Title
Estimation of Wildfire Ignition Conditional Intensity Parameters Via the Stoyan-Grabarnik Statistic

Author
Coelho, Abigail Bernice

Publication Date
2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Estimation of Wildfire Ignition

Conditional Intensity Parameters

Via the Stoyan-Grabarnik Statistic

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

Abigail Coelho

2023

ABSTRACT OF THE THESIS

Estimation of Wildfire Ignition
Conditional Intensity Parameters
Via the Stoyan-Grabarnik Statistic

by

Abigail Coelho
Master of Science in Statistics
University of California, Los Angeles, 2023
Professor Frederic R. Paik Schoenberg, Chair

Point process modeling is used to predict the conditional intensity of wildfire ignitions in California from 2008 to 2012. Weather variables from the closest station record are used as model covariates. To reduce computation requirements of maximum likelihood estimation, the Stoyan-Grabarnik statistic is employed. Models of various combinations and number of variables are examined, as well as models for specifically naturally occurring fires caused by lightning and fires caused by humans.

The thesis of Abigail Coelho is approved.

Mark Stephen Handcock

Chad J. Hazlett

Yingnian Wu

Frederic R. Paik Schoenberg, Committee Chair

University of California, Los Angeles

2023

For James, my constant support

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

Point process modeling has been shown to be a useful method to estimate wildfire hazard with weather variables [2], but estimation via maximum likelihood is computationally expensive. This paper instead utilizes the Stoyan-Grabarnik statistic to reduce the computational requirements of model fitting. This paper seeks to estimate the parameters of linear point process models to predict the intensity of the ignition of wildfires. Parameter estimation begins with models with many variables and reduces to essential variables. Each model is evaluated by the Stoyan-Grabarnik statistic value and by comparing the observed rate of fires ignited to the estimated intensity for random space-time locations.

CHAPTER 2

Data

2.1 Wildfire

The fire occurrence dataset is from the USDA Forest Service website for 2005 to 2012 in California [4]. The dataset includes all known fires of any size ignited on National Forest System Lands, and the variables include the ignition date, coordinates, and cause. The categorical cause variable includes lightning and human-causes, such as campfire, equipment use, smoking, children. Using this variable, the data is divided into two additional datasets for natural fires—caused by lightning—and unnatural fires—caused by humans.

2.2 Weather

The daily weather data is from the National Oceanic and Atmospheric Administration for weather stations in California from 2008 to 2012 [3]. The variables of interest in the datasets are daily maximum temperature, minimum temperature, average wind speed, evaporation, and precipitation. Of the 1,233 weather stations recording, not every weather station records every variable (table 2.1), and not every weather station records every day.

Since the accumulation of precipitation likely has a stronger impact on the rate of fires, the precipitation variable is transformed into a rolling average of precipitation:

$$y_t = 0.4 * x_t + 0.3 * x_{t-1} + 0.2 * x_{t-2} + 0.1 * x_{t-3}$$

Extreme, nonsensical weather values are removed from the data for the temperature and average wind speed variables. For example, a maximum temperature recording of 150 degrees Fahrenheit indicates a data entry or reading error as the maximum temperature recorded in California is 134 [5]. For maximum and minimum temperatures, values beyond the 0.001 and 0.999 percentiles were removed. For average wind speed, values beyond the 0.9999 percentile of the dataset were removed.

Weather Variable	Num of Stations Recording
Average Wind Speed (AWND)	83
Maximum Temperature (TMAX)	724
Minimum Temperature (TMIN)	724
Evaporation (EVAP)	17
Rolling Precipitation (PRCP_ROLLING)	930

Table 2.1: Number of stations recording each weather variable

2.3 Merging

To match ignited fires with the closest weather reading, each weather variable is taken from the closest weather station with a reading for that variable on the day the fires was ignited. This process is completed for each weather variable separately because not every station has every variable (table 2.1). For example, only 83 weather stations have readings for average wind speed, but 724 stations have readings for temperature.

To match each of the over 7,000 fires to the closest weather reading, a less computationally intensive method than calculating the distance to each of the hundreds of weather stations. Therefore, k-dimensional trees for each fire ignition are used to find the closest weather station with the variable's measurement on the ignition day. A k-d tree is a space-partitioning data structure, which allows for quicker determination of the closest point for each fire. Instead of calculating every distance between fires and weather stations, the points (the weather station coordinates) are arranged with respect to the coordinates of the fires.

California is divided into 100 regions by dividing the minimum and maximum latitude and longitude of California into 10 evenly spaced bins. Each fire is assigned to the region in which it occurred. For each region, the background rate was calculated using 5,036 fires from 2005 – 2007. These years are otherwise not used in the estimation of the model, so that the background rate and covariate effects may be estimated separately, as in Ogata [1]. The remaining 7,179 fires are used for model estimation. The background rate is the min-max normalization of the count of wildfires ignited in each region during the three years.

Of the 100 divided areas of California, 40 regions contained wildfire ignitions in National Forests. This division results in regions which are 0.0119528 megameters squared. Regions without wildfires are excluded from the parameter estimation using Stoyan-Grabarnik estimation, as the value of the Stoyan-Grabarnik statistic is undefined for such regions.

2.4 Baseline

In order to compare the estimated spatial-temporal pattern of wildfires, a baseline comparison dataset was formed using simulated uniformly randomly selected locations in each region. For each region with fires, 100 random points and days were generated, and then, in the same manner as matching the fire coordinates to the closest weather reading, matching weather variables were assigned to each point.

CHAPTER 3

Methods

The models considered evaluate the spatial-temporal conditional intensity at time t for location (x, y) as the background rate $m(x, y)$ plus a linear combination of the weather covariates $W(x, y, t)$.

$$\lambda(t, x, y) = \gamma m(x, y) + \alpha W(t, x, y)$$

In a sequential, step-wise fashion, parameter estimation begins by testing a model using all the independent variables to predict the intensity. Sequential models remove inconsequential variables to improve upon the SG statistic and mean absolute residual.

This paper aims to estimate the conditional intensity parameters using the Stoyan-Grabarnik (SG) statistic. Optimizing for SG allows the point process model to be estimated without calculating computationally expensive integrals [6]. SG is the inverse of the conditional intensity.

$$\overline{m} = \frac{1}{\lambda}$$

In order optimally to estimate the model parameters, the sum of squares method over a partition was used as suggested in [6]. For this simplified Stoyan-Grabarnik estimator, $|I_j|$ is the size of the partition, which in this paper is the number of years times the size of the regions in megameters squared.

$$\hat{\theta}_0 = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{j=1}^{p} \left(\sum_{i:\tau_i \in I_j} \frac{1}{\lambda(\tau_i; \theta)} - |I_j| \right)^2$$

To evaluate the estimated parameters, the SG statistic and a numerical approximation of the integral of the lambda over the square are examined. The approximation of the integral is calculated by estimating the conditional intensity of the randomly chosen points for each region. Additionally, the residual of the point process model is evaluated by subtracting the integral approximation from the observed rate of fires within each region.

The order of the initial parameter estimates for each model is chosen by fitting the model for orders from 100,000 to 0.1. The order with the lowest SG value is used for all the model's initial parameter estimates.

In addition to comparing models using the SG and residual statistics, a constant model is fitted to the dataset for comparison, estimating the only parameter β.

$$\lambda(t, x, y) = \beta$$

To reduce variance due to different climates, modeling is repeated for Southern California as defined by the middle value between the maximum and minimum latitude of California.

CHAPTER 4

Results

4.1 Rate of Ignited Fires

To comparing the effectiveness of proceeding models, the constant model is tested. The constant model produces an SG value of 0.074 (table 4.1) and a high mean absolute residual value of 4,087. Note that the median rate of ignited wildfires is about 2,600 per megameter squared per year.

Variables	SG	Mean Absolute Residual	Fail to Converge
BKGD_RT	0.01906	610	
BKGD_RT, EVAP	0.05200	1,328	
BKGD_RT, TMAX	0.07421	2,420	
BKGD_RT, TMAX, EVAP	0.07010	2,426	
BKGD_RT, AWND, TMAX, EVAP	0.07456	3,661	
BKGD_RT, AWND, TMAX, TMIN, EVAP, PRCP_ROLLING	0.07389	4,070	yes
constant	0.07438	4,087	
BKGD_RT, AWND, TMAX, EVAP, PRCP_ROLLING	0.06563	23,073	yes
BKGD_RT, AWND, TMAX	0.07270	42,745	

Table 4.1: Results of estimating models for ignited fire rates

After the constant model, the first model tested uses all the weather variables. However, this model does not converge. The final, non-converged parameter estimates of this model produce statistics very similar to the constant model: an SG value of 0.074 and mean absolute residual of 4,087 (table 4.1). As the model does not converge, the parameter estimates do not accurately detect regions with high rates of fires and those with low. Heatmaps in figure 4.1

show predicted regions in Northern California (top right, purple) have low rates compared to the rest of California, but the observed heatmap (top left, green) shows moderate or high rates in the north. The absolute residual heatmap shows Southern California as the region with the poorest estimation (bottom left, grey). Furthermore, the residuals density plot show many overestimates and are not centered around 0 (bottom right).

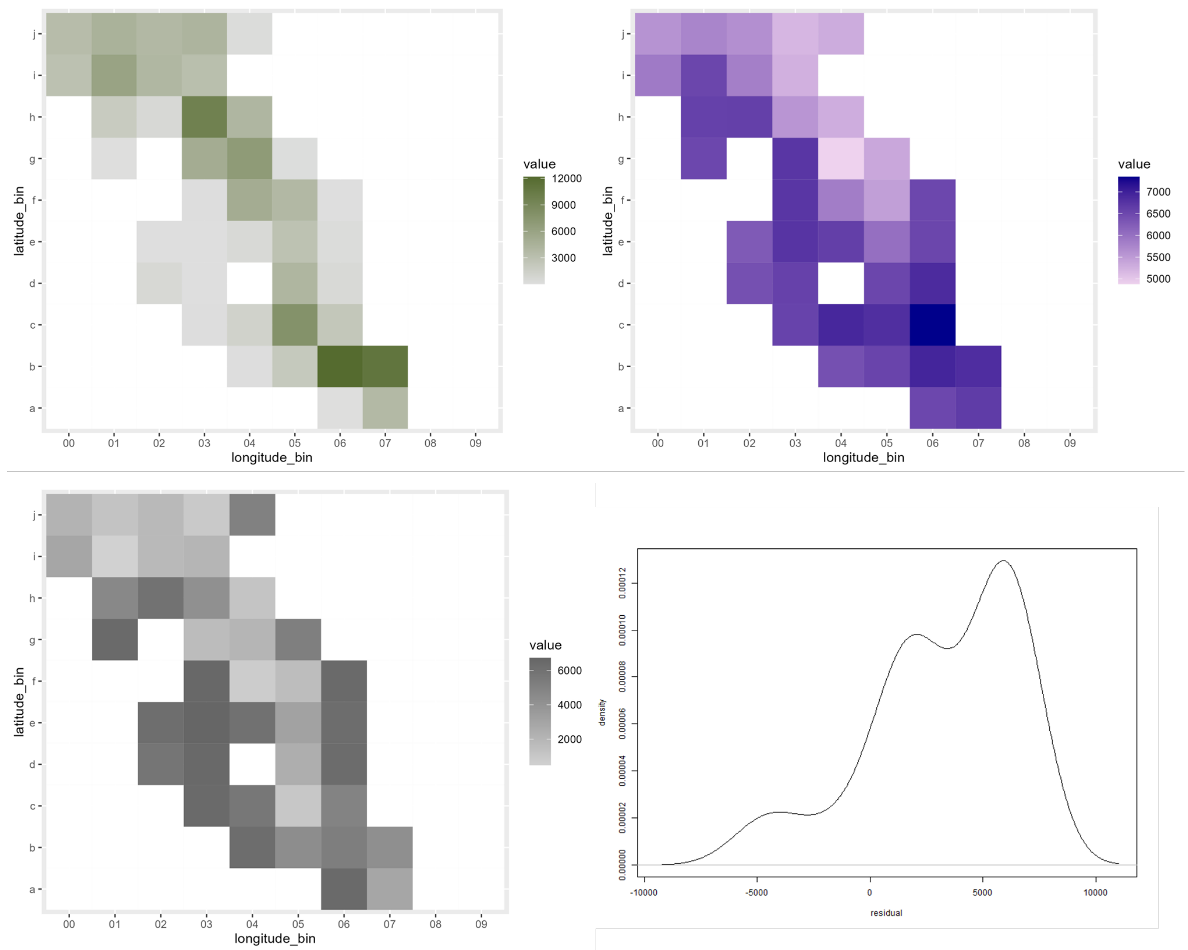

Figure 4.1: Heatmaps of observed, predicted, and absolute residual ignition rate and region residual values density plot for model using every weather covariate

Modeling continues in a step-wise function. For the next model tested, minimum temperature is removed from modeling because it strongly correlates to maximum temperature (figure 4.2), which in the first model was given a larger parameter estimate of about 73 compared to 22 for minimum temperature. Additionally, maximum temperature is assumed to be of more importance to igniting wildfires than the day's minimum temperature. The second model of the average wind speed, maximum temperature, evaporation, and rolling precipitation still failed to converge. Although the same initial parameter order of 100 is used for this model as for the first model, parameter estimates produce a mean absolute residual value of 23,000, but the SG value is minimized to a smaller value of 0.066 (table 4.1).

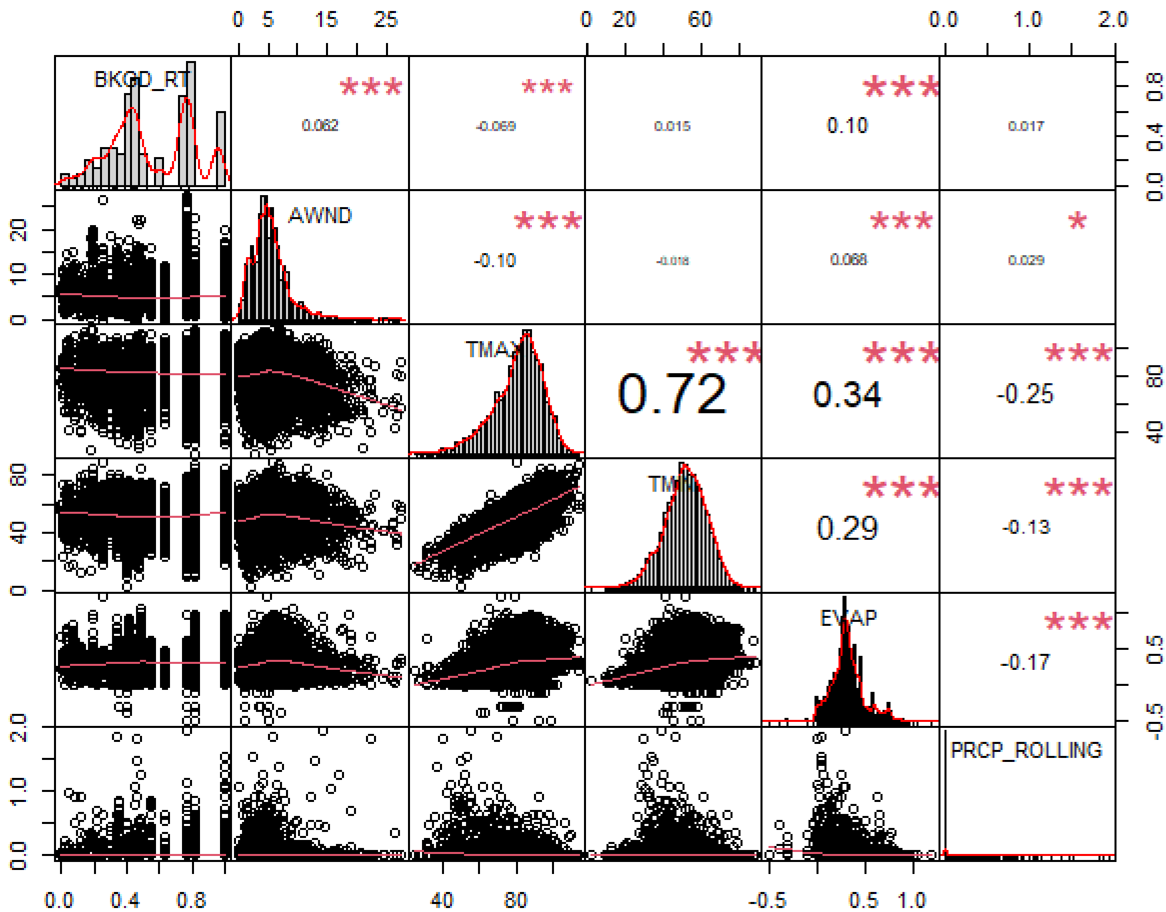

Figure 4.2: Correlation matrix chart of variables

After the model using four weather covariates, rolling precipitation is removed, since rolling precipitation is often zero within the observations (figure 4.2). At the 75th percentile, the rolling precipitation is zero, and at the 80th percentile, 0.008 inches of rain.

The proceeding models utilize a combination of the three remaining weather variables: average wind speed, maximum temperature, and evaporation. The remaining models all converge. A model using maximum temperature and evaporation produces an SG value of 0.070, improving upon the constant model's value of 0.074. Furthermore, the mean absolute residual value for this model is 2,426, which is about a 40% decrease (table 4.1). In comparison, the models which include average wind speed as a value produce a mean absolute residual value greater than the constant model, and about the same SG value but with only a 10% decrease in mean absolute residuals (table 4.1). Based on these results, average wind speed is not used as a covariate in the next models.

The model producing the most optimal results with historical weather uses evaporation as the only covariate. The parameter estimation for this model produces an SG statistic of 0.052, about a 30% decrease from the constant model, and a mean absolute residual of 1,328, a 68% decrease (table 4.1). Further examining, the predictions of this model creates a heatmap of the predicted intensity (figure 4.3, top right, purple) similar to the heatmap of the observed intensity (top left, green). However, the intensity predictions range up to only about 8,000 fires per megameter per year whereas the maximum observed rate range is about 12,000; the absolute residual heatmap shows the magnitude of this underestimate (bottom left). The density plot of the residuals appear to be centered around zero (bottom right), but is shifted right, again, due to the underestimate in magnitude.

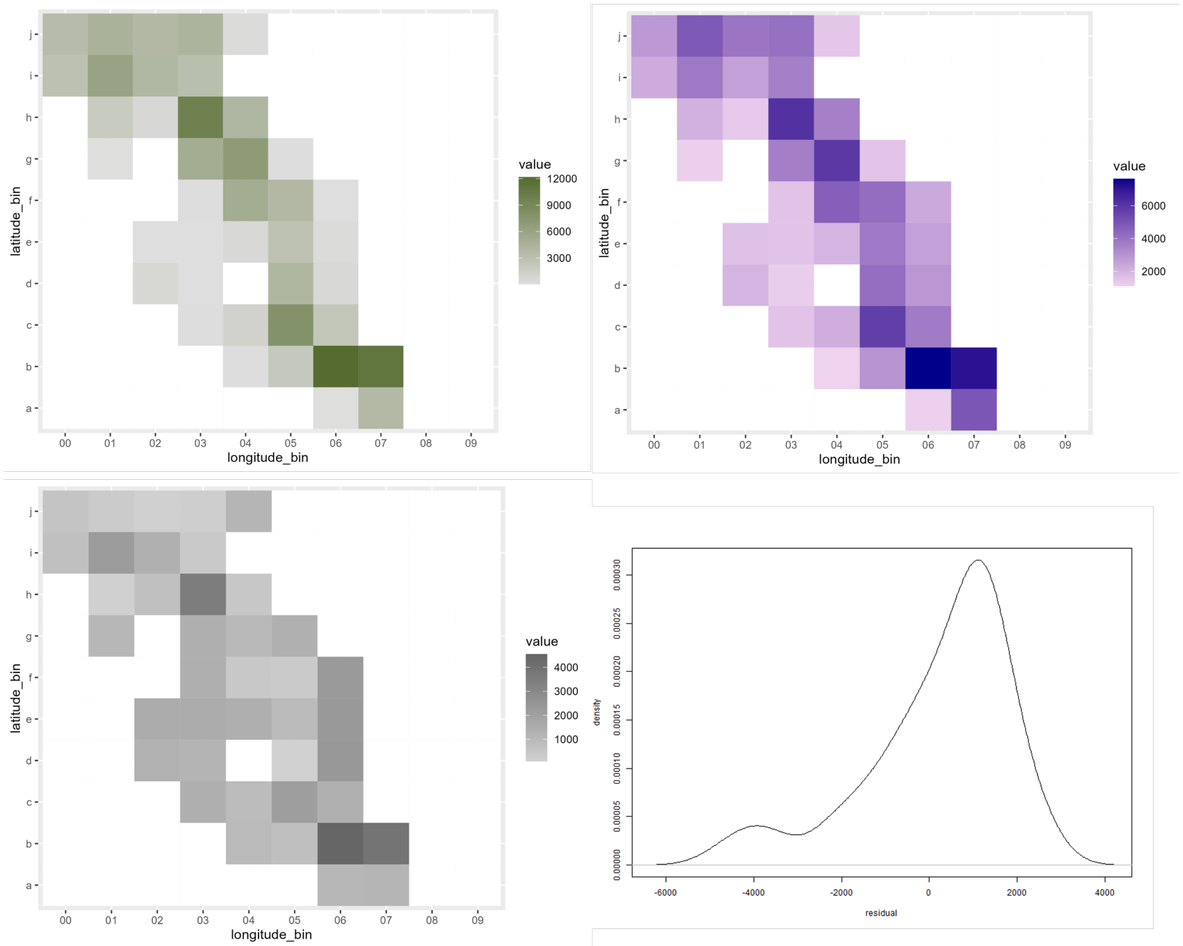

Figure 4.3: Heatmaps of observed, predicted, and absolute residual ignition rate and region residual values density plot for model of all fires, using background rate and evaporation covariates

Removing all the weather variables and leaving only the background rate, which is used to control for the historical rate, produces the best results of all the models. This model has an SG value of 0.019, about a 75% improvement from the constant model, and mean absolute residual value of 610, an 85% decrease from the constant model (table 4.1). The heatmaps of the observed and predicted rates appear to have similar patters in high intensity regions and low intensity regions (figure 4.4, top). However, for this model, the magnitude of the rate is better captured, hence the residuals are closer to zero and there is less of a shift in the density plot towards the right towards underestimate residuals (figure 4.4, bottom right).

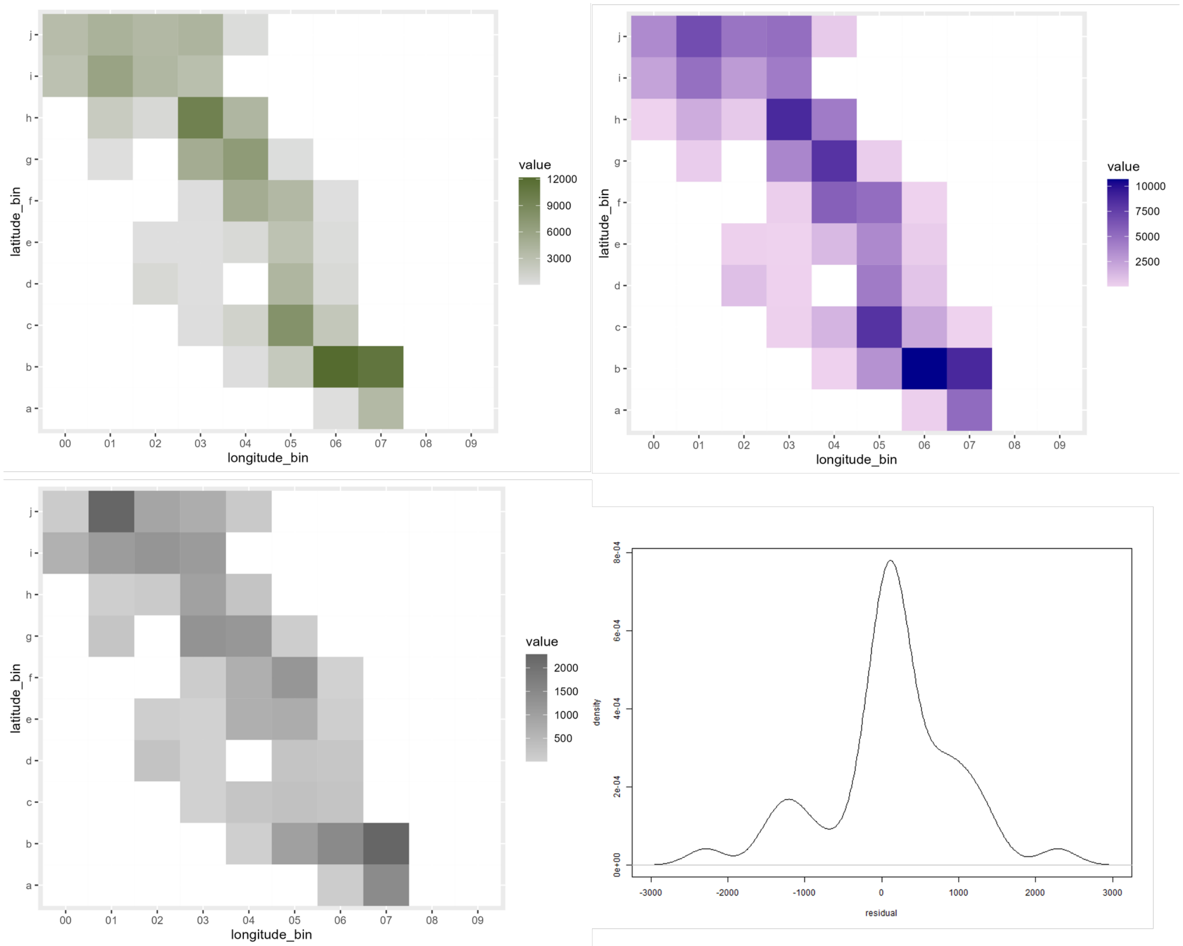

Figure 4.4: Heatmaps of observed, predicted, and absolute residual ignition rate and region residual values density plot for model of all fires, using only background rate

4.2 Rate of Natural vs. Human Ignited Fires

The models for estimating the rate of naturally ignited fire rates have a median mean absolute residual of 1,368 fires per megameter squared per year (table 4.2), which is about 40% smaller than the median value of 2,312 for human-caused fires (table 4.3). The models of naturally ignited fires' median SG value of 0.048 is also smaller than human-caused fires, 0.082. The dataset contains 2,690 observations of natural wildfires, which is 40% less than the 4,489 observations of human-caused fires.

Variables	SG	Mean Absolute Residual	Fail to Converge
BKGD_RT, EVAP	0.03975	844	
BKGD_RT	0.02632	852	
BKGD_RT, TMAX	0.04773	1,028	
BKGD_RT, AWND, TMAX	0.04818	1,151	
constant	0.04800	1,368	
BKGD_RT, AWND, TMAX, EVAP, PRCP_ROLLING	0.04817	1,387	
BKGD_RT, AWND, TMAX, TMIN, EVAP, PRCP_ROLLING	0.04833	1,478	yes
BKGD_RT, AWND, TMAX, EVAP	0.04694	1,618	
BKGD_RT, TMAX, EVAP	0.04773	3,828	

Table 4.2: Results of estimating models for naturally ignited fire rates

Variables	SG	Mean Absolute Residual	Fail to Converge
BKGD_RT	0.03760	800	
BKGD_RT, EVAP	0.05740	1,299	
BKGD_RT, TMAX	0.08188	1,665	
BKGD_RT, TMAX, EVAP	0.08446	1,803	
BKGD_RT, AWND, TMAX	0.08414	2,312	
BKGD_RT, AWND, TMAX, EVAP, PRCP_ROLLING	0.07081	2,597	yes
BKGD_RT, AWND, TMAX, EVAP	0.08300	2,710	
BKGD_RT, AWND, TMAX, TMIN, EVAP, PRCP_ROLLING	0.08063	3,275	
constant	0.08523	3,639	

Table 4.3: Results of estimating models for human-caused ignited fire rates

The model using only evaporation preforms the best of the point process models with weather as a covariate for both causes of wildfires. For the naturally occurring fires, this model has a mean absolute residual value of 844, smaller than the 852 mean for the model of only the background rate (table 4.2). The SG value is about 50% higher for the evaporation model of natural fires than model with using only the background rate.

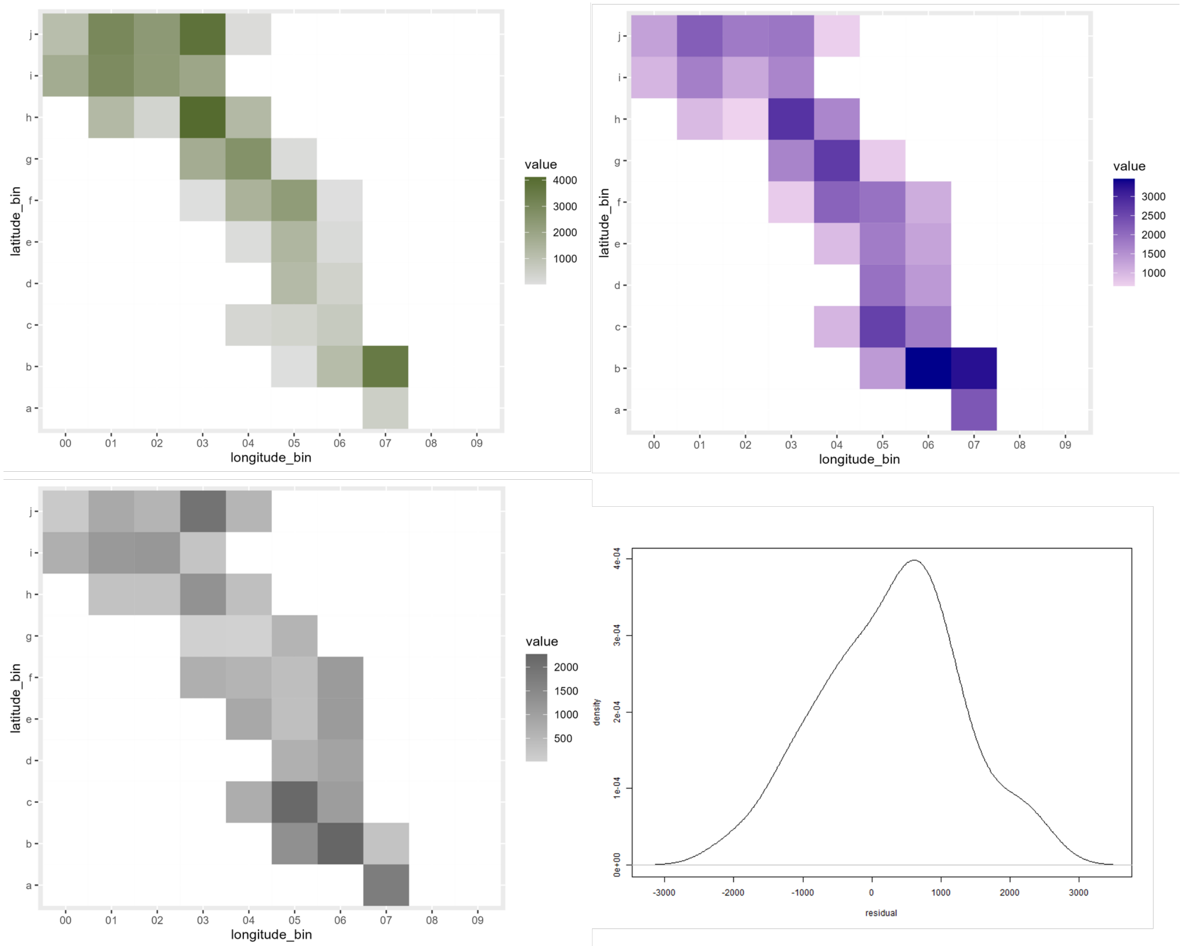

Figure 4.5: Heatmaps of observed, predicted, and absolute residual ignition rate and region residual values density plot for model of natural fires, using background rate and evaporation covariates

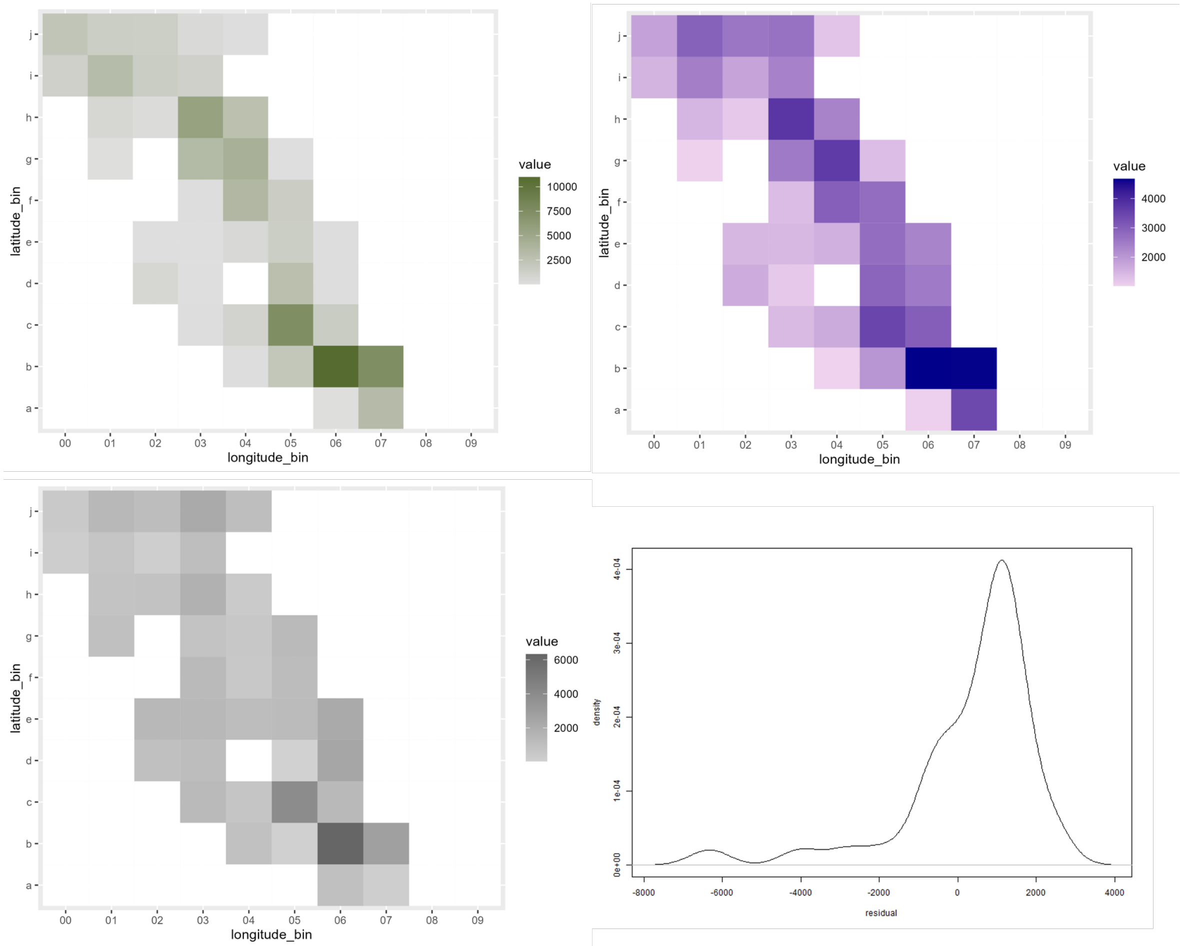

Figure 4.6: Heatmaps of observed, predicted, and absolute residual ignition rate and region residual values density plot for model of human-caused fires, using background rate and evaporation covariates

The heatmaps of the observed and the predicted rates appear to have similar patterns in areas with high rates of fires compared to areas with low rates of fires (figure 4.5, 4.6), for both natural and human-caused evaporation models. However, again for both portions of the observations, this model does not fully capture the range of the rate values. The evaporation model for natural fires does not contain values as high as the observed rate of 4,000 fires per megameter year. Some estimates are about 2,000 smaller in magnitude than the observed intensities (bottom left).

The same model for human-caused fires also does not capture the full magnitude of the rate (figure 4.6). The density plot of the residuals for this human-caused fire model has a large tail for the overestimates (bottom right). Examining the observed and predicted heatmaps of California (top left, right), Northern California is overestimated compared to the rest of California. This model overestimates by up to 6,000 in the fire rate (bottom left), but more commonly by about 2,000.

4.3 Rate of Ignited Fires in Southern California

For estimating Southern California ignited fire rates, the model using average wind speed and maximum temperature preforms similarly to the model using only evaporation for the weather covariates (table 4.4). The SG values are both about 0.032, but the mean absolute residual value is about 33% higher for the model with average wind speed and maximum temperature. This improved performance for the average wind speed and maximum temperature model is also seen in the model for natural fires (table 4.2). Natural fires are about 20% of the 2,982 fires in Southern California, compared to 37% of the fires in all of California.

Variables	SG	Mean Absolute Residual	Fail to Converge
BKGD_RT	0.01094	508	
BKGD_RT, EVAP	0.03222	1,312	
BKGD_RT, AWND, TMAX	0.03170	1,748	
BKGD_RT, TMAX	0.04400	2,780	
BKGD_RT, AWND, TMAX, EVAP, PRCP_ROLLING	0.03561	3,103	yes
BKGD_RT, AWND, TMAX, EVAP	0.04471	4,661	
constant	0.04477	5,931	
BKGD_RT, AWND, TMAX, TMIN, EVAP, PRCP_ROLLING	0.03434	11,381	yes
BKGD_RT, TMAX, EVAP	0.03205	117,570	

Table 4.4: Results of models for all fires in Southern California

The model using only evaporation for weather covariates preforms the best of all the models. The density plot of the residuals for this model shows that the rate estimates are often around 1,000 overestimated or underestimated (figure 4.7), which is a smaller range of residuals than for previous models (figure 4.3, 4.5) and similar to the common residual values for human-caused fires (figure 4.6).

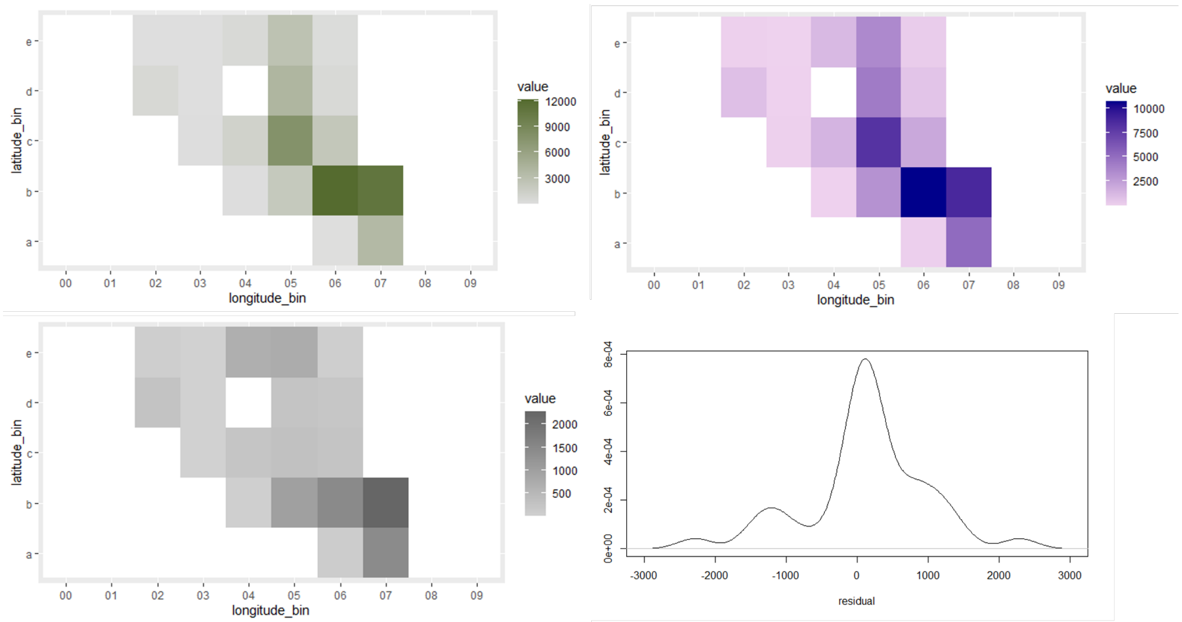

Figure 4.7: Heatmaps of observed, predicted, and absolute residual ignition rate and region residual values density plot for model of Southern California fires, using background rate and evaporation covariates

For the models of naturally occurring fires in Southern California, two models with weather covariates preform similarly to the model using only the background rate (table 4.5). The SG statistic is about 0.015 for both the background model and the model including evaporation, which has a mean absolute residual of 800, 23 less than the background model. There is a jump in the mean absolute residuals to 1,364 for the constant model, but the constant model preforms better than the remaining models.

Variables	SG	Mean Absolute Residual	Fail to Converge
BKGD_RT, TMAX	0.01724	564	
BKGD_RT, EVAP	0.01459	800	
BKGD_RT	0.01458	823	
constant	0.02378	1,324	
BKGD_RT, AWND, TMAX, TMIN, EVAP, PRCP_ROLLING	0.01871	1,616	yes
BKGD_RT, TMAX, EVAP	0.02335	2,309	
BKGD_RT, AWND, TMAX, EVAP	0.01696	5,503	
BKGD_RT, AWND, TMAX	0.02182	8,249	
BKGD_RT, AWND, TMAX, EVAP, PRCP_ROLLING	0.01257	31,105	yes

Table 4.5: Results of models for natural fires in Southern California

Variables	SG	Mean Absolute Residual	Fail to Converge
BKGD_RT	0.01699	499	
BKGD_RT, EVAP	0.03382	1,374	
BKGD_RT, TMAX	0.04381	2,152	
BKGD_RT, AWND, TMAX, EVAP, PRCP_ROLLING	0.04179	2,790	yes
BKGD_RT, AWND, TMAX	0.04634	4,056	
constant	0.04627	5,148	
BKGD_RT, AWND, TMAX, TMIN, EVAP, PRCP_ROLLING	0.04336	5,584	
BKGD_RT, TMAX, EVAP	0.04630	14,500	
BKGD_RT, AWND, TMAX, EVAP	0.04206	9,497,498	

Table 4.6: Results of models for human-caused fires in Southern California

For the models of human-caused fires in Southern California, the background model preforms the best (table 4.6) with an mean absolute residual of 499 and SG statistic of 0.017. The next highest preforming model includes the evaporation rate, but the SG value is twice as much and the mean absolute residual is almost three times as much.

CHAPTER 5

Discussion

Overall, the models with the smaller number of variables preform better than those with many weather covariates. This is true for each subset of the wildfire dataset, whether split by cause or reduced to only Southern California fires. The background rate model preforms better than models including any weather variables, expect for the two natural subsets of the fire dataset (table 4.2, 4.5). Evaporation and maximum temperature are the weather variables that produce the best results. Additionally, models with four or five weather covariates typically do not converge.

When different models are trained for fires ignited by lighting rather than humans, the natural dataset models perform better than the human-caused dataset models (table 5.1). The naturally occurring fires is the smaller proportion of observed fires, 37% of the 7,179 ignited fires in California. The median mean absolute residual value for natural fires is 41% smaller, 1,368 compared to 2,312 for human-caused; the SG statistic for natural fires is also 41% smaller with a median of 0.048 compared to 0.082 for human-caused. Perhaps this is indicative that naturally occurring fires are easier to predict than those caused by human disruptions, or perhaps because human-caused fires have a wider range of ignited wildfire rates amoung the regions.

Dataset	Size	Model Median	
		SG	Mean Absolute Residual
All	7,179	0.073	3,661
Natural	2,690	0.048	1,368
Human-Caused	4,489	0.082	2,312
Southern California	2,982	0.034	3,103
Southern California, Natural	600	0.017	1,616
Southern California, Human-Caused	2,382	0.043	4,056

Table 5.1: Medians of model results by subset of data

For fires in Southern California, this pattern was again repeated (table 5.1). The median SG rate was 60% smaller at 0.017 compared to 0.043 for human-caused fires. The median mean absolute residual value was also 60% smaller at 1,616 compared to 4,056. In Southern California, naturally occurring fires are 20% of the number of ignited fires.

Similar models preform well for the subset of wildfires in Southern California (table 4.4), but again the SG statistic is smaller at about half the value across the models than the full dataset (table 4.1), yet the median mean absolute residual is only about 500 higher for all of California than for the median value of about 3,100 for Southern California 5.1). This difference in the SG value could be because southern California had a lower median of the count and therefore rate of fires in each region than all of California, making the SG value easier to minimize.

CHAPTER 6

Conclusion

The Stoyan-Grabarnik statistic reduces the computational requirement and produces models better than those fitted to a constant. An increase in the number of model covariates causes modeling performance to decrease. The weather variables that proved to be most valuable were evaporation and maximum temperature. The dataset of natural fires preformed better than the dataset of unnatural fires.

Future studies could utilize interpolation of weather variables along with using the SG statistic. Additionally, the predictability of natural compared to unnatural fires could be further explored by reducing the datasets to avoid model underfitting to wide ranges in intensity between regions. For specific models, the initial parameter estimates should be more closely examined to avoid overfitting to produce the lowest SG without evaluating other statistics.

Furthermore, the uncertainty within the models should be further examined as various models showed deviation in the evaluation statistics. The uncertainty can be measured by bootstrapping the data, and examining the variance in the SG and residual values. Further researching the region size could also lead to more stable evaluation statistics. The various region sizes can be further tested to find the appropriate dimensions. Other methods of dividing the regions could be explored, such as choosing non-uniform region divisions so that each contains the same number of ignited fires.

Bibliography

[1] Y. Ogata. "Space-Time Point-Process Models for Earthquake Occurrences". In: *Annals of the Institute of Statistical Mathematics* 50.2 (1998), pp. 379–402.

[2] H. Xu and F.P. Schoenberg. "Point process modeling of wildfire hazard in Los Angeles County, California". In: *Annals of Applied Statistics* 5.2a (2011), pp. 684–704.

[3] National Oceanic and Atmospheric Association. *Climate Data Online*. data retrieved from National Centers for Environmental Information, `https://www.ncei.noaa.gov/cdo-web/search`. 2022.

[4] USDA Forest Service. *Fire Occurrence FIRESTAT Yearly*. data retrieved from, `http://data.fs.usda.gov/geodata/edw/datasets.php`. 2022.

[5] *Death Valley Weather*. Accessed on March 8, 2023. 2023. URL: `https://www.nps.gov/deva/learn/nature/weather-and-climate.htm`.

[6] C. Kresin and F Schoenberg. "Parametric estimation of spatial-temporal point processes using the Stoyan-Grabarnik statistic". In: *AISM* (to appear).