# Report of MA678 Midterm Project

Lihao Liao

Dec 10, 2021

## Abstract

In this report, I have explored the distribution of ground-level Ozone concentrations at 89 different locations in the United States. Based on the exploratory data analysis and linear mixed-effect regression, I have discovered that the ground-level Ozone concentrations are strongly correlated with locations where they are measured and the time when they are measured. In addition, the NO3 concentrations also affect the ground-level Ozone concentrations. This has implications for the prediction of future ground-level Ozone concentrations.

## Introduction

Ozone, as a natural substance, plays an important role in environmental sustainability (Staehelin et al. 2001). It can be used as a water disinfectant and natural pesticide in agriculture. However, ground-level ozone can have a harmful effect on humans since it is mostly formed of "fog." Ozone pollution has been found to be a major health hazard worldwide(Stathopoulou et al. 2008). Daily temperatures, relative humidity, and wind speed can affect ozone levels. In general, warm dry weather is more conducive to ozone formation than cool wet weather. Wind can affect both the location and concentration of ozone pollution(EPAR 2020). United Environmental Protection Agency (EPA) considers anything over 70 ppb to be unhealthy for human health and welfare. Therefore, for this project, I decided to study the distribution of ground-level ozone in the last ten years (2011-2020) in the United States. The aim of this project is to build a multilevel model with ground-level ozone concentrations being the response and time, location, other air pollutants being predictors to predict the future distribution of ground-level ozone concentrations in the United States. In the rest of this article, Ozone concentrations indicate ground-level Ozone concentrations unless told otherwise.

## Method

### Data Cleaning and Processing

The data used for analysis are ground-level ozone concentrations and concentrations of pollutants at more than 85 rural sites throughout the United States collected by United Environmental Protection Agency (EPA) from 2010 to 2021. All of the data were downloaded at https://java.epa.gov/castnet/clearsession.do. I have used two datasets for this study including the 8-hour daily maximum Ozone concentration data and the measured concentrations for each pollutant (such as SO2 and NO3) averaged over weekly. Since the concentrations of ground-level Ozone are mostly correlated to the man-made emissions of nitrogen oxides(Zhang, Wei, and Fang 2019; Stathopoulou et al. 2008), I have decided to only keep variables of locations of monitor site, date, and concentration of NO3 (Nitrogen Oxidation 3) for the data of concentrations for pollutants. These two data were merged by monitor sites and the time when they were collected, which were represented by year and week number in the year. Since there was no information on the week in which the concentrations of Ozone was measured, I have first converted the original date variable ("DDATE") into a format of date in R language and then convert it into the number of weeks. In addition, ground-level ozone concentrations are heavily affected by temperature and humidity(Zhang, Wei, and Fang 2019; Stathopoulou et al. 2008), which is why ground-level ozone concentrations are also found to be seasonal. Therefore, I have created another variable, named Season, to indicate the season when the ozone concentrations and NO3 were measured based

on the date of measurement. Further, since we would like to study the effect of time (year) on the ozone concentrations and predict the pattern of ozone concentrations in the future, we decided to create another variable, named "yeard10," which indicates the difference of year observed from the year 2010. For instance, $yeard10 = 1$ for year 2011. Thus, the final study data contain variables of monitor location, year, year increased from 2010, season, week, day, ozone concentrations, and NO3 concentrations.

There are a total of 336704 observations in the original dataset of ozone concentrations and a total of 54847 observations in the data of pollutants concentrations. After merging the two datasets, there are a total of 323528 observations in the joint data. There are a few missing data (n=6705). Since there were no particular patterns for the missing data and the amount of missing data is relative smaller given the sample size, I decided to remove them from the study data. Further, since the data for 2021 were not complete yet, I decided to remove the data in 2021 and use the rest of the observations, which are measured in the year 2010-2020, to train (build) the model. There are a total of 296274 observations in the training set that were collected from 89 monitor sites for 3998 different days from 2010-2020.

### Exploratory Data Analysis

The average Ozone concentration across all the monitor sites from the year 2010 to 2020 is 41.265. The spread of the Ozone concentrations is large (variance = 135.93). The overall average and median value of Ozone concentration across all the monitor sites have decreased since 2010 (See Figure 1a). Similarly, the overall average and median value of NO3 concentration across all the monitor sites have also decreased since 2010 (See Figure 1b). Since there were a total of 89 monitor sites included in the data, it will make the plot too complicated if we plot them all. Thus, I have decided to randomly select 10 monitor sites and plot the Ozone distribution only at the selected sites for the purpose of simplicity. As we can see from Figure 3, the distribution of Ozone concentrations is quite different among the 10 selected sites. For each site, the ozone concentration tends to increase and then decrease over time (week), which can be partially explained by the fact the ground-level ozone concentrations typically are high on hot days, which are the middle of the year for most locations. Further, the distribution of Ozone concentrations over Week is not a straight line, which suggests a non-linear association or/and possible interaction with confounders. From Figure 3, we can see that the distributions of Ozone concentrations at the selected monitor sites are different in different seasons. Ozone concentrations tend to be higher in Summer and Fall than they are in Spring and Winter, which is reasonable since Ozone concentrations are usually higher when the temperature is high. Thus, we suspect that there might be an interaction effect between Season and Week on the Ozone concentrations. The scatter plot of Ozone and NO3 concentrations across 10 years (2010-2020) for the selected locations (See Figure 4) suggests that there might be a positive linear association between them.

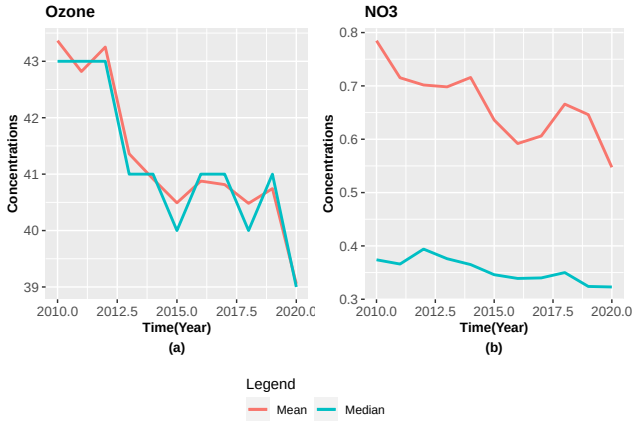



Figure 1: Average OZONE concentration and NO3 concentrations across monitor locations from Year 2010 to Year 2020


Figure 2: Average Ozone concentrations across year over time at selected minitor sites

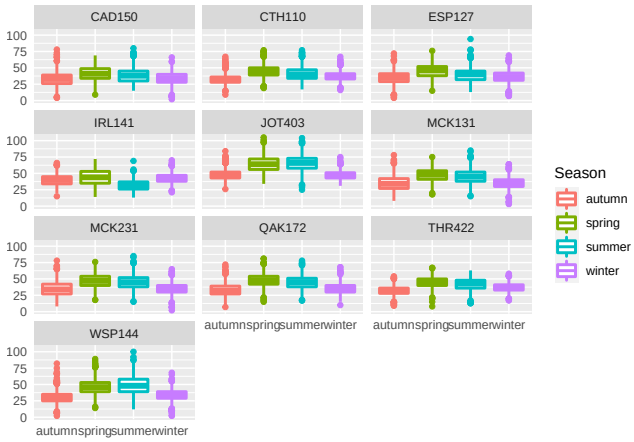



Figure 3: Boxplot of Ozone concentrations at each season for selected locations

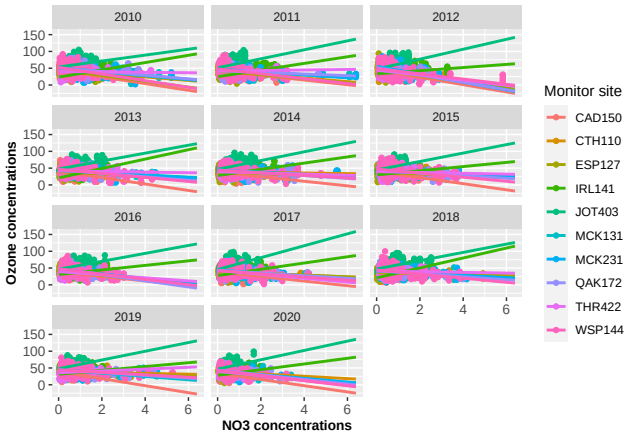



Figure 4: Scatter plot Ozone and NO3 concentrations for selected locations

### Model Fitting

Based on the exploratory data analysis results, I have decided to fit the study data using a linear mixed-effect regression model with Ozone concentrations being the response variable. Week, Season, Year different from 2010, NO3 concentrations, monitor sites, and the interaction term of Week and Season are included in the model as fixed effects. Further, to account for the correlation between Ozone concentrations measured at the same site, the monitor site is included as a random intercept. Similarly, Week is included as random effects to account for the correlation between Ozone concentrations measured within the same week. ANOVA test suggests that all the fixed effects are significant and thus are all included in the final model (See Table 2). Further, likelihood ratio tests for the random effect and the random intercept suggest that both of them are statistically significant at the confidence level of 0.05 (p_value $< 0.001$). Therefore, the final model includes Week, Season, Year different from 2010, NO3 concentrations, monitor sites, and the interaction term of Week and Season as fixed effects. Week and monitor sites are included as random effects.

Table 1: ANOVA table of the fixed effects

| | npar | Sum Sq | Mean Sq | F value | Pvalue |
|---|---|---|---|---|---|
| Mointor sites | 88 | 7142476.8064 | 81164.5092 | 1006.10209 | $< 0.001$ |
| Week | 1 | 695.7118 | 695.7118 | 8.62393 | $< 0.001$ |
| Season | 3 | 5702866.5731 | 1900955.5244 | 23563.93634 | $< 0.001$ |
| NO3 Concentrations | 1 | 297124.9087 | 297124.9087 | 3683.11217 | $< 0.001$ |
| Year | 1 | 446624.5539 | 446624.5539 | 5536.28553 | $< 0.001$ |
| Week:Season | 3 | 315392.7980 | 105130.9327 | 1303.18599 | $< 0.001$ |

## Result

The detailed results of the final model can be found in Appendix. Week, year difference, and NO3 concentrations are all found to be negatively associated with Ozone concentrations. That is probably caused by a very large fixed intercept of the final model, which is over 60. In addition, the ozone concentrations are the highest in Fall compared to the other 3 seasons, while holding everything else the same. From the plot of random effects (see Figure 5), we can see that the random intercept for each monitor site is quite different from each other, while the random effect of Week seems not to be very noticeable. Figure 6 shows fitted Ozone concentrations with respect to different predictors and 95% confidence interval (shaded area in the plot) according to the proposed model. The prediction plots are consistent with the estimated regression effects we have seen in the model. With one unit increase in SO3 concentrations, the the predicted Ozone concentrations will decrease by $-1.255$ units. During Fall, the Ozone concentrations will decrease by 0.565 units (ppb) every one week.

## Discussion

The estimated coefficients of fixed effect Week and NO3 concentrations are different from what we have seen in the explanatory data analysis. NO3 concentrations seem to be positively associated with Ozone concentrations at some monitor sites in the explanatory data analysis while NO3 concentrations are found to be negatively associated with Ozone concentrations in the proposed model. This actually reasonable since the marginal effect of NO3 can be different than the conditional effect of it while conditioning on other predictors. However, the linear mixed-effect regression does not perform well when it comes to nonlinear relationships between Week and Ozone concentrations. Further study can be conducted on models and methods for the nonlinear relationship between Week and Ozone concentrations. In addition, since the Ozone concentrations are densely measured over time, instead of longitudinal data, we could actually view Ozone concentrations as a functional covariate of time (Week or day) in a future study.

The variable, Season, is included in the model for accounting for the effect of temperature and humidity on Ozone concentrations(Zhang, Wei, and Fang 2019; Stathopoulou et al. 2008). However, it might not be

a very accurate indicator of temperature and humidity since whether varies, especially under the current situation of climate change. Information on temperature and humidity might be needed for future study.

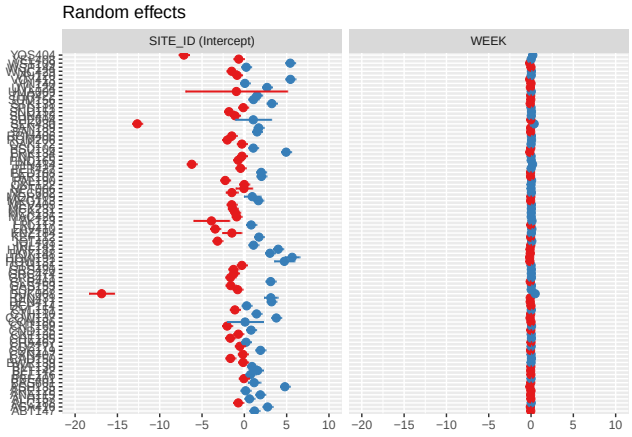



Figure 5: Random Effects of the final model

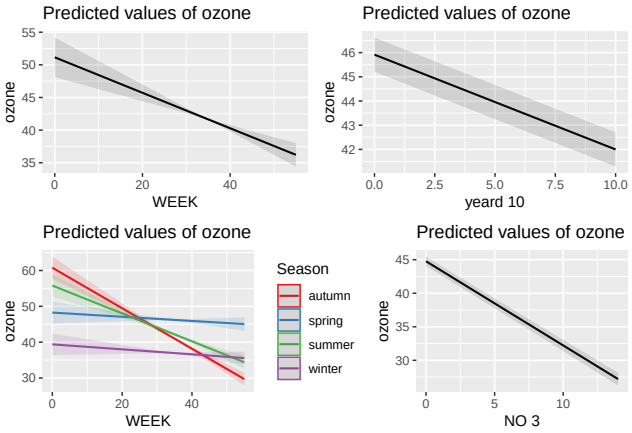



Figure 6: Prediction plot

## Reference


EPAR. 2020. "Report on the Environment." 2020. https://www.epa.gov/report-environment.

Staehelin, Johannes, Neil RP Harris, Cristof Appenzeller, and J Eberhard. 2001. "Ozone Trends: A Review." *Reviews of Geophysics* 39 (2): 231–90.

Stathopoulou, E, G Mihalakakou, M Santamouris, and HS Bagiorgas. 2008. "On the Impact of Temperature on Tropospheric Ozone Concentration Levels in Urban Environments." *Journal of Earth System Science* 117 (3): 227–36.

Zhang, Junfeng Jim, Yongjie Wei, and Zhangfu Fang. 2019. "Ozone Pollution: A Major Health Hazard Worldwide." *Frontiers in Immunology* 10: 2518.

## Appendix

### More EDA

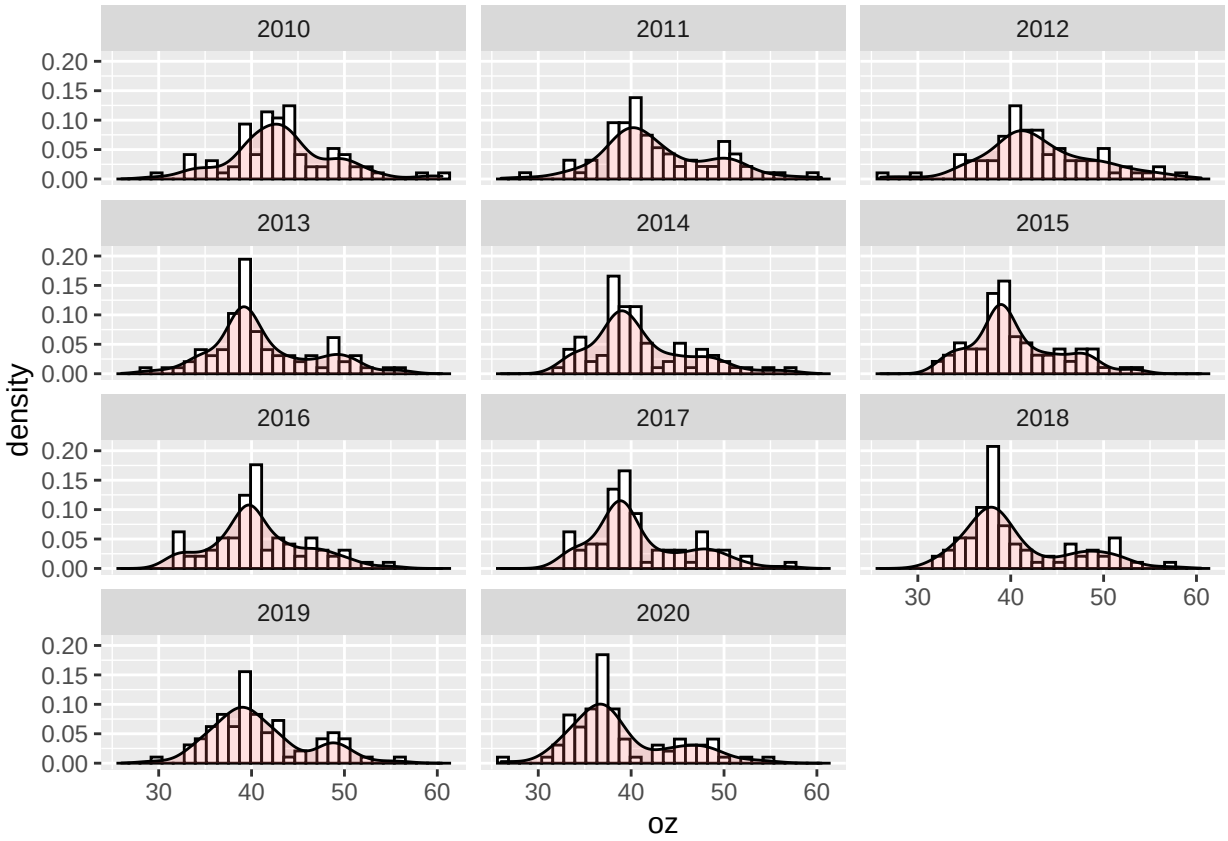



Figure 7: Histogram of average ozone across locations for each year

**model checking: residual plot**

```
plot(m1)
```

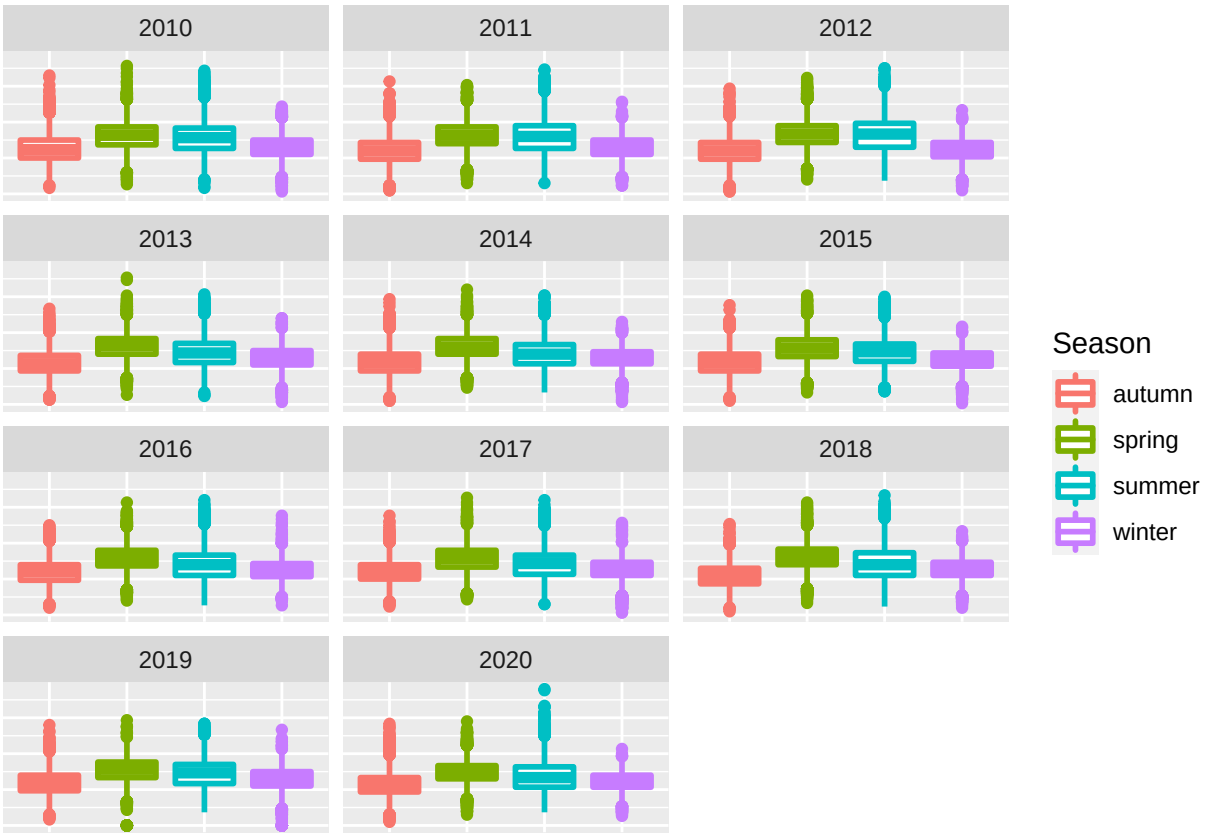



Figure 8: box plot over season for each year

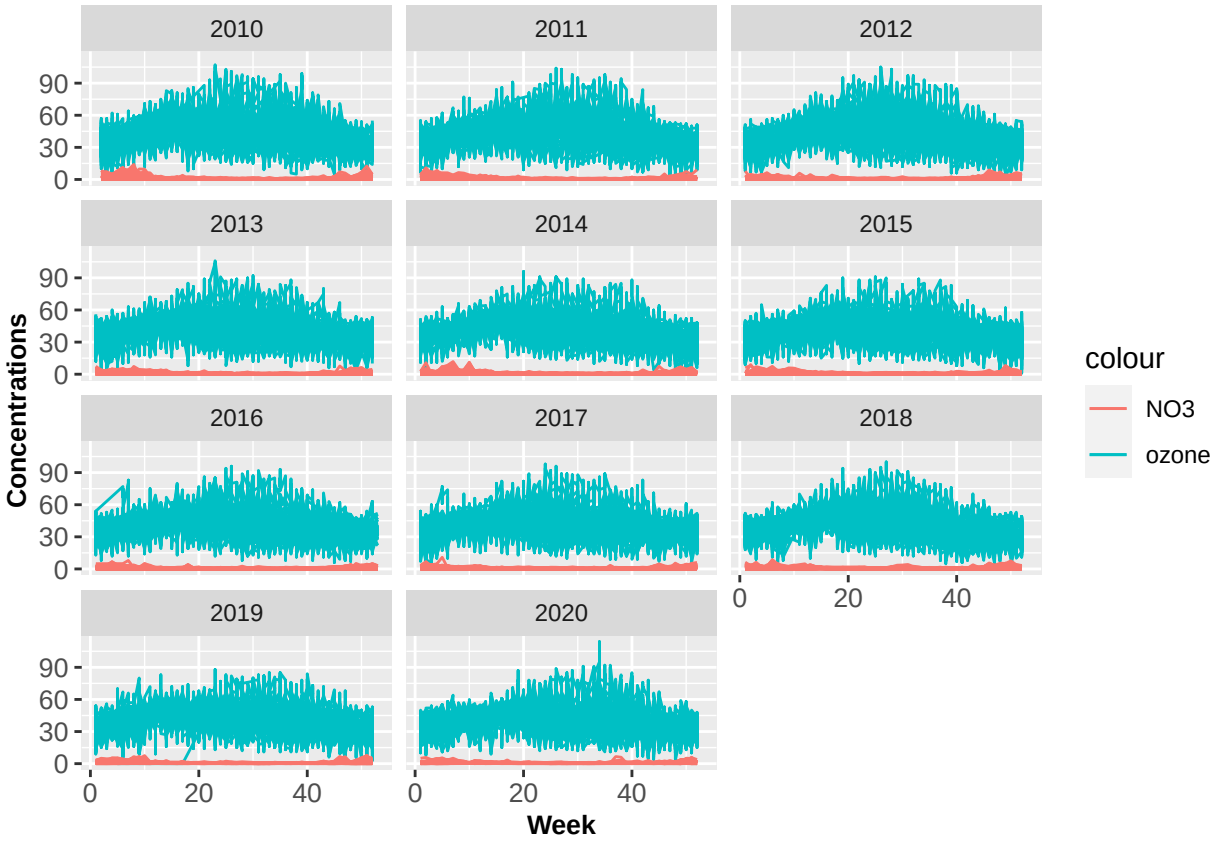
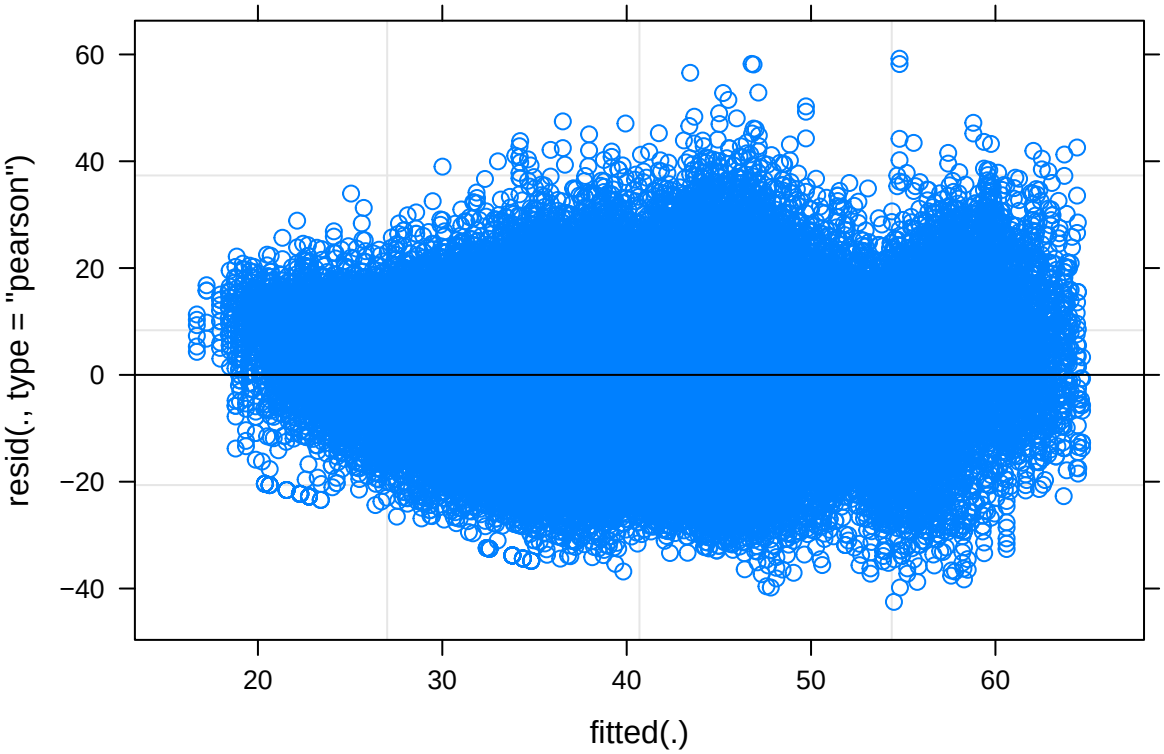



Figure 9: Average Ozone and NO3 concentrations across locations over time from 2010 to 2020


resid(., type = "pearson")
60
40
20
0
−20
−40
20
30
40
50
60
fitted(.)

**Full model results**

Table 2: Final Model selected

| | Estimate | Std. Error | t value |
|---|---|---|---|
| (Intercept) | 60.9929793 | 1.6016063 | 38.0823808 |
| ABT147 | -1.9529288 | 0.2350338 | -8.3091400 |
| ACA416 | -1.8401900 | 0.2432511 | -7.5649804 |
| ALC188 | 2.1296723 | 0.2368487 | 8.9917010 |
| ALH157 | 0.5142531 | 0.2391156 | 2.1506464 |
| ANA115 | 2.8879974 | 0.2384547 | 12.1113033 |
| ARE128 | -6.9724235 | 0.2352080 | -29.6436542 |
| ASH135 | 2.1340543 | 0.2623319 | 8.1349397 |
| BAS601 | 5.6050627 | 0.2398926 | 23.3648831 |
| BBE401 | 1.6625374 | 0.2395598 | 6.9399694 |
| BEL116 | 0.3485477 | 0.2458048 | 1.4179860 |
| BFT142 | 0.2828800 | 0.2384447 | 1.1863549 |
| BVL130 | 2.8971332 | 0.2356933 | 12.2919611 |
| BWR139 | -2.0198562 | 0.2383100 | -8.4757516 |
| CAD150 | 9.0212563 | 0.2365166 | 38.1421626 |
| CAN407 | -5.2901252 | 0.2360367 | -22.4123027 |
| CDR119 | 1.4793165 | 0.2379281 | 6.2174949 |
| CDZ171 | 9.5864528 | 0.2361222 | 40.5995440 |
| CHA467 | 2.8161464 | 0.2374609 | 11.8594128 |
| CHE185 | 0.7140380 | 0.2363426 | 3.0211990 |
| CKT136 | 0.2492811 | 0.2368985 | 1.0522695 |
| CND125 | 11.6376475 | 0.2437601 | 47.7422250 |
| CNT169 | 18.2829915 | 0.5638182 | 32.4271027 |
| CON186 | -3.7381602 | 0.2365035 | -15.8059392 |
| COW137 | -0.0589346 | 0.2359825 | -0.2497415 |
| CTH110 | -2.8362713 | 0.2369044 | -11.9722192 |
| CVL151 | 2.0538152 | 0.2377998 | 8.6367419 |
| DCP114 | -4.8912404 | 0.2343952 | -20.8674945 |
| DEN417 | 8.6258333 | 0.2720619 | 31.7054041 |
| DIN431 | -0.1169335 | 0.4304384 | -0.2716614 |
| DUK008 | 0.2174072 | 0.2348256 | 0.9258241 |
| ESP127 | 2.5848556 | 0.2369314 | 10.9097237 |
| GAS153 | -5.3868136 | 0.2385189 | -22.5844309 |
| GLR468 | 9.3092431 | 0.2361375 | 39.4229809 |
| GRB411 | 9.6652998 | 0.2352923 | 41.0778462 |
| GRC474 | 5.2403858 | 0.2393259 | 21.8964418 |
| GRS420 | 9.4506922 | 0.2371594 | 39.8495282 |
| GTH161 | -8.5994285 | 0.3907896 | -22.0052660 |
| HOW132 | -7.0494520 | 0.2901490 | -24.2959713 |
| HOW191 | -1.1040875 | 0.2362996 | -4.6724045 |
| HOX148 | -5.8928852 | 0.2389106 | -24.6656469 |
| HWF187 | 1.5580788 | 0.2393726 | 6.5090102 |
| IRL141 | 18.8434003 | 0.2348745 | 80.2275252 |
| JOT403 | 0.1489646 | 0.2353115 | 0.6330531 |
| KEF112 | 5.2342938 | 0.3865260 | 13.5418931 |
| KNZ184 | 6.7068090 | 0.2389151 | 28.0719306 |
| LAV410 | -0.0658163 | 0.2386110 | -0.2758311 |
| LRL117 | 4.6122776 | 0.8502939 | 5.4243335 |
| LYK123 | 2.5592004 | 0.2362084 | 10.8345010 |
| MAC426 | 2.5996433 | 0.2354822 | 11.0396610 |
| MCK131 | 2.6945870 | 0.2356963 | 11.4324553 |
| MCK231 | 10.0527926 | 0.2366247 | 42.4841137 |
| MEV405 | 0.5941824 | 0.2357675 | 2.5202045 |
| MKG113 | -12.4543007 | 0.3523258 | -35.3488145 |