STDA Final Project

SHAFFER MICHAL, 2020321753

2021/11/15

Comparing models to analyze and predict air quality data

Contents

1 Introduction

In this project I explored some air quality data in California, USA over the years 2010 through 2020. At the begining of 2020, many countries and cities had mandatory lockdowns due to the COVID-19 pandemic. With less people driving to work and factories being shut down, I wondered if the lockdowns had a significant effect on the air quality in major cities. With this question in mind, I explored and compared various methods of analyzing and predicting air quality in California over the years 2010 through 2021. With each method I tried to analyze or predict various aspects of the data.

2 Methods

2.1 Data

I used air quality data from the United States Environmental Protection Agency. The agency provides various data sets including meteorological data, daily, and yearly air quality concentration data sets, as well as daily AQI (air quality index) datasets. I chose to focus on two measures: PM2.5 and Ozone levels. I chose to use the datasets that provide the AQI values for these measures instead of the raw data, as the AQI gives an immediate impression of the severity of the pollution levels.

The AQI is divided into six categories:

AQI (Air Quality Index) Chart		
AQI values	Levels of Health Concern	Colors
0-50	Good	DarkGreen
51-100	Moderate	Light Green
101-150	Unhealthy for Sensitive Groups	Yellow
151 to 200	Unhealthy	Orange
201 to 300	Very Unhealthy	Red
301 to 500	Hazardous	Purple

2.2 Methods

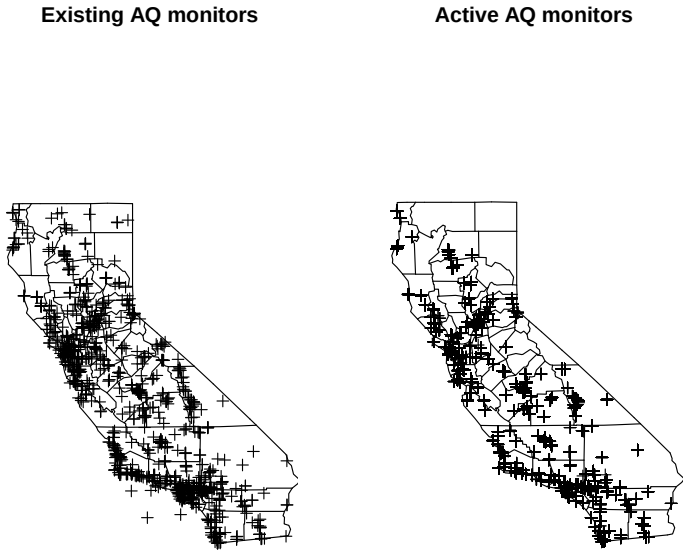
I used the following methods for my various analyses: Likelihood-based inference, kriging, point-wise regression models, and generalized linear spatio-temporal regression.

For the inital exploratory data anlaysis, I looked at the average AQI levels over the entire time period of my study, and used both likelihood based inference and kriging methods to interpolate the AQI levels for existing, but inactive air-quality monitors. I then used a point-wise linear regression model to obtain linear models for each air-quality monitor using data from 2010-2019. I then used those models to predict the AQI values for 2020 (assuming it was a normal year - without COVID lockdowns) and compared those values to the actual 2020 AQI values to determine whether the difference in values was significant. Finally, I used generalized linear spatio-temporal regression to analyze the data a different way and observe the residuals for evidence of spatial and/or temporal correlation.

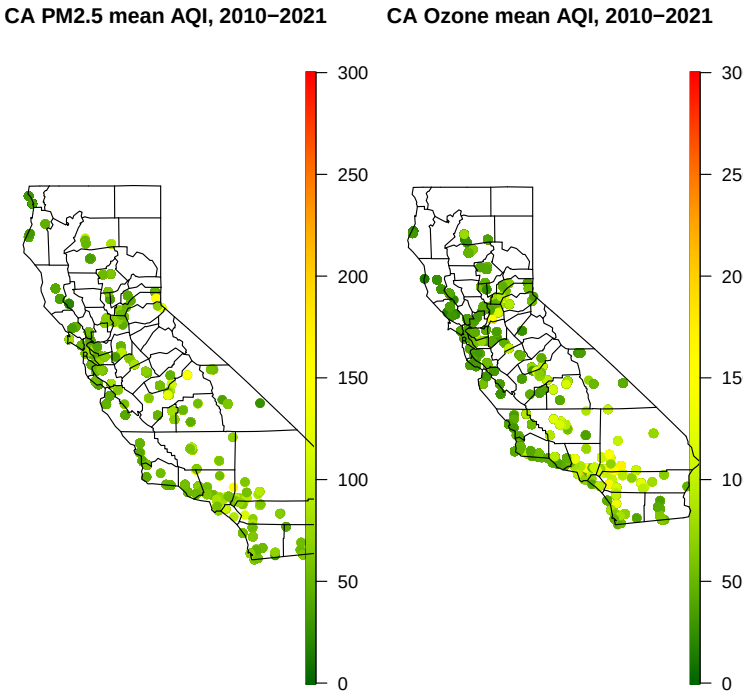
3 Initial Data Exploration

I first looked at the available data:

Figure 1 shows all of the existing air-quality monitors in California while the second figure shows the monitors that are active.



Figures 3 and 4 show the average AQI values for PM2.5 and Ozone, respectively, for each active monitor (active = has data) over the entire time period of the study (2010-2021).



It is apparent from these plots that there are more monitors reporting Ozone levels than those reporting PM2.5 levels. We also see some clustering of slightly higher AQI levels i n the southern areas near Los Angeles county, a big metro area, and lower levels along the northern coast-lines.

4 Likelihood based estimation and prediction using kriging

To begin my analysis, I wanted to see the AQI levels for all the inactive monitors to get a better idea of the AQI levels across the entire state. To conduct this interpolation, I began with a likelihood based estimation. I first ran a function for each year individually as I intended to

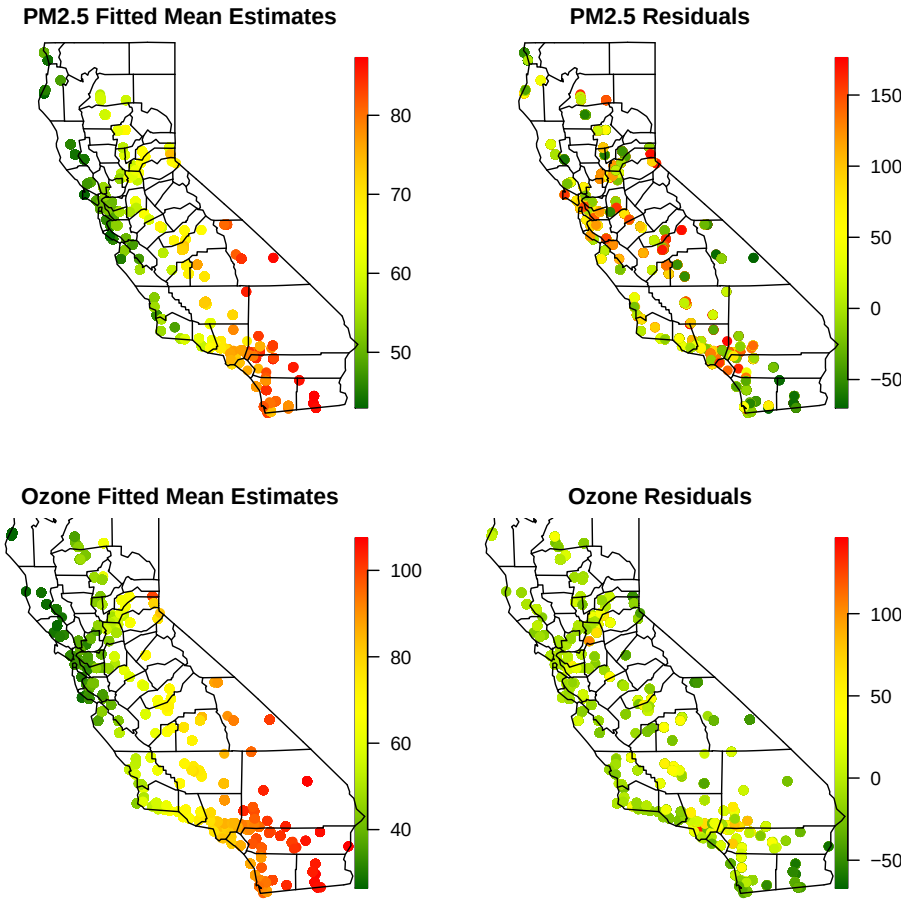
use this data for the point-based prediction later on, but ultimately that data was not used in my analysis. Therefore, I did not include the yearly plots of the MLE fit, residuals, and the kriging results for each year. I also conducted the estimation and prediction for the average values across all the years, which I present here:

The linear model used parameters that showed significance of their effect on AQI. The parameters include the monitor geolocation, elevation, and the type of location (0 - Urban, 1 - Suburban, 2 - Rural)

$AQI\ Longitude + Latitude + Elevation + Location.SettingNum$

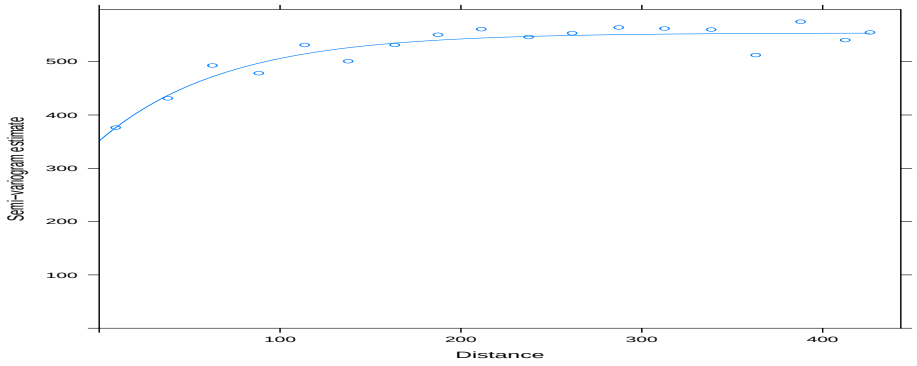
Fitted values and residuals were found using the following formulas: $\hat{Y} = X\hat{\beta}\ \hat{e} = Y - \hat{Y}$

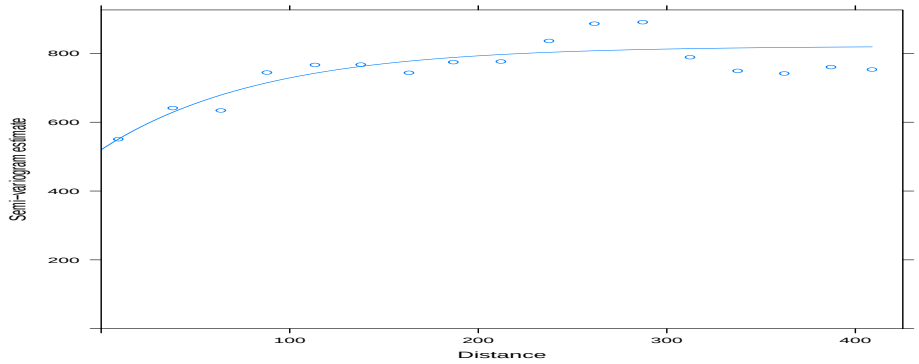
The fitted and residual plots:



From a subjective viewing of the fitted and residuals plots you can see some areas of clustering, indicating spatial association. The variagrams also showed indication of spatial association. The exponential variogram formula used:

$$\gamma(h) = C(0) - C(h) = \tau^2 + \sigma^2(1 - exp\{\frac{-||h||}{\rho}\})$$





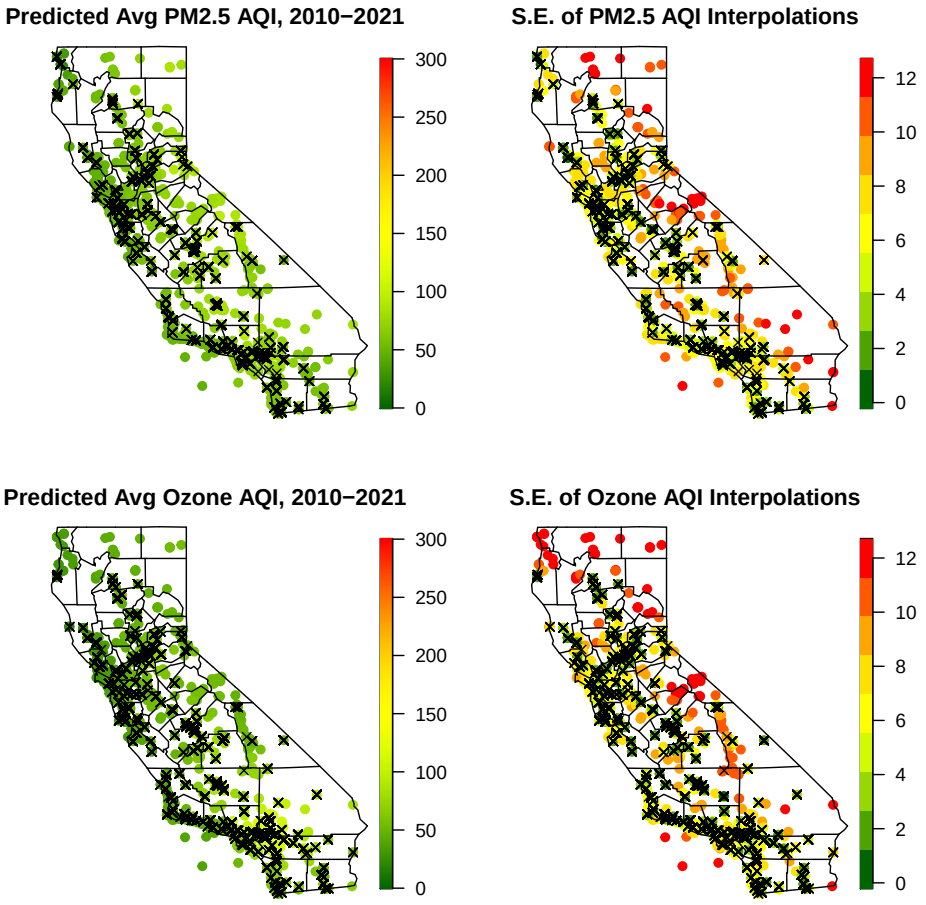
5 Kriging

For interpolation of the AQI data for inactive monitors, I used the $\hat{\beta}_{gls}$ obtained earlier and the following kriging formulas:

$$\hat{Y}(s_0) = m_0 + \gamma' \Sigma^{-1}(Y - m)$$

$$Var[\hat{Y}(s_0) - Y(s_0)] = \sigma^2 - \gamma' \Sigma^{-1} \gamma$$

You can then see the predicted interpolations as follows:



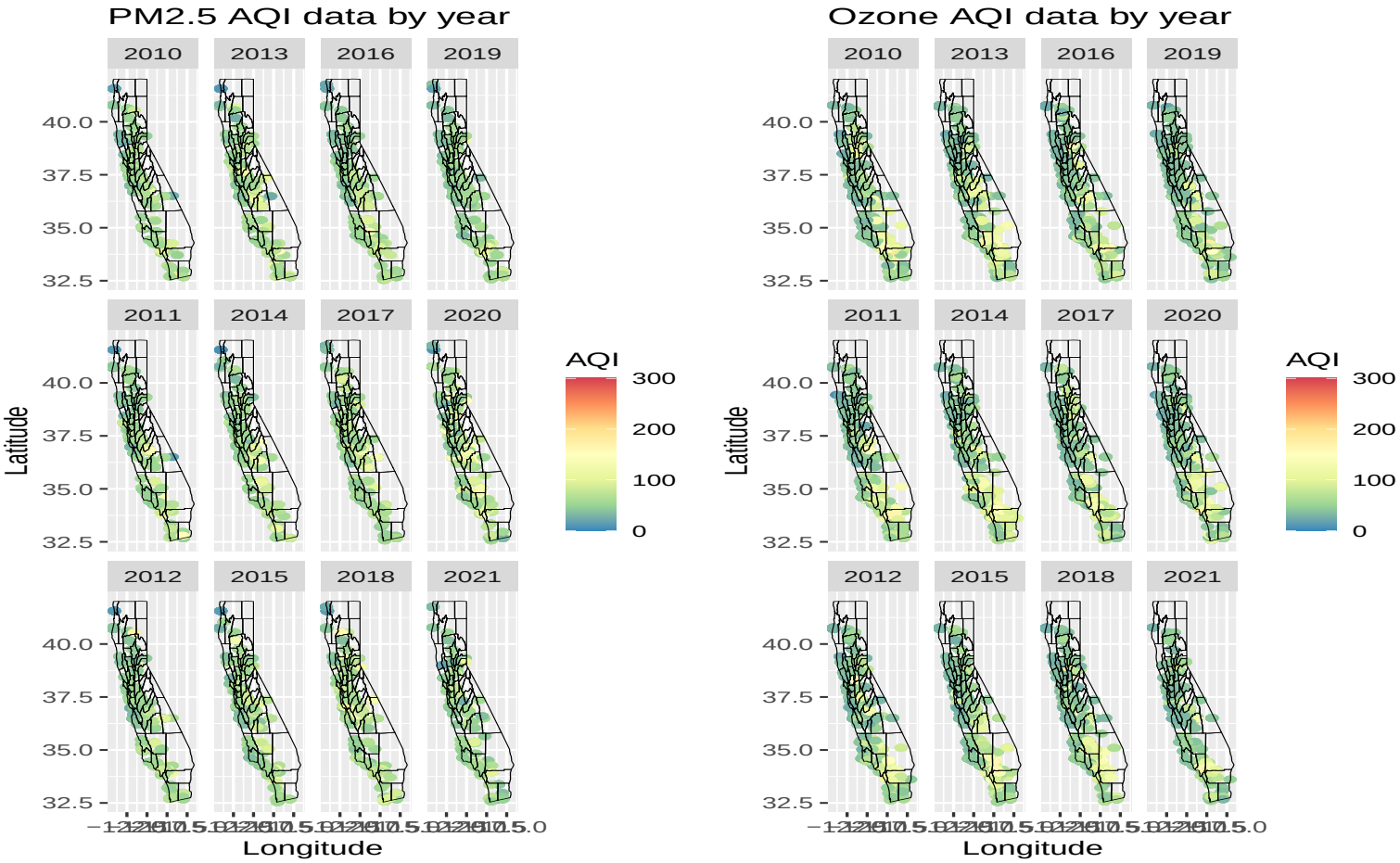
You can see from the plot that the standard error is higher when there are fewer active monitors surrounding the location of interest. The fewer monitors within close distance, the less confident the predictions will be.

6 Point-wise linear regression analysis and prediction

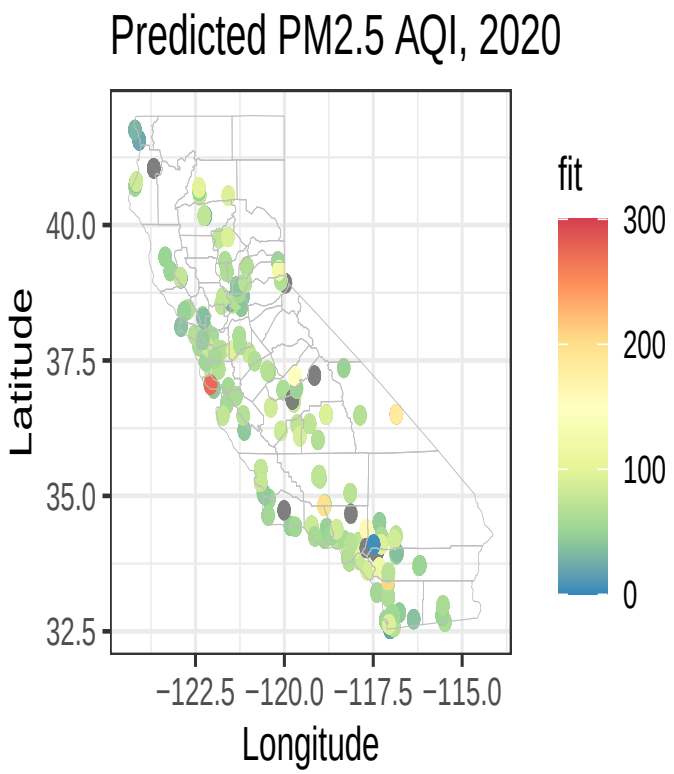
The first method I used to predict AQI values for the year 2020 was point-wise regression analysis and prediction. I first filtered the data to only consider the years 2010-2019 and grouped

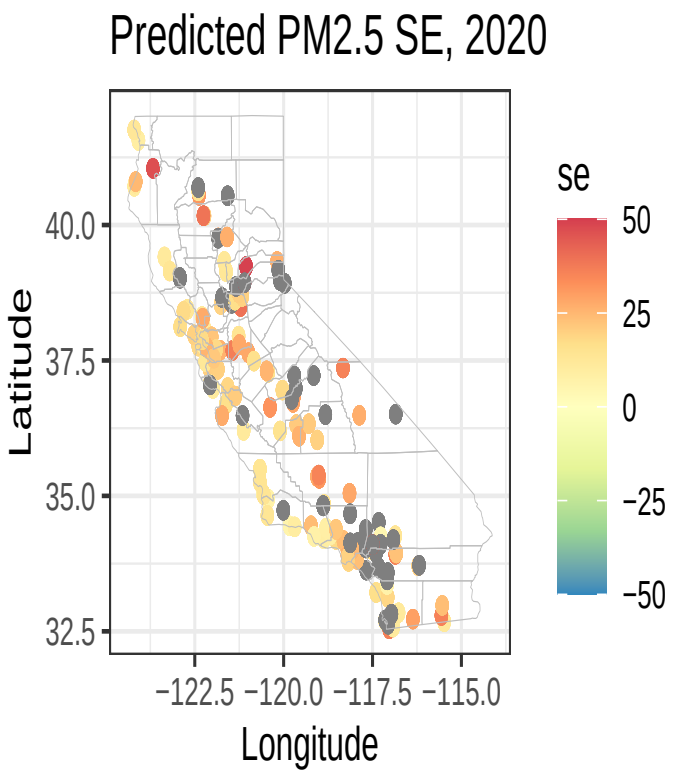
the data by monitor location. I then found a linear model for each monitor individually using the following formula:
$AQI = 1 + year + month$
This formula used the temporal information and AQI data for each monitor location to create an appropriate regression model. I then used these models to predict AQI values for each month of the year 2020 at each monitor location.

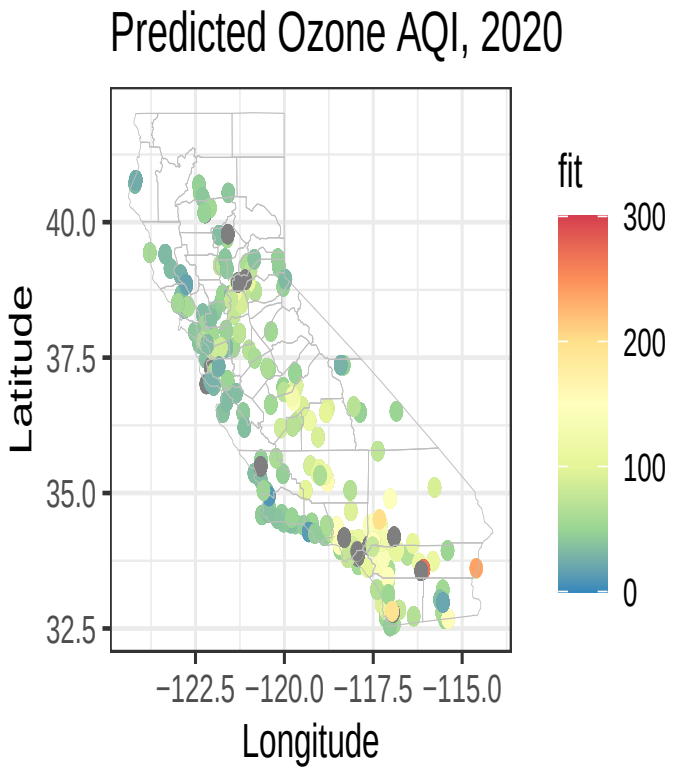
First we can look at the AQI values for each year:

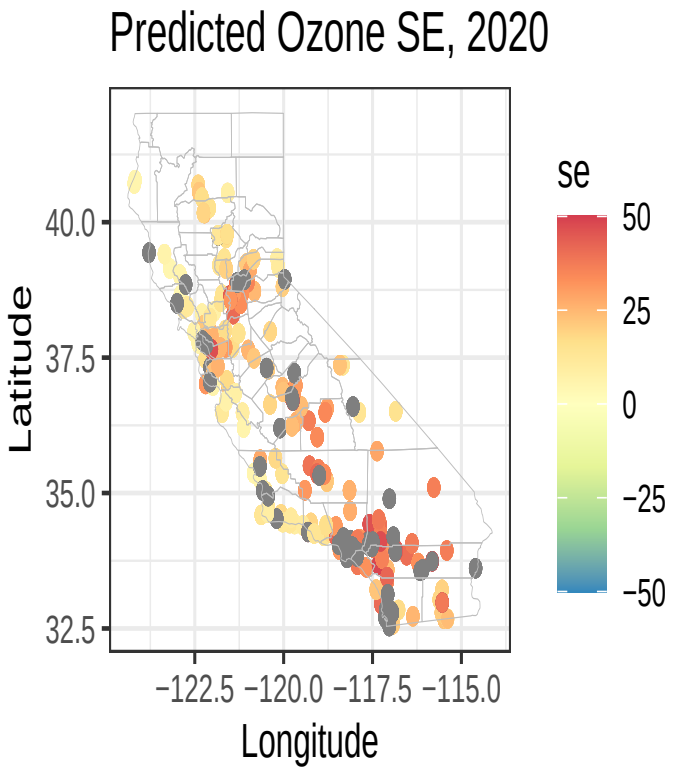


After using the point-wise regression models to come up with 2020 predictions: The AQI predictions for 2020:

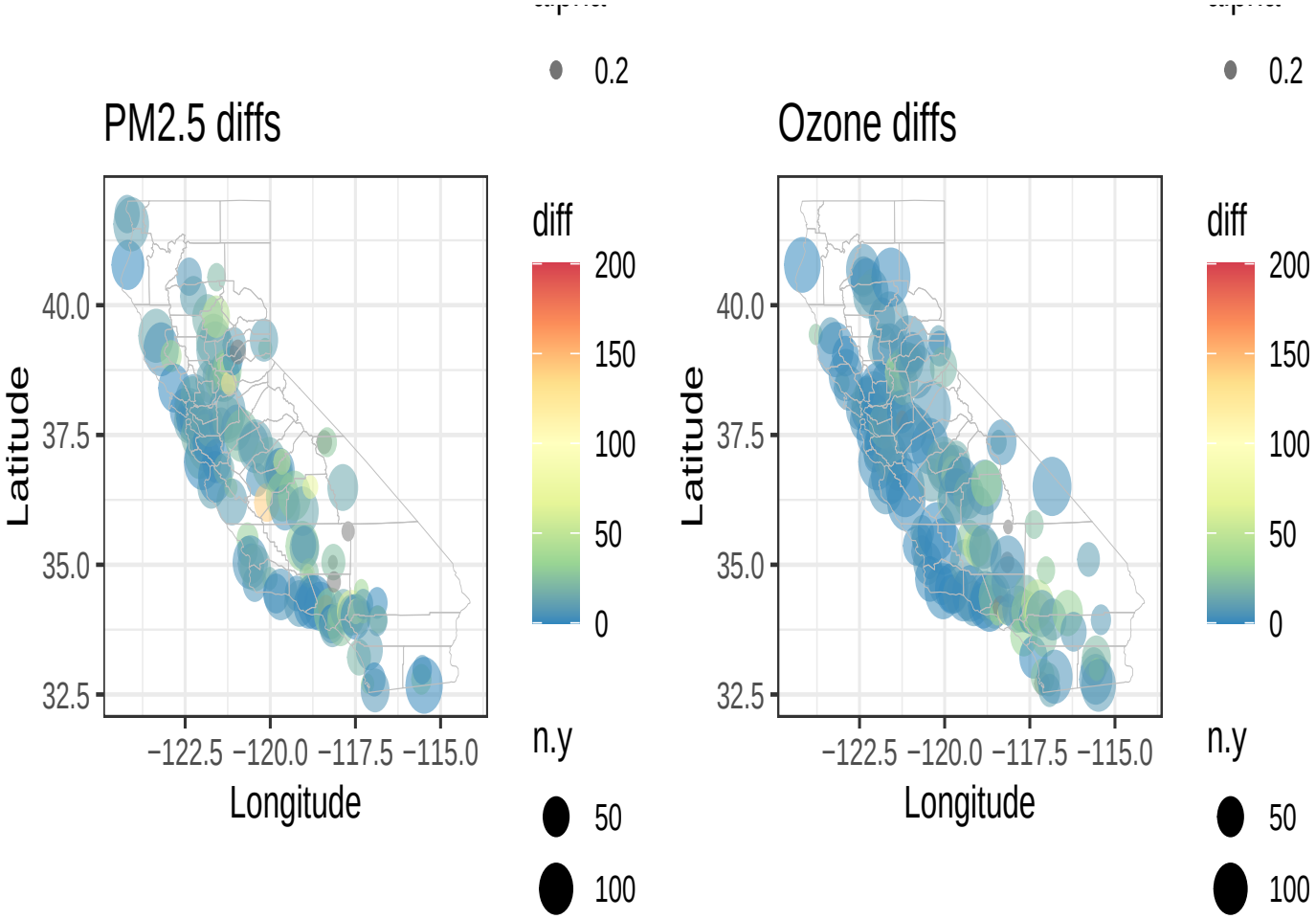

Predicted PM2.5 AQI, 2020
Latitude
40.0
37.5
35.0
32.5
−122.5 −120.0 −117.5 −115.0
Longitude
fit
300
200
100
0


Predicted PM2.5 SE, 2020
Latitude
40.0
37.5
35.0
32.5
−122.5 −120.0 −117.5 −115.0
Longitude
se
50
25
0
−25
−50


Predicted Ozone AQI, 2020
Latitude
40.0
37.5
35.0
32.5
−122.5 −120.0 −117.5 −115.0
Longitude
fit
300
200
100
0


Predicted Ozone SE, 2020
Latitude
40.0
37.5
35.0
32.5
−122.5 −120.0 −117.5 −115.0
Longitude
se
50
25
0
−25
−50

Finally, we can compare the predicted fitted values with the actual values recorded in 2020 for each monitor. Below are plots showing the difference in the predicted and actual values. Each point is also sized according to the number of observations there are for that monitor.

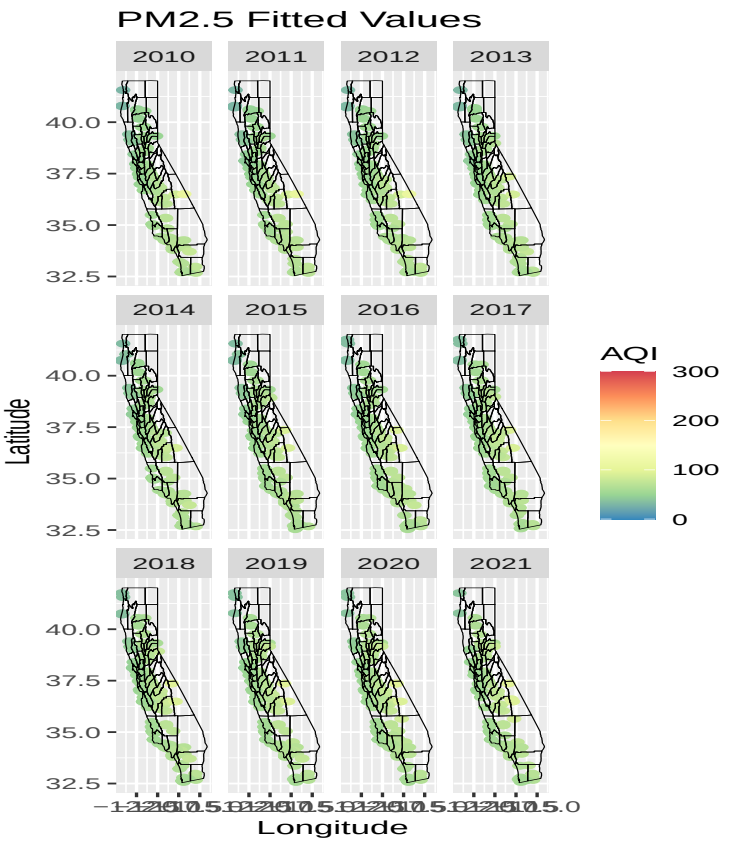


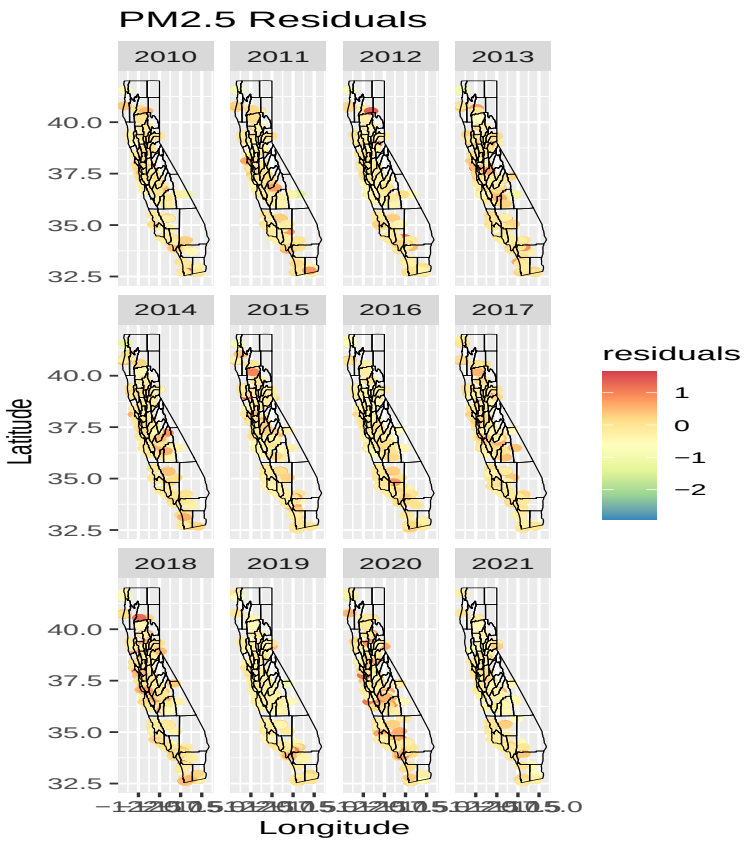
Looking at the plots, you can observe the big differences occur where there are fewer observations, so the model doesn't fit as well. However, most of the differences are near zero where there were a significant amount of observations.

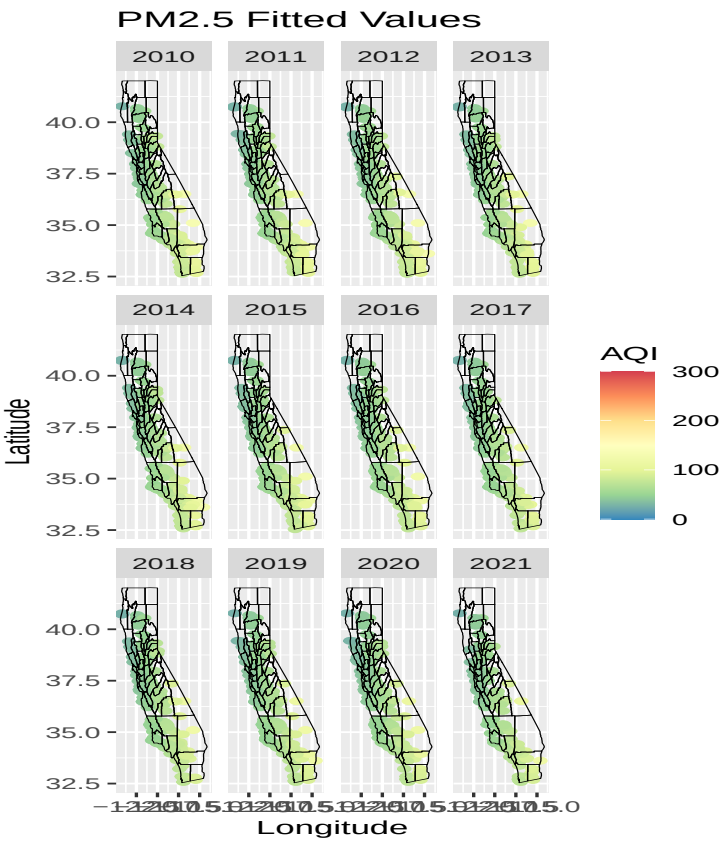
7 Spatio-Temporal GLM

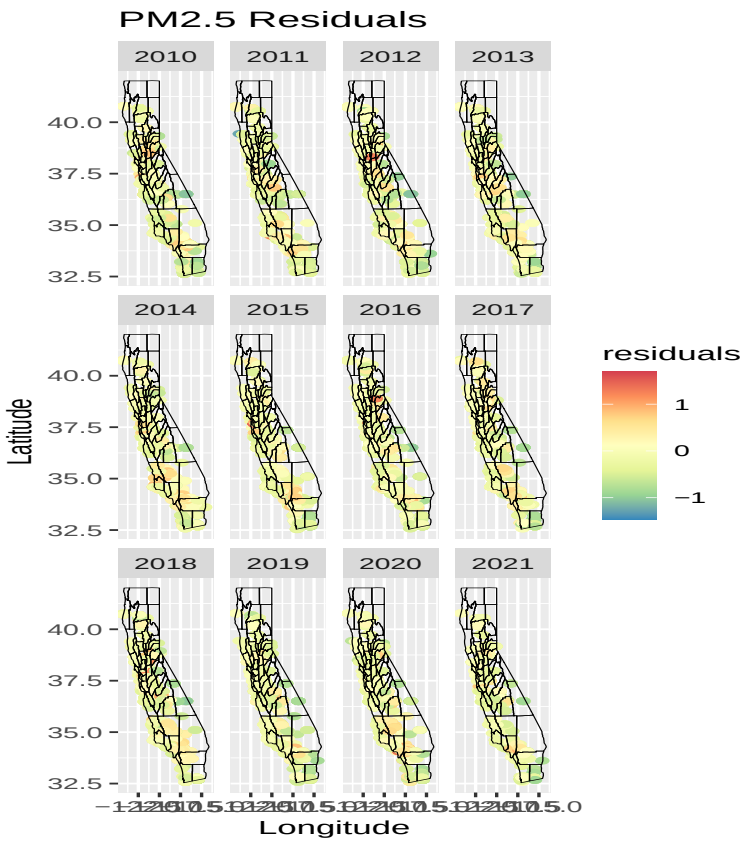
The GLM equation I used takes into account the spatial and temporal data as well as elevation at each monitor:
$AQI = (Longitude + Latitude + year)^2 + Elevation$
After confirming the p-value showing significance for no overdispersion, I then created a basis function to evaluate at each spatio-temporal point where I wanted a prediction.

The predictions for each year are shown in the following plots:









Looking at the plots and sumaries for both PM2.5 and Ozone predictions, you can see the clustering as evidene of spatio-temporal associations.

I also ran Moran's I test for both sets of predictions to see the significance of the spatio correlations.

```
## [1] "PM2.5 Predictions Moran's I summary"
##     observed        expected             sd           p.value
##  Min.   :0.045   Min.   :-0.0018   Min.   :0.023   Min.   :0.0e+00
##  1st Qu.:0.156   1st Qu.:-0.0017   1st Qu.:0.024   1st Qu.:0.0e+00
##  Median :0.234   Median :-0.0016   Median :0.025   Median :0.0e+00
##  Mean   :0.227   Mean   :-0.0016   Mean   :0.025   Mean   :3.8e-03
##  3rd Qu.:0.314   3rd Qu.:-0.0016   3rd Qu.:0.025   3rd Qu.:1.6e-05
##  Max.   :0.359   Max.   :-0.0013   Max.   :0.027   Max.   :4.2e-02
## [1] "Ozone Predictions Moran's I summary"
##     observed       expected             sd           p.value
##  Min.   :0.27   Min.   :-0.0012   Min.   :0.018   Min.   :0
##  1st Qu.:0.32   1st Qu.:-0.0011   1st Qu.:0.019   1st Qu.:0
##  Median :0.35   Median :-0.0011   Median :0.020   Median :0
##  Mean   :0.35   Mean   :-0.0011   Mean   :0.020   Mean   :0
##  3rd Qu.:0.37   3rd Qu.:-0.0010   3rd Qu.:0.020   3rd Qu.:0
##  Max.   :0.40   Max.   :-0.0010   Max.   :0.022   Max.   :0
```

As evident in the summaries, when looking at a 5% level of significance, the null hypothesis (of no spatial correlation in these deviance residuals) is rejected, proving spatial correlation.

8 Conclusion

Each method of analysis and prediction provided different aspects that can be necessary for different study varieties and focuses.

The likelihood based estimation and kriging works well for interpolating spatially correlated data, but the formulas I used fall short when the data includes a temporal dimension. Although kriging is possible for spatio-temporal data, it is quite compuationally expensive. Therefore, I ignored the temporal dimension in my analysis and used different models to explore that part of the data.

The point-wise linear regression was a good model to use for spatio-temporal data, allowing for each geo point to have its own regression model. It works well for spatial locations with many observations and provides good predictions with low SE, but it falls short in locations with fewer observations. This model bases the predictions mostly on temporal trends and doesn't well incorporate the spatial association in the models.

The GLM model is different than the point-wise model, focusing on spatial associations and less on temporal associations. Choosing the proper response model is important. For example, in this study, when assuming a Poisson response and a log link there was the problem of over-dispersion. Therefore, I changed the response model to Negative-Binomial to solve the over-dispersion issue. Looking at the plots of the predictions and errors of the GLM model it is usually fairly easy to determine the strength of the spatial and temporal associations.

9 Limitations and Further Study

- The initial idea for this project was to explore this data for New York City, but I found that the NYC data was very limited, with only about 13 active air-quality monitors reporting data in the NYC area. I therefore moved my focus to the entire state of California, where there are a lot of metro areas as well as many parks and non-residential areas, hoping to see a bigger variety in air quality data across the state.

- For further studies, analyzing other air pollution measures (CO2, PM10, etc) would be interesting. In addition, expanding the analysis to the entire country, or looking at this data for other big cities or counties would be advisable.

- Although meteorological data was available for the monitor locations, it was very sparse and unreliable. In future study, finding the temperature, wind speed, and barometric pressure for each location at each time point would add to the accuracy of the models, as some meteorological elements may have an effect on air-quality measurements.

10 References

The United States Environmental Protection Agency `https://aqs.epa.gov/aqsweb/airdata/download_files.html`

NYCOpenData `https://opendata.cityofnewyork.us/`

Rafael Borge, Weeberb J. Requia, Carlos Yague, Iny Jhun, Petros Koutrakis, Impact of weather changes on air quality and related mortality in Spain over a 25 year period [1993-2017], Environment International, Volume 133, Part B, 2019, urlhttps://doi.org/10.1016/j.envint.2019.105272.

Wikle, C. K., Zammith-Mangion, A., & Cressie, N. (2019). Spatio-temporal statistics with R. CRC Press.