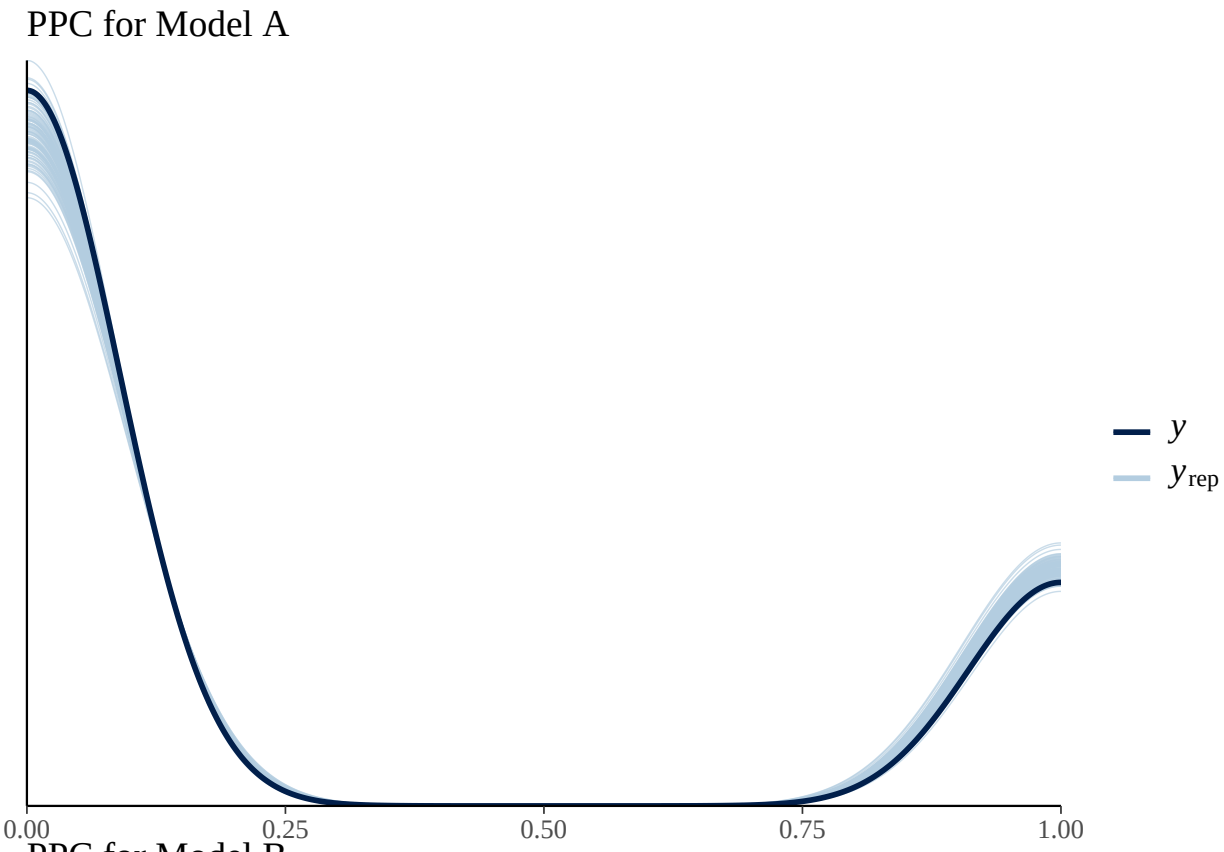
$$f(longitude_i, latitude_i) = \sum_{q=1}^{K_2}\sum_{l=1}^{K_1} b_l(longitude_i) b_q(latitude_i)\beta_{l,q}$$

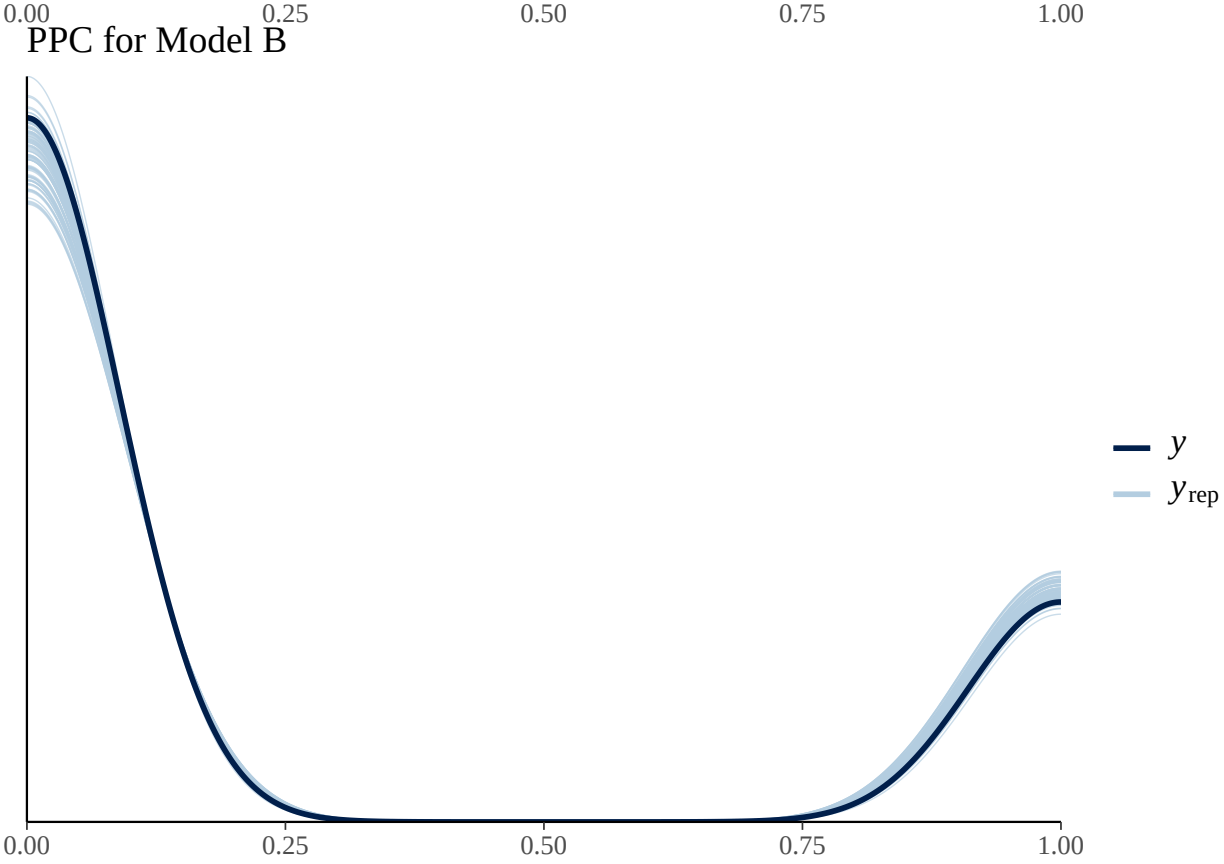
where $b_l(longitude_i)b_q(latitude_i)$ is the tensor product of marginal B-splines evaluated at $(longitude_i, latitude_i)$. This is already implemented in `R` package `mgcv`, and thus Model C will be fitted with `mgcv` using the smoothing structure discribed above, and everything else similar to Model A and Model B. Besides, we also include an additional covariate `AcresBurned` to Model C, as the acres burned by a fire incident is intuitively related to the severity.

Results

Model checks and visualization

In this subsection, we apply various model checking techniques on the models fitted above and provide visualizations when applicable. First, we conduct posterior predictive check (PPC) on Model A and Model B. PPC means comparing the predictive distribution y_{rep} to the observed data y, it can be used to look for systematic discrepancies between real and simulated data. Below are the posterior predictive check plots for Model A and Model B:





Both PPC plots look fine as they are all close to the observed data, indicating that both Model A and Model

Table 2: Smoothing term estimates for Model C

term	edf	ref.df	statistic	p.value
te(Longitude,Latitude)	11.322394	13.508086	80.57217	<0.001
s(ArchiveYear)	1.907415	1.991101	28.61008	<0.001

B did a reasonably good fit to the dataset. Next, there are some diagnostics and visualization for the GAM model fitted by `mgcv` (Model C):

Below are the plots for the output from Model C. The first panel is a plot of the tensor product of longitude and latitude, where the lighter color indicates major fire incident; the second panel is the effect of archive year, it indicates the number of major incidents is decreasing by year; and finally the last panel shows the effect of acres burned, it is quite clear that major incidents will have more acres burned compared to non-major incidents.

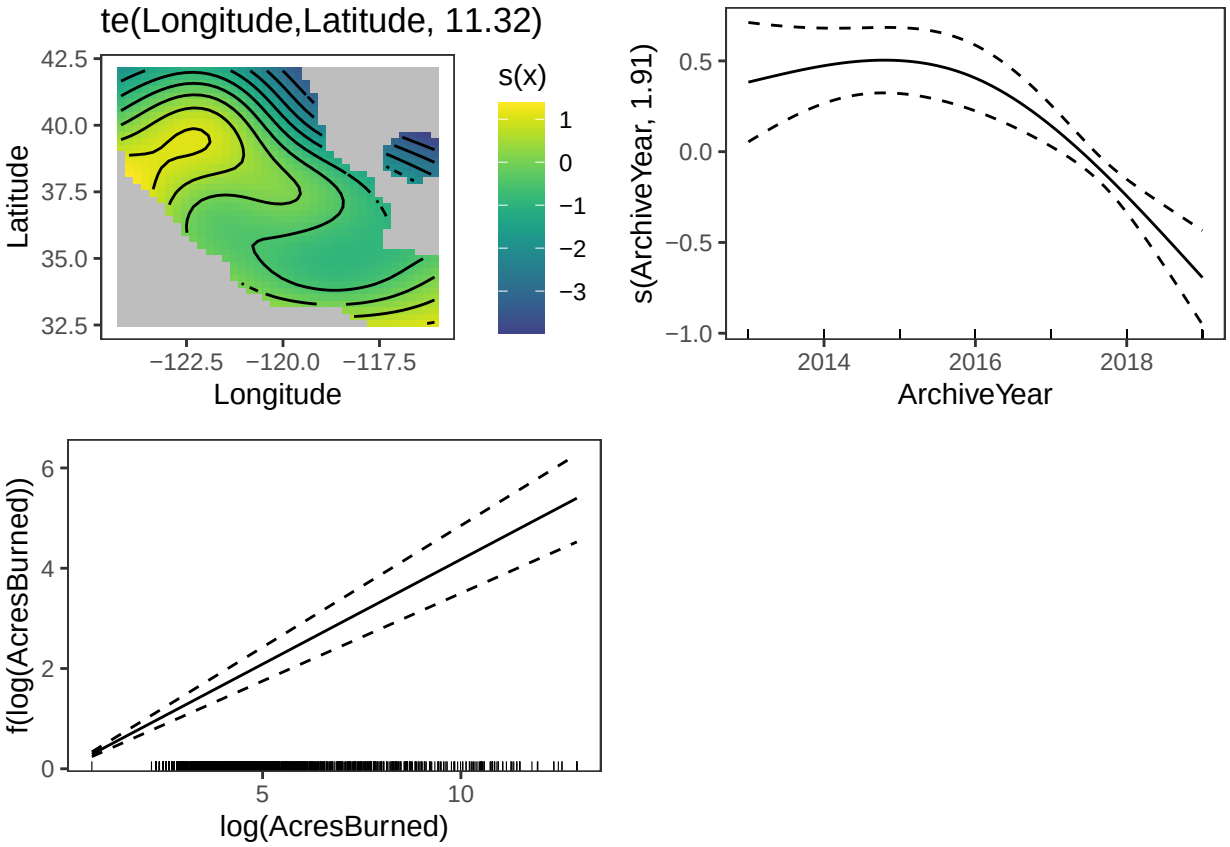
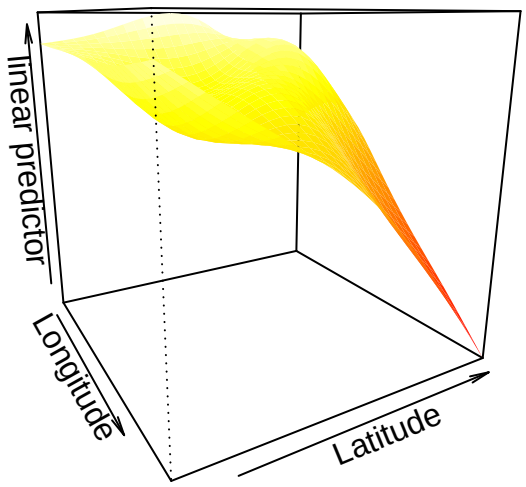


Table 3: Comparison of ELPD

	elpd_diff	se_diff	elpd_loo	se_elpd_loo	p_loo	se_p_loo	looic	se_looic
model2	0.000000	0.000000	-737.0943	19.12850	62.57150	2.466737	1474.189	38.25700
model1	-7.982909	4.116007	-745.0772	19.15689	44.03791	1.933006	1490.154	38.31379

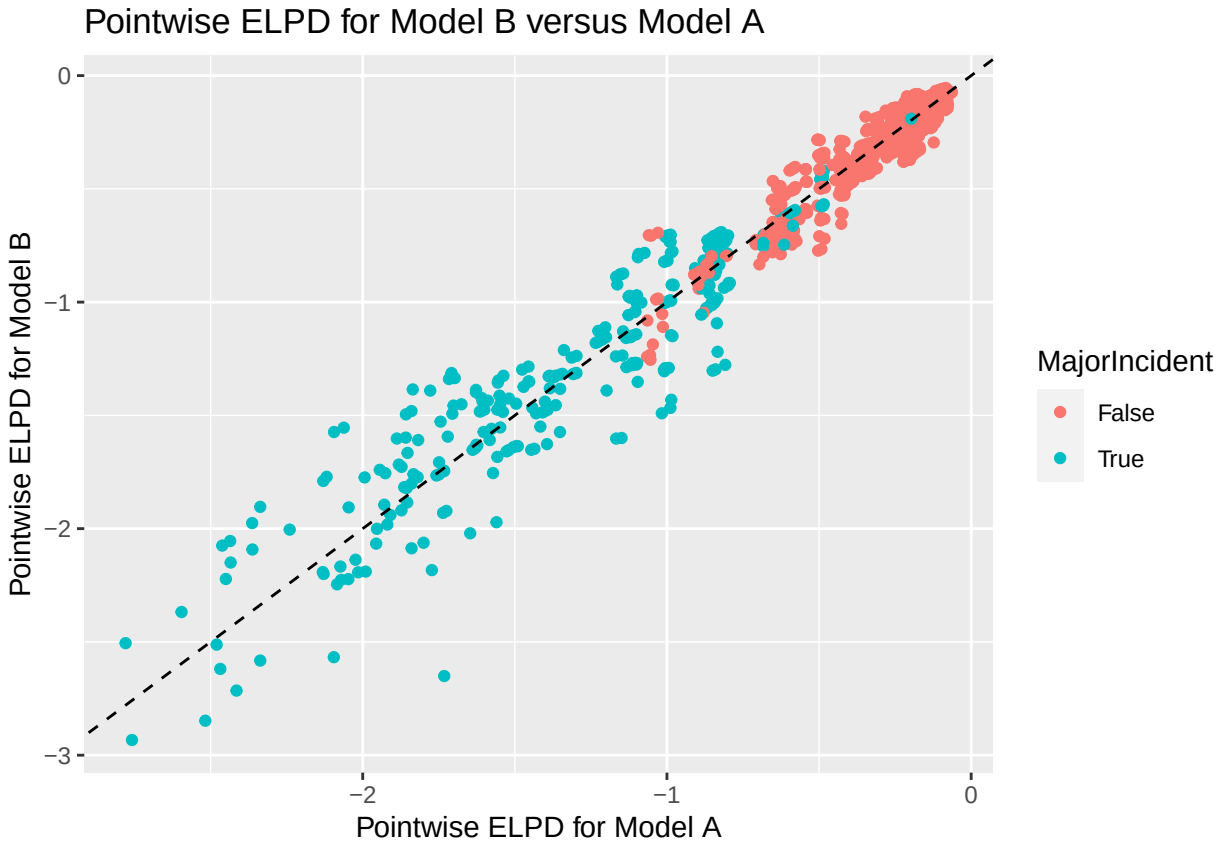
3D visualization of smoothed longitude and latitude



Model comparison

First, to compare the two Bayesian models Model A and Model B, we will make use of the expected log pointwise predictive density (ELPD).

It is clear that Model B has a larger $\sum_i ELPD_i$ (the difference is about 7.39). Therefore, based on this model checking method, we would prefer Model B over Model A. In addition, the pointwise ELPD for these two models is also provided below:



The $ELPD_i$'s seem to line up reasonably well between the two models at larger values, regardless of whether the fire incident is major or not. However, the models begin to disagree for lower values of $ELPD_i$, given that the particular wildfire is a major incident; the pattern becomes obvious for values between -2 and -1. Model A tends to perform better than Model B, since a slightly more number of those extreme values are below the diagnoal line. In other words, when predicting the probability of major incident for the wildfires that actually are major incidents, Model A is more certain about the outcome; it is therefore less likely to have false negatives.

Predictive accuracy

Finally, we present the predictive accuracy for the major incidents by Model B, in all counties. Overall, the prediction accuracy varies a lot across the counties, and the accuracy is the highest in those counties with low proportion of major fire incidents, which may indicate that longitude, latitude and year are not enough to characterize the major incidents.

Discussion

In this project, a set of logistic GAMs are used to classify the wildfire in California from 2013 to 2019 based on the spatial and temporal attributes. Although various model checking metrics are fine for these models and they do capture some of the trends, the prediction accuracy in many counties are not good enough. This is because the study of wildfire also require extensive knowledge of the underlying ecosystem, like for example the fuel type, availability of water source, and human activity are all important (Vilar et al. (2010)). Due to the limitation of available dataset (most of the variables provided are descriptive), these variables are not included in the current models. Besides, since the dataset has only 6 years available, it would be also interesting to collect more year's data in the future in order to forecast the number of wildfires in particular regions in the following years.

Table 4: Predicted proportion of major incidents vs actual proportion, and the prediction accuracy for Model B

Counties	actual	predict_modB	accuracy
Alameda	0.0740741	0.0000000	0.9259259
Amador	0.5454545	1.0000000	0.5454545
Butte	0.3750000	0.1250000	0.7142857
Calaveras	0.4500000	0.3000000	0.5500000
Colusa	1.0000000	1.0000000	1.0000000
Contra Costa	0.1111111	0.0000000	0.8888889
Del Norte	0.0000000	0.0000000	1.0000000
El Dorado	0.4411765	0.0000000	0.5588235
Fresno	0.2264151	0.0000000	0.7735849
Glenn	0.5000000	0.1666667	0.6666667
Humboldt	0.2500000	0.0000000	0.7500000
Inyo	0.6000000	1.0000000	0.6000000
Kern	0.0727273	0.0000000	0.9272727
Kings	0.2000000	0.0000000	0.8000000
Lake	0.4545455	0.2045455	0.5681818
Lassen	0.0625000	0.0000000	0.9375000
Los Angeles	0.1538462	0.0000000	0.8461538
Madera	0.3225806	0.0000000	0.6774194
Marin	0.1666667	0.0000000	0.8333333
Mariposa	0.3913043	0.0869565	0.6956522
Mendocino	0.3043478	0.0000000	0.6956522
Merced	0.0000000	0.0000000	1.0000000
Mexico	0.0000000	0.0000000	1.0000000
Modoc	0.1071429	0.0000000	0.8928571
Mono	0.0000000	0.0000000	1.0000000
Monterey	0.1428571	0.0000000	0.8571429
Napa	0.6363636	0.8636364	0.7727273
Nevada	0.2941176	0.0000000	0.7058824
Orange	0.1000000	0.0000000	0.9000000
Placer	0.1538462	0.0000000	0.8461538
Plumas	0.0000000	0.0000000	1.0000000
Riverside	0.1269841	0.0000000	0.8730159
Sacramento	0.0000000	0.0000000	1.0000000
San Benito	0.1666667	0.0000000	0.8333333
San Bernardino	0.2115385	0.0000000	0.7884615
San Diego	0.3333333	0.0000000	0.6666667
San Joaquin	0.0000000	0.0000000	1.0000000
San Luis Obispo	0.1694915	0.0000000	0.8305085
San Mateo	0.3333333	0.0000000	0.6666667
Santa Barbara	0.1538462	0.0000000	0.8461538
Santa Clara	0.0588235	0.0000000	0.9411765
Santa Cruz	0.5000000	0.2500000	0.7500000
Shasta	0.4333333	0.2333333	0.6000000
Sierra	0.0000000	0.0000000	1.0000000
Siskiyou	0.1739130	0.0000000	0.8260870
Solano	0.3529412	0.0588235	0.7058824
Sonoma	0.5625000	1.0000000	0.5625000
Stanislaus	0.0000000	0.0000000	1.0000000
State of Nevada	0.0000000	0.0000000	1.0000000
State of Oregon	0.0000000	0.0000000	1.0000000
Sutter	0.0000000	0.0000000	1.0000000
Tehama	0.2244898	0.0000000	0.7755102
Trinity	0.2500000	0.0000000	0.7500000
Tulare	0.1612903	0.0000000	0.8387097

Reference

Greco, Fedele, Massimo Ventrucci, and Elisa Castelli. 2018. "P-Spline Smoothing for Spatial Data Collected Worldwide." *Spatial Statistics* 27. Elsevier: 1–17.

Vilar, Lara, Douglas G Woolford, David L Martell, and M Pilar Martı'n. 2010. "A Model for Predicting Human-Caused Wildfire Occurrence in the Region of Madrid, Spain." *International Journal of Wildland Fire* 19 (3). CSIRO Publishing: 325–37.

Woolford, Douglas G., David L. Martell, Colin B. McFayden, Jordan Evens, Aaron Stacey, B. Michael Wotton, and Dennis Boychuk. 2021. "The Development and Implementation of a Human-Caused Wildland Fire Occurrence Prediction System for the Province of Ontario, Canada." *Canadian Journal of Forest Research* 51 (2): 303–25. https://doi.org/10.1139/cjfr-2020-0313.