Final Project Report

Mingzhi Ye

Introduction:

Objective

In this project, the objective is to analyze (1) how did the concentration of SO2 in CA change because of the Camp Fire in 2018 and (2) whether there is any specific relationship between the square of the distance of the county to the center of the fire and the increase in concentration of SO2 in 2018 because of the Camp Fire.

Background

The SO2 is one of the main pollutants produced by fires and Camp Fire in 2018 was the deadliest and most destructive wildfire in California's history. So it is probable that this big fire significantly increased the concentration of the SO2 in CA in 2018.

Introduction to the data

The data was from the EPA website and contains the data about the concentration of SO2 collected by every monitor in CA. And I select the data from 11/8 to 11/25 in each year from 2013 to 2018. This is because that the Camp Fire lasted from 11/8 to 11/25. I planned to use GET() function rather than downloading manually, in fact I successfully used GET() to obtain the data in my midterm project, but the connection through R to the American website was not available in the last week because I was in mainland China. There are difference between the data obtained by downloading manually and the data obtained by GET(). The former only has "Daily Max 1-hour SO2 Concentration" and the later only has "monitor_concentration". It is probable that "monitor_concentration" is not the maximum of the 1-hour SO2 concentration in a single day.

Methods:

Data source

The data was downloaded manually from the EPA website. I planned to use GET() function rather than downloading manually, in fact I successfully used

GET() to obtain the data in my midterm project, but the connection through R to the American website was not available in the last week because I was in mainland China. There are difference between the data obtained by downloading manually and the data obtained by GET(). The former only has "Daily Max 1-hour SO2 Concentration" and the later only has "monitor_concentration". It is probable that "monitor_concentration" is not the maximum of the 1-hour SO2 concentration in a single day.

Required tools

R packages including "data table", "dplyr", and "dplyer" are used to check, and manipulate the data.

Data Preparation

There are three continuous variables useful for the analysis, I use summary() to check and there are some abnormal minus value in the "concentration" variable, so I delete these rows.

I renamed some variables for easier reference.

I extracted the year as well as month & day information from the "Date" variable in the form of "xx/xx/xxxx" and create two extra variables "year" and "date", "date" only contains month and day information.

Results:

The website is on https://mingzhiye16.github.io/Final_Project/

Figure 1

The figures below display how did the concentration of SO2 in CA change from 2013 to 2018. The x axis represents the date from 11/8 to 11/25, which is when the Camp Fire lasted in 2018.

The first figure displays the difference between the concentration in 2018 and the average concentration in previous years, and the second figure displays concentration in each year from 2013 to 2018.

We can see that the concentration in 2018 is higher.
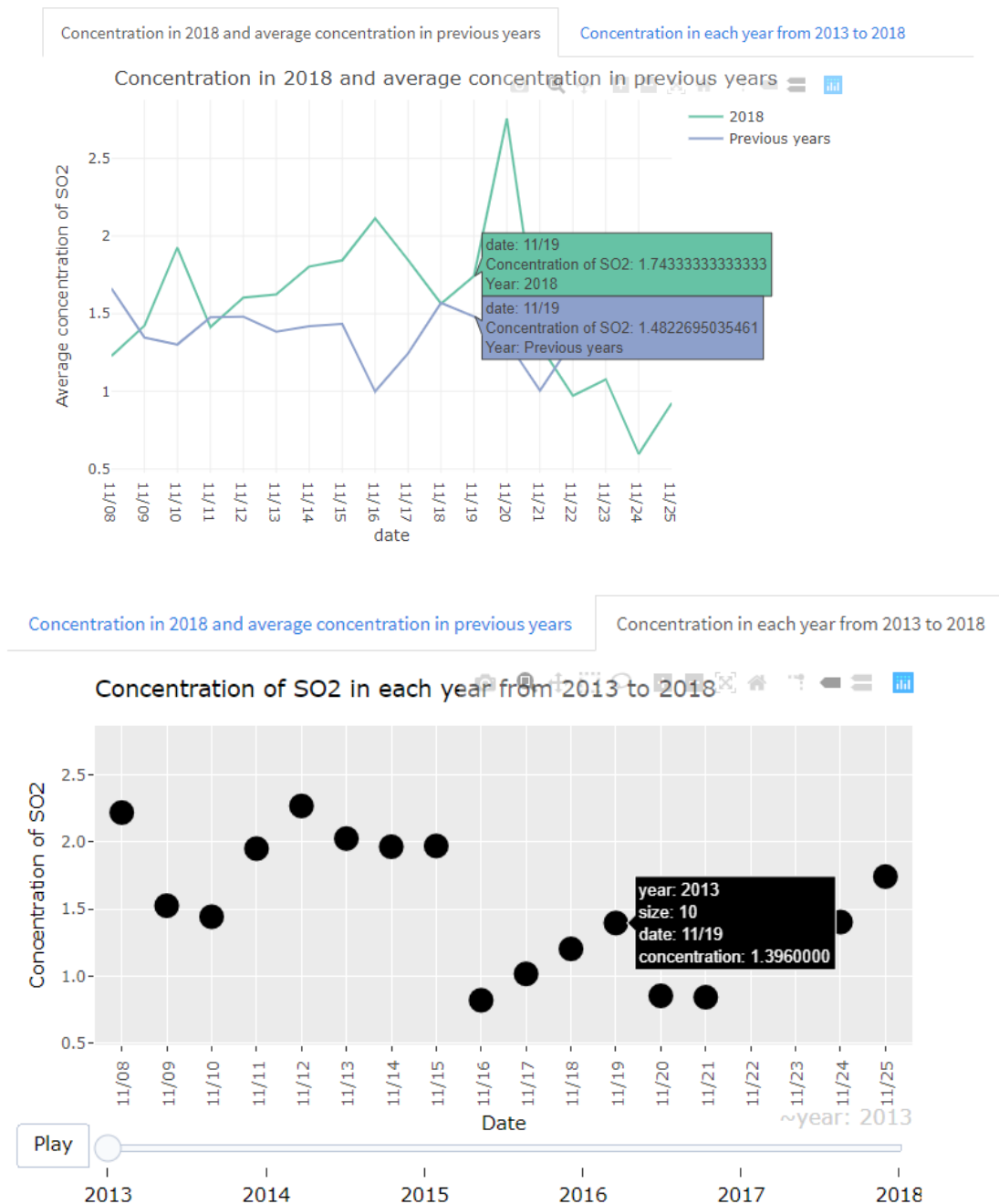
Figure 2

The heatmap below displays how did the concentration of SO2 in different counties change from 2013 to 2018, we can see that the change is not regular and in some county the concentration in 2018 even decreased.

I built a matrix of hover information text so that the interactive plot can show the information for each grid well.

There are three years that Inyo doesn't have any concentration data, which means the data of Inyo is not reliable, so I excluded it from later analysis.

When it comes to Orange county, there is no concentration data in 2018 when the wild fire happened, so I also excluded Orange from later analysis.
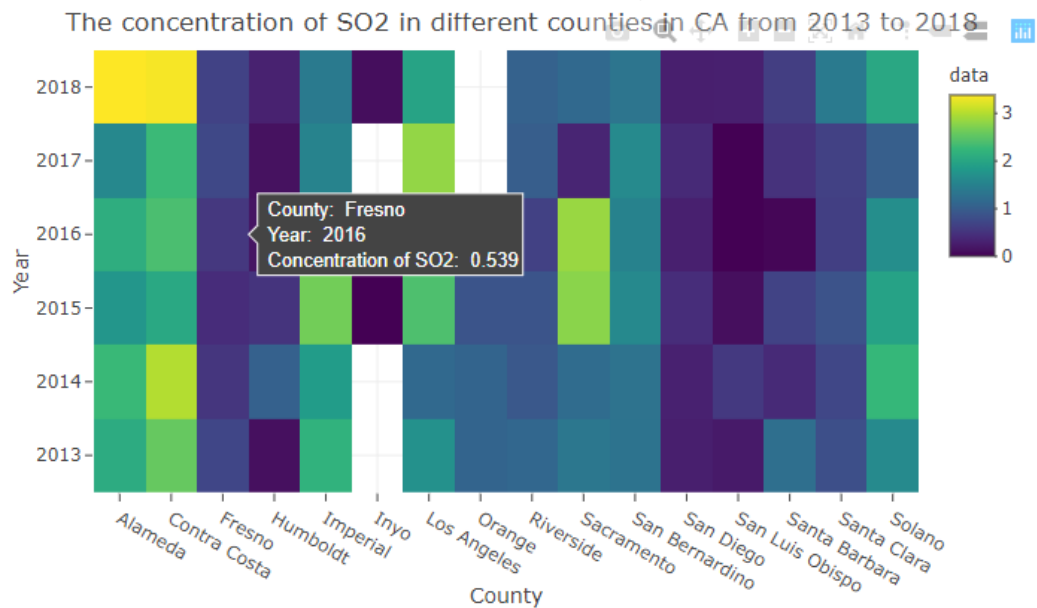
Figure 3

The map below displays the increase in concentration of SO2 in 2018 compared to previous years. Each circle represents a county. The bigger the size of the circle, the greater the increase.


Figure 4

The figure below displays Relationship between the square of the distance to the center of the fire and the increase in the concentration of SO2 in 2018 compared to previous years. The bigger the size of the circle, the greater the increase.

We can see that in general the closer the county is to the center of the fire, the greater the increase in concentration of SO2.

I also tested whether there is a linear relationship between the square of the distance and the increase in concentration, and the result is that there is no significant linear relationship (p-value = 0.06512).
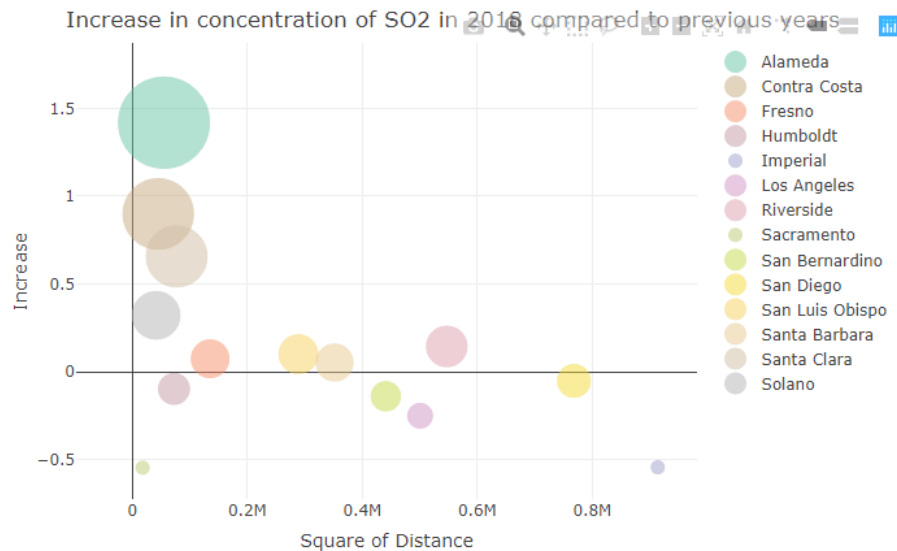

```
##
## Call:
## lm(formula = increase ~ square_of_distance, data = data_sx)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95784 -0.19934 -0.06367  0.23168  1.04173
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.275e-01  1.901e-01   2.249   0.0441 *
## square_of_distance -9.294e-07  4.578e-07  -2.030   0.0651 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4847 on 12 degrees of freedom
## Multiple R-squared:  0.2556, Adjusted R-squared:  0.1936
## F-statistic: 4.121 on 1 and 12 DF,  p-value: 0.06512
```

Figure 5

I use the concentration data in previous years (2013-2017) in each county to build linear models, assuming that there is a linear relationship between the year and the concentration. Then I use the models to predict the estimated each county's concentration of SO2 in 2018 when there was no Camp Fire. Then I calculate how much the actual concentration in 2018 is higher than the predicted concentration in 2018 in each county. In this way I can reduce the influence of confounders and focus on the influence of the Camp Fire on the concentration of SO2.

The figure below displays the relationship between the increase from the

predicted concentration to the actual concentration and the square of the distance from the county to the center of the fire. The bigger the size of the circle, the greater the increase.

We can see that in general the closer the county is to the center of the fire, the greater the increase in concentration of SO2.

I also tested whether there is a linear relationship between the square of the distance and the increase in concentration, and the result is that there is no significant linear relationship (p-value = 0.08821).

The increase of the actual concentration compared to the predicted concentration
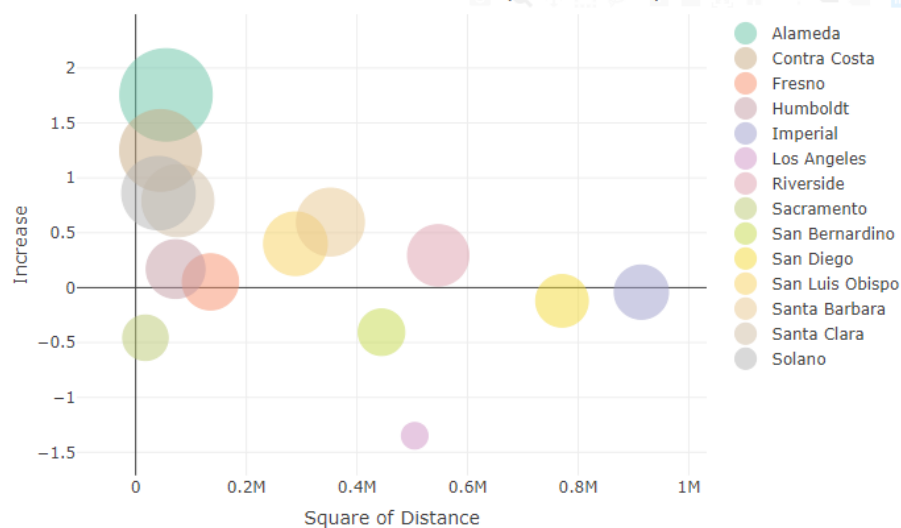
```
##
## Call:
## lm(formula = increase ~ square_of_distance, data = data_compare)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3713 -0.4200  0.2123  0.3701  1.1738
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         6.503e-01  2.792e-01   2.329   0.0381 *
## square_of_distance -1.245e-06  6.708e-07  -1.856   0.0882 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7118 on 12 degrees of freedom
## Multiple R-squared:  0.223,  Adjusted R-squared:  0.1582
## F-statistic: 3.444 on 1 and 12 DF,  p-value: 0.08821
```

Conclusion:

1. The concentration of SO2 in CA in 2018 is significantly higher than that in previous years
2. The amount of increase in concentration of SO2 is different in different counties.

3. In general, the closer the county is to the center of the fire, the greater the increase in concentration of SO2.
4. There is no significant **linear** relationship between the square of the distance from the county to the center of the fire and the increase in concentration of SO2 (p-value = 0.08821).
5. Probably the linear model can't effectively predict the estimated each county's concentration of SO2 in 2018 when there was no Camp Fire. Because probably the relationship between the year and the concentration of SO2 is not linear in previous years.