Katrina Altawil & Zeynep Cetin
12/06/2024
Dr. Jun Ye
Spatial Temporal Statistics 586

Final Project:

Brightness Temperature of Fire across the United States

Part 1: Introduction

This project investigates and analyzes near real-time (NRT) active fire data collected by National Aeronautics and Space Administration's (NASA) Moderate Resolution Image Spectroradiometer (MODIS). NASA collects and aggregates this thermal activity detected by MODIS sensors on the Aqua and Terra satellites, separating based on country and by year [Source 1]. The goal here is to explore the variability between the spatial-temporal characteristics in the data and the brightness temperature of fire, specifically in the United States in the year 2023.

Part 2: Explanation of the Data

The dataset, named "modis_2023_United_States", has a total of 14 variables, including: **latitude** and **longitude**, **brightness temperature (Kelvin) detected by radiation**, along scan and track actual pixel size, **acquisition date** and **time**, satellite (aqua and terra), confidence (low-, nominal-, or high-confidence fire), version (collection and source), brightness temperature from channel 31 (Kelvin), fire radiative power (FRP in megawatts), inferred hot spot type (vegetation, volcano, other static land source, offshore), and day-/night-time. (The variables of interest for spatial-temporal analysis are **bolded**.) The data set, excluding any rows with missing information, has a total of 83,490 observations [Source 2]. In an effort to lower this number and avoid interpolating over the Pacific Ocean, only the continental United States is considered (i.e., longitude values above -130°).

A preliminary look into the dataset reveals the brightness over the entire year has a mean (SD) of 321.37 K (19.02 K) and a median (IQR) of 316.70 K (17.70 K) and has the following distribution:

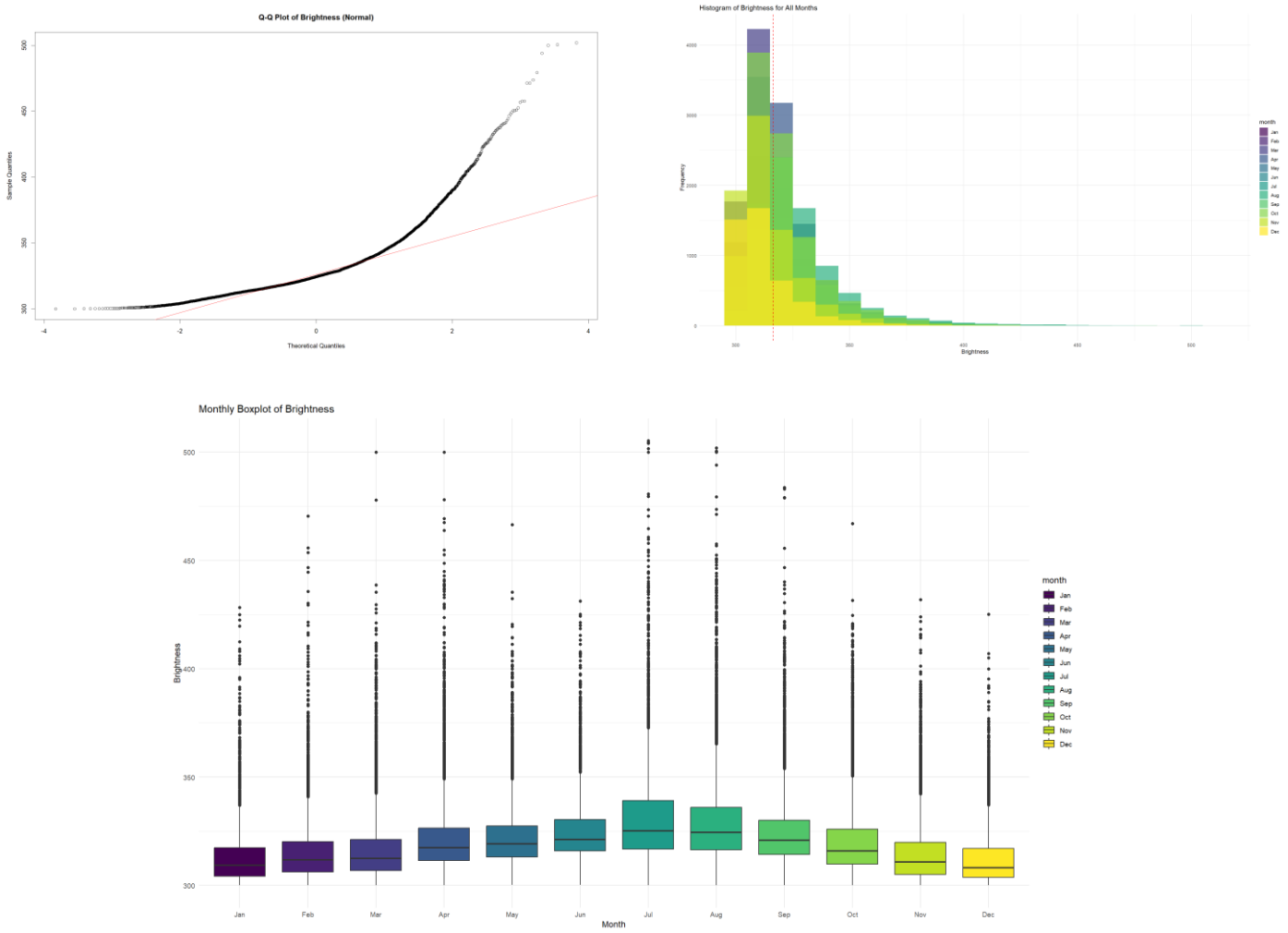

Figure 1. Q-Q plot with theoretical normal line in red (left), histogram with median dotted red line (right), and boxplot (below) by month for brightness of fires in Kelvin in continental USA.

It becomes very obvious that the data is very right skewed, which is affirmed by the mean being higher than the median. As a way of ensuring the outliers are minimized, only brightness Kelvin values at or below 350 are considered:

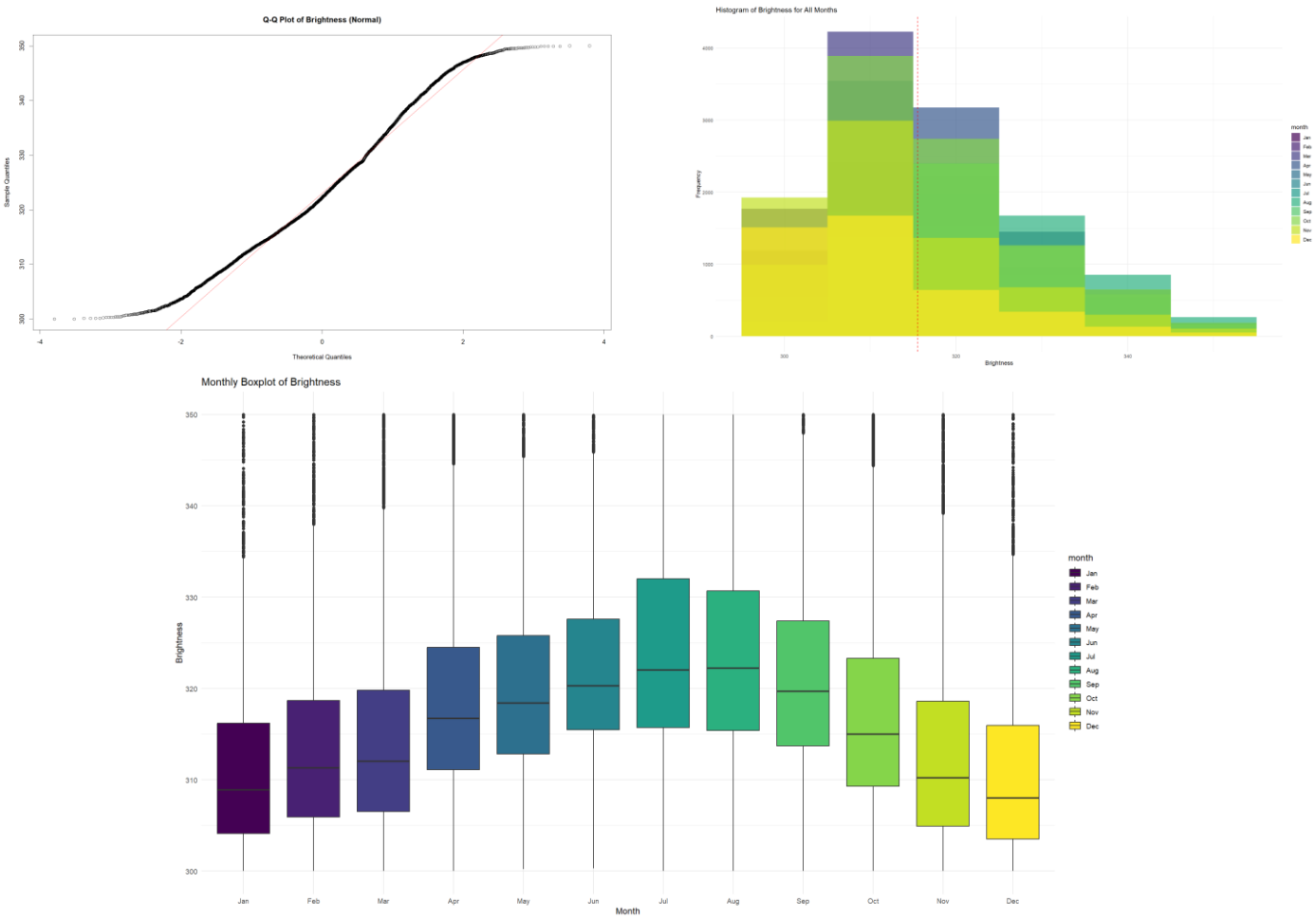

Figure 2. Q-Q plot with theoretical normal line in red (left), histogram with median dotted red line (right), and boxplot (below) by month for brightness of fires in Kelvin that are less than or equal to 350 K in continental USA.

There is still some skewness to the data, but this limitation greatly improves the skewness of the data, which now has mean (SD) of 317.42 K (11.23 K) and median (IQR) of 315.50 K (15.30 K). In a further effort to limit the number of observations in the data, only the month of August, which occurs during a higher point in the month, is considered:

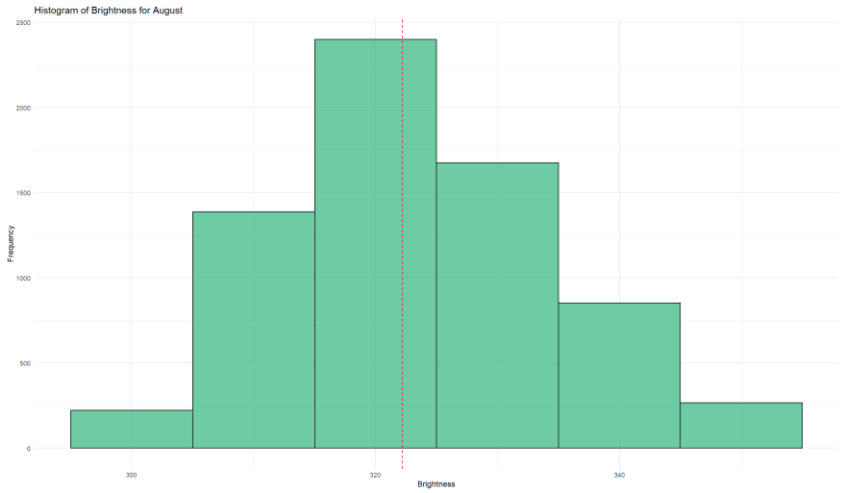

Figure 3. Histogram of brightness of fires (Kelvin) for the month of August with median in dotted red line. The month of August has mean (SD) brightness of 323.42 K (10.93 K) and median (IQR) of 322.20 K (15.27 K).

A further limitation of the dataset is to only conduct spatial-temporal analysis on the first 8 days of August, which allows for more readable interpolation and kriging dynamics as described in the sections below.

Part 3: Methods

To investigate the patterns of brightness over space (continental United States) and time (August 1st through 8th, 2023), interpolation using inverse distance weighting (IDW) and Gaussian kernel are used.

An initial spatial-temporal visualization is created to see the observational data over the 8 day period, where the true observations of the brightness are placed on a scatterplot of longitude and latitude for each of the 8 days.

For a spatial-temporal analysis to predict values at those locations which may not have actual values, two different interpolation techniques are used and compared. The first is IDW interpolation, which works on the assumption that values that are close together are more alike than those that are further apart. It estimates an unknown point by using a weighted average of all points, where points closer to the unknown point are assigned larger weights. The formula for IDW interpolation is as follows:

$$x_i = \frac{\sum_{j=1}^{n} \frac{x_j}{d_j^p}}{\sum_{j=1}^{n} \frac{1}{d_j^\alpha}}$$

where x_i is the unknown value, n is the number of points, x_j is the known value, d_j is the distance between the ith unknown value and the jth known value, and α is the power [Source 3]. The second interpolation technique is Gaussian kernel interpolation. It smooths the data by applying a Gaussian function to the weights and increases the influence of nearby points to the point at which the value is unknown. The Gaussian form of the kernel is defined below as:

$$K(x, x_i) = \exp\left(\frac{-\left|\left|x - x_i\right|\right|_2^2}{\theta}\right)$$

where $||\cdot||_2$ is Euclidean vector-norm and θ is a parameter to control the width of the kernel [Source 4].

The accuracy of both of the models constructed will be assessed by using leave-one-out cross-validation (LOOCV), which is an n-fold cross-validation. This is a method that uses each point in the data set as testing data (so there are n splits), with every other point as a training data. The prediction error is then calculated for the true value at that point and the predicted value, with the error aggregated as mean square error (MSE):

$$CV_n = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

Since LOOCV is based on MSE, it is logical that smaller LOOCV indicates a better predictive fit to the data [Source 5].

Finally, the parameters in each model, α for IDW interpolation and θ for Gaussian kernel, will need to be optimized. This is done by testing the LOOCV values for a prescribed sequence of parameter values, then finding the parameter value for each model that has the minimum LOOCV score.

Part 4: Results

As a first look into the data, the spatial-temporal relationship of the brightness is visualized over each day for the first eight days of August:

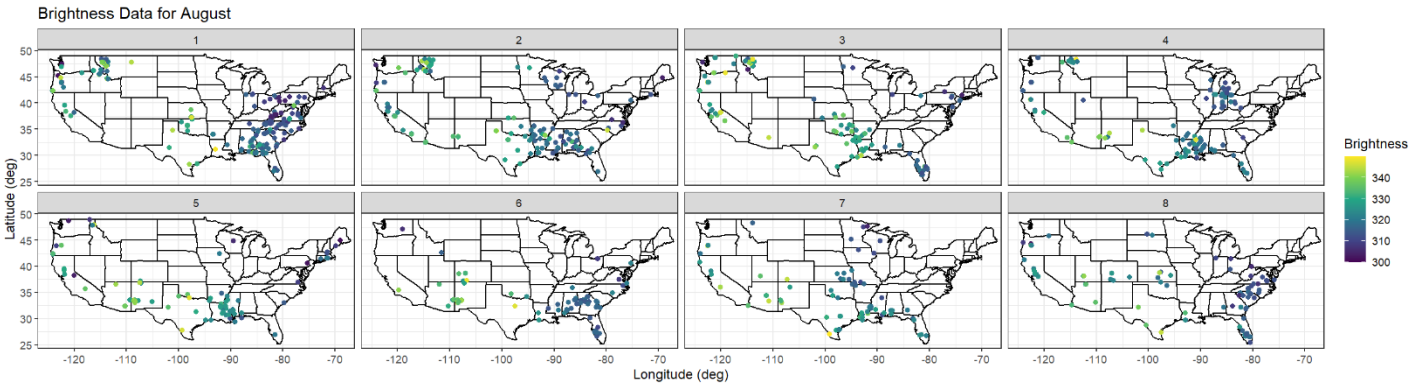

Figure 4. Spatial and temporal distribution of fire brightness (Kelvin) across longitude and latitude (degrees) for the first eight days of August with the map of the U.S. states overlaid.

Just from looking at the plots, it is clear to see that there are a lot of areas across the space that are do not have data/observations for specific day(s). However, from the data that is available, it seems as though there are darker colors (which correspond to cooler fire temperatures) the more east and north, while the western and southern observations are much lighter (which corresponds to hotter/more intense fire temperatures). (Note: again, to limit the outliers in the data, the brightness was capped at 350 K, so this is an insight into those temperatures that are at or below this range).

In order to predict the theoretical values at these other locations, a prediction grid was created using the coordinate and the day ranges. Then, an interpolation grid was created, here by using IDW interpolation:

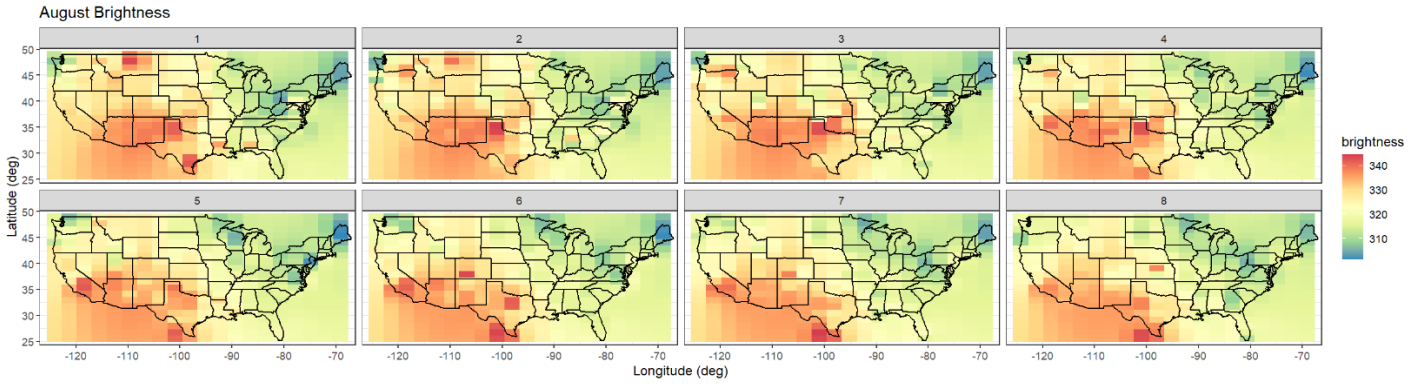

Figure 5. IDW interpolation of fire brightness (Kelvin) across longitude and latitude (degrees) for the first eight days of August.

The prediction grid created once again enforces the conclusion made visually that the fires that occur in the southwest are more intense than those in other areas of the country.

Next, the distance matrix of the interpolant versus the observation locations is made. The differences between the two sets of predictions are made using the IDW interpolation and the IDW interpolation with normalized weights. The summary statistics about these differences reveal that the maximum difference between any of the two interpolations is 2.615×10^{-12} and the median is 0, which indicates the two sets of predictions (normalized weights and IDW function) are virtually the same and the interpolant/prediction calculations are consistent between the two approaches.

The LOOCV scores are calculated for the initial interpolations for the IDW and Gaussian kernel methods, with prescribed parameter values of 5, which are 105.29 and 77.45, respectively. This indicates that the Gaussian kernel model performed better at this parameter value and was able to have smaller prediction errors on average than the IDW model. To consider the meaning of these values with respect to the problem, we consider the root mean square error (RMSE), which are in the same units as the brightness value (K), which are 10.26 K and 8.80 K, respectively. This means that, on average, the predictions for the IDW model deviate about 10.26 K from the actual brightness value (2.9% of the range) while the Gaussian kernel predictions deviate about 8.80 K (2.5% of the range).

In order to optimize the predictive methods, we find the optimal parameter values through trial by varying the theta values and finding the one that minimizes LOOCV. By sequencing 21 theta values starting from 0.1 until 4, we can visualize the parameter values versus their respective LOOCV:

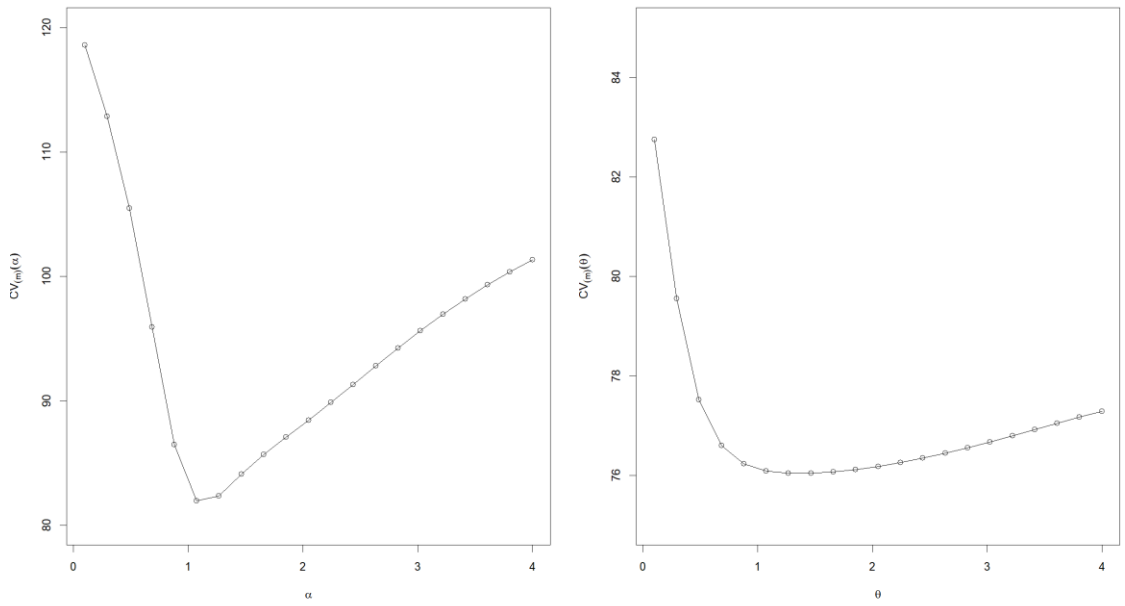

Figure 6. Parameter value for interpolation techniques, IDW (left) and Gaussian kernel (right), versus their respective LOOCV score.

Visually, it is clear at which point that the LOOCV (y-axis) is minimized, which is somewhere right around alpha/theta = 1 for both methods. This is confirmed by finding the minimum value explicitly:

Table 1. Comparison of LOOCV for IDW and Gaussian kernel interpolation methods at different parameter values. Range of brightness calculated by RMSE/350 K * 100%.

	IDW	Gaussian kernel
$\theta = 5$	105.29	77.45
RMSE	10.26	8.80
Range of Brightness	2.93%	2.51%
Optimal θ	81.95 (α = 1.075)	76.04 (θ = 1.465)
RMSE	9.05	8.72
Range of Brightness	2.59%	2.49%

By optimizing the parameter values for the different methods, the LOOCV score is also minimized. At the initial theta value, the Gaussian outperformed the IDW method by a lot. However, at their respective optimal parameter value, there is still a smaller LOOCV score for the Gaussian method, but the discrepancy between the two is much smaller. The Gaussian kernel having a higher parameter value also indicates that it optimizes the prediction accuracy by considering points that are slightly further than those considered by IDW interpolation.

Part 5: Discussion and Conclusions

This dataset of fires across the U.S., which includes man-made and natural fires (including volcanos), highlights the importance of accounting for space and time in statistical. Using the MODIS data, the spatial-temporal models were effective at analyzing and predicting brightness across the U.S. The two interpolation methods used, IDW and Gaussian kernel, indicated that the Gaussian kernel, which smooths weights for nearby points, slightly outperformed the IDW model at their respective optimal parameter values. The prediction errors for both were minimal, representing less than 3% of the overall range.

Some insightful conclusions from these interpolation models enforce the spatial dependency on fire intensity. Fires in the Southwestern states, which are desert climates, tend to be more intense compared to other regions. The Eastern states seem to experience a higher frequency of less-intense fires. Additionally, those states in the mountain ranges in the north seem to have no fire occurrences at this range.

These conclusions are insightful, but there are considerations and limitations to acknowledge. First, the analysis limited brightness values to under 350 K. While this limitation excluded outliers to create a more focused study, but including the full range of data would provide a more applicable and comprehensive model for nationwide conclusions. Similarly, the exclusion of Hawaii and Alaska limits the scope of the results to only the contiguous 48 states. Finally, the dataset only encompasses the first week of August, so the results may vary for other seasons/times of the year.

Part 6: References

1) "Active Fire Data Attributes for MODIS and VIIRS | NASA Earthdata." *NASA Earthdata*, 31 Oct. 2024, www.earthdata.nasa.gov/data/tools/firms/active-fire-data-attributes-modis-viirs.
2) "NASA-FIRMS." *Firms.modaps.eosdis.nasa.gov*, firms.modaps.eosdis.nasa.gov/country/.
3) Zhuang, Xiahai, and Yipeng Hu. "Statistical Deformation Model: Theory and Methods." *Elsevier EBooks*, 1 Jan. 2017, pp. 33–65, https://doi.org/10.1016/b978-0-12-810493-4.00003-1.
4) James, Gareth, et al. *An Introduction to Statistical Learning: With Applications in R*. Springer, 2013, https://www.statlearning.com/.
5) GISGeography. "Inverse Distance Weighting (IDW) Interpolation." *GIS Geography*, 22 May 2016, gisgeography.com/inverse-distance-weighting-idw-interpolation/.