ISYE-6414 Final Report - Air Quality Index

Ishika Arora, Ian Jiang, Tess Leggio

November 30, 2018

1 - Problem Statement

Air quality is a factor that is cited as both contributing to and being affected by climate change, and has been shown to have a direct impact on human health. As a factor with such important consequences, our team decided to study factors that impact air quality. For our project, we focus on answering the following question: *How is air quality across the United States related to demographic, emissions, and weather factors?* Specifically, we chose to investigate the impact of factors related to economic output and climate, including temperature, population, per capita income, greenhouse gas emissions, and industrial power consumption and generation. Our response variable is the Air Quality Index (AQI), a measure of local air quality calculated by the Environmental Protection Agency (EPA) based on the concentrations of 6 major classes of ground-level pollutants: ozone, carbon monoxide, nitrogen dioxide, sulfur dioxide, and two measures of particulate matter (PM2.5 and PM10). Upon building initial models of overall AQI, we found that the residuals differed based on the defining parameter (pollutant class) of the AQI. Thus, we ended up building models for the 6 individual pollutant AQIs in addition to the overall AQI.

The scope of our analysis ultimately included county-level data for all US counties with available AQI data. To limit the size of our dataset to a size manageable within R, we used daily data from 2016, the most recent year with data available for each of the factors included in our analysis.

In the following analysis, we give results for analysis preformed to generate one of the 7 models, the initial model for overall AQI. We used a similar approach to generate each of the models for the individual AQI defining parameters.

2 - Data Collection

Most of the data we used in our analysis is available from government websites. Data is available for download as CSV files from the following sources: per capita income data from bea.gov, population data from census.gov, powerplant energy consumption and generation data from eia.gov, and AQI data from aqs.epa.gov.

Weather data were significantly more challenging to collect. Daily weather sensor data is available from aqs.epa.gov, however the county data is far less complete than AQI data, so to avoid omitting a large portion of data from analysis, which could introduce bias into the model we instead collected temperature data from api.mesowest.net, an API which allows straight-forward programmatic extraction of data. By setting parameters in the URL, daily temperature and other weather averages can be scraped for any given day, state and county. We exploit this feature by looping through all counties in the AQI data and all days within our date range, sending a "get" request for each iteration. Data is available in JSON format, and includes daily averages for all stations within a given county. Therefore, a global measure of temperature and other weather factors is easily calculated by converting the JSON format into a Python dictionary and averaging across all stations. The speed of the script is highly dependent on local factors and internet speed, but on average it takes about 10 minutes per county for each of the 1053 counties included in our analysis.

3 - Data Cleaning

Upon collecting data from each of the 6 identified data sources, we were faced with the task of merging the datasets together. First and foremost, we had to ensure that the datasets merged together well. We

had to merge data based on state name, county name, year, month, and date, so it was essential that these fields were in the exact same format across data sources to avoid loss of data. In particular, we spent much time ensuring that county names were in the proper format. Since county name is a string field, it was not uncommon to see several different formats (for example, Saint John vs. St. John vs. StJohn). To avoid having several sparsely populated observations for each county, we wrote python scripts to format each dataset to use the same format and merge the datasets together.

After merging the datasets, we noted that there were several columns that were only sparsely populated, especially within weather data such as wind speed, relative humidity, and precipitation. As these data were sparse, we decided not to use them in our analysis because it may have introduced bias into the model. As such, we dropped the columns from our dataset. The remaining dataset was large and did not have a large number of missing values, so we decided to omit NAs from analysis rather than imputing missing data.

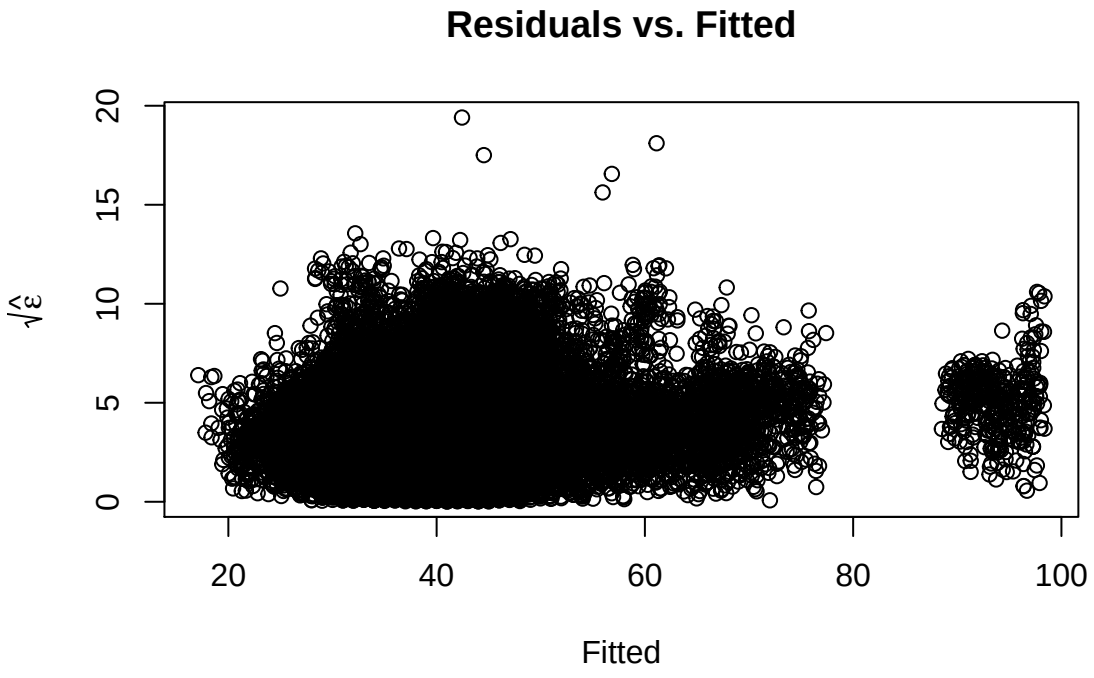
Next, we checked our merged data to ensure data quality. One of our more effective methods for investigating data quality was to plot individual predictors versus time to easily spot any clear outliers. Using this method, we were able to quickly spot clear outliers, such as the county Iberville, Louisiana, which reported temperatures of around -38C in the middle of the summer, which is not a reasonable value and did not align with checks on other websites. These values and few others were thus identified as clear data quality issues and were omitted from analysis.

Finally, our initial dataset across multiple years included more than a million observations and was too large a file to easily work with within R, so our team made the decision to trim our dataset to a more reasonable size of approximately 330,000 observations by including only data from 2016.

4 - Model Diagnostics

Checking Constant Variance Assumption

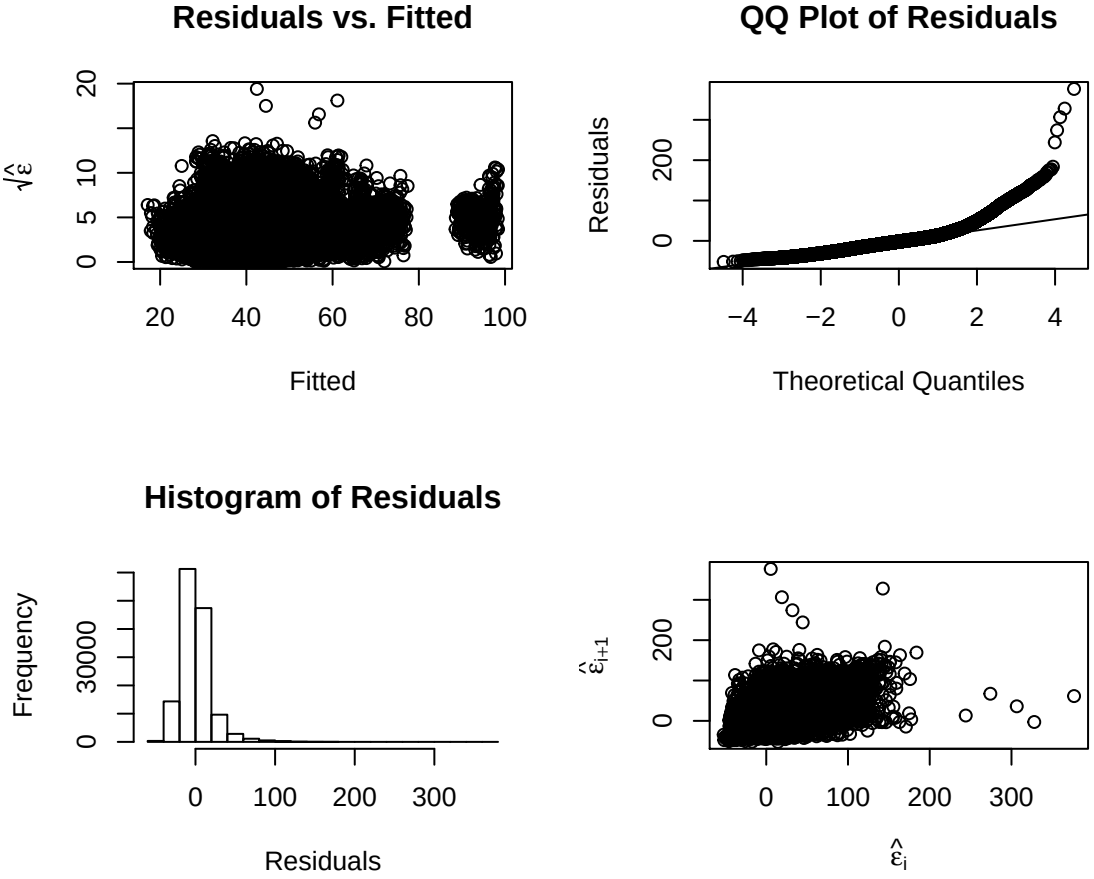
We check for the constancy of variance by examining the relationship of the residuals to the fitted values of the full model. If residuals follow a normal distribution, then their magnitudes would be half-normally distributed. Therefore, in order to mitigate the effects of the skewedness of this distribution, we also take the square root of the magnitudes, and plot it against the corresponding predictions.



When the square root residuals are fitted on the estimated response, we find a significant p-value, indicating that there is a dependence of the residuals. Therefore, transformation of either the response or factors are needed. For the moment, we leave the response alone, in order to proceed with the diagnostics.

Checking Normality Assumption

Next, we investigate whether or not the residuals of the model are normal distributed. For this, we visually examine the histogram of residuals to gain intuition around the distribution of residuals (though a histogram alone is not sufficient to make a determination about the normality of residuals), as well as the qq-plot of residual quartiles against the corresponding Gaussian quartiles.



As seen in the above figures, the residuals of the model appear to be right-skewed. The non-normality of the residuals are confirmed using a Shapiro-Wilks test. The p-value for this test is highly significant at the 5% level, so we reject the null hypothesis that the errors are normally distributed. Further investigation into ways to resolve non-normality will be needed.

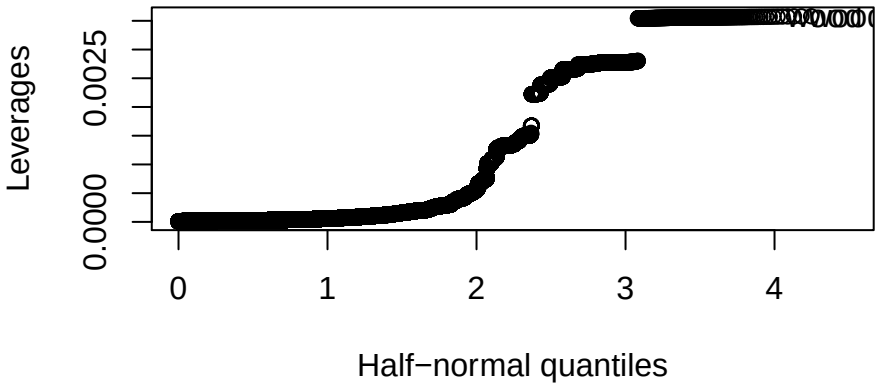
```
##
##  Shapiro-Wilk normality test
##
## data:  lmod$residuals[1:4999]
## W = 0.86576, p-value < 2.2e-16
```

Checking for Serial Correlation

Next, we check the serial correlation in the error, to determine if there is any serial dependence of the residuals. This is achieved by regressing each residual on the previous to determine if there is any trend. The R squared p-values show that this result is significant, and we can conclude that there is marked serial correlation among the residuals which may be caused by a missing covariate. We plot successive pairs of residuals to visualize this serial correlation in the bottom right of the following set of figures.

Unusual Points

Next, we address unusual points. As a point of clarification, this is to be distinguished from the examination of extreme points in the data cleaning. In the cleaning, we remove spurious results that are likely due to data entry or data collection errors (such as a failed sensor), while the remaining extreme points are assumed to be naturally-occuring. We first examine the leverages. We produce a qqplot of the residual quantiles against half-normal quantiles to investigate high-leverage points.



In order to preserve the stability and representability of the model, we try removing points with extremely large leverage (h > 0.2). As we have a very large dataset, this does not much change the model. Next, in order to identify outliers, we calculate the studentized residuals of each point, and compare them to the Bonferroni-corrected 5% quartile of a t-distribution with n - p - 1 degrees of freedom. Here we investigate potential outliers with studentized residuals greater than the Bonferroni criterion, which represent about 0.3% of our dataset.

```
#Calculating Studentized Residuals
stud <- rstudent(lmod)
n = nrow(data)
df = lmod$df
bonferroni <- abs(qt(.05/(n*2),df))
length(stud[abs(stud) > bonferroni])/n
```

```
## [1] 0.003054399
```

Multicollinearity

To check collinearity of the predictors, we first examine the variable inflation factors for each predictor to identify those with a high (>5) VIF. A VIF that is high indicates that the standard error is much larger than it would be without collinearity. We identify 7 factors with high VIF values, so it is clear some predictors are collinear. Using this model with highly collinear predictors leads to unstable models, which leads to drastic changes in the coefficient estimates with even small perturbations to the response. We produce and investigate the correlation matrix to better understand which predictors are related to one another. Armed with this understanding, we remove one high-VIF predictor at a time and rerun analysis so as to ensure we are only

only removing predictors that provide redundant information to the model. Using this process, we choose to remove the following variables: Short.Lived.Compounds, CO2, Total.Emissions, pp_net_gen_MWh, Other.Fluorane, and Nitrous.Oxide.

```
##             Other.GHG       Total.Emissions                   CO2
##             460.52648             185.77343            183.35615
##                   HFE    pp_consumed_MMBtu Short.Lived.Compounds
##             323.22094              66.33428             827.89949
##        pp_net_gen_MWh
##              63.88733
```

```
#Calculating correlation matrix
corr <- round(cor(data[,c("AQI", "HFC", "Other.GHG", "SF6", "Stationary.Combustion", "Biogenic.CO2", "H
```

5 - Model Selection

A) Stepwise selection

We tried running stepwise regression on the model before removing the collinear variables from the model, but the results were not very satisfactory. It removed only two factors out of the 20 predictors. However, when we looked at the pairwise correlation factors (and plots) and the VIFs, several factors seemed to be strongly correlated. This suggested that we had to remove more than just two factors from the model to address multicollinearity.

B) Backward selection

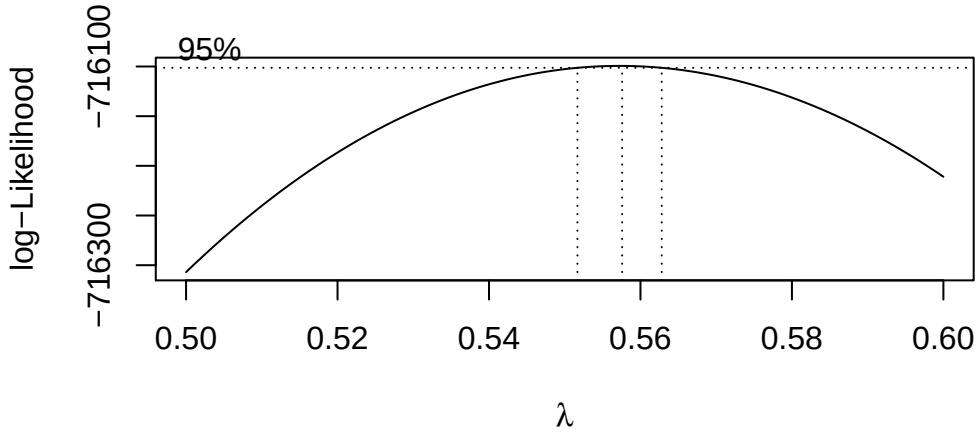
Because of the complex interaction of the factors, we decided to use backward elimination for removing non-significant factors from the model. We started with 13 predictors in the model and then removed the predictor with highest p-value greater than 0.05. We removed two more factors from the model and finally used the following factors to train our model - - *other greenhouse gases, hydrofluorocarbons, biogenic CO2, population, perfluorinated chemicals, hexafluorethane, stationary combustion, industrial power consumption, Temperature, sulfur hexafluoride, and per capita income.*

6 - Transformations

Based on the model diagnostics, specifically the violations of constant error variance and normality, it was clear we needed to modify the initial model. Several transformations were performed and tested to see if any transformations resulted in a better-fitting model. Some of these included square root transformations, adding polynomial terms for the predictors, investigating interactions between predictors, using log transformations, and others. Adding polynomial terms for Temperature improves the model fit. Most of the other transformations tested only improved model fit marginally while making it more difficult to interpret the model. Ultimately, we used Box Cox to identify the best transformation on the response.

Box Cox identified a lambda value of 0.56, so for model interpretability we used a square root transformation on the response. This helped resolve some, but not all of the error assumption violations while improving the R squared value. All predictors remain significant in this model. We have a relatively large dataset, and in this case it is not uncommone for small deviations from non-normality to be detected. We continue, however, with this model because we have a large sample size and can lean on the central limit theorem. The errors are not exactly normal, so the least squares estimate may not be optimal, though it will remain the best linear unbiased estimate. Investigations into robust regression methods, which are used especially when errors are not normally distributed, reveal residual standard errors similar to that of the transformed model.

Upon completing all of these steps, we repeat analysis steps as necessary as we build and update each of the models for the 6 air quality defining parameters.




```
##                            Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)             6.0011e+00  1.7706e-02 338.9232 < 2.2e-16
## Other.GHG               1.5655e-04  2.8053e-05   5.5803 2.405e-08
## HFC                     3.0298e-07  2.7492e-08  11.0205 < 2.2e-16
## Biogenic.CO2           -2.4644e-07  1.0623e-08 -23.1983 < 2.2e-16
## Population              4.2803e-07  6.1724e-09  69.3462 < 2.2e-16
## PFC                    -1.0594e-06  6.7784e-08 -15.6294 < 2.2e-16
## HFE                    -4.8943e-06  9.1965e-07  -5.3219 1.029e-07
## Stationary.Combustion   8.2435e-09  1.9423e-09   4.2441 2.196e-05
## pp_consumed_MMBtu      -8.6269e-09  5.2637e-10 -16.3896 < 2.2e-16
## poly(Temperature, 2)1   1.1588e+02  1.4904e+00  77.7506 < 2.2e-16
## poly(Temperature, 2)2   7.0885e+01  1.4705e+00  48.2031 < 2.2e-16
## SF6                     3.9913e-06  2.5091e-07  15.9072 < 2.2e-16
## Income                  3.0867e-06  3.8428e-07   8.0324 9.634e-16
##
## n = 137829, p = 13, Residual SE = 1.46290, R-Squared = 0.11

## Robust Regression residual standard error: 1.10453
```

7 - Final Model

Based on the diagnostics and transformation results (as shown above), we built the final models for each defining parameter by transforming the response as square root. We used the following set of factors to train the model, which were selected based on the results of multicollinearity tests in the diagnostics section and backwards elimination - *nitrogen trifluoride, other greenhouse gases, hydrofluorocarbons, biogenic CO2, population, perfluorinated chemicals, hexafluorethane, stationary combustion, industrial power consumption, Temperature, methane, sulfur hexafluoride, and per capita income.* The R-squared value of the full final model comes out to be 10.7%. All the predictors are highly statistically significant in this model as shown in the model summary below.

```
lmod.T <- lm(formula = AQI ^ 0.5 ~ Other.GHG + HFC +  Biogenic.CO2 + Population + PFC + HFE + Stationary

sumary(lmod.T)
```

```
##                            Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)             5.8952e+00  1.8266e-02 322.7460 < 2.2e-16
## Other.GHG               1.6091e-04  2.8940e-05   5.5603 2.698e-08
```

```
## HFC                     3.1022e-07  2.8361e-08  10.9383 < 2.2e-16
## Biogenic.CO2           -2.4791e-07  1.0959e-08 -22.6220 < 2.2e-16
## Population              4.3650e-07  6.3674e-09  68.5521 < 2.2e-16
## PFC                    -1.0678e-06  6.9925e-08 -15.2703 < 2.2e-16
## HFE                    -5.1409e-06  9.4871e-07  -5.4188 6.009e-08
## Stationary.Combustion   8.5422e-09  2.0037e-09   4.2632 2.017e-05
## pp_consumed_MMBtu      -8.9469e-09  5.4300e-10 -16.4770 < 2.2e-16
## poly(Temperature, 2)1   1.1899e+02  1.5375e+00  77.3914 < 2.2e-16
## poly(Temperature, 2)2   7.1503e+01  1.5170e+00  47.1346 < 2.2e-16
## SF6                     4.1047e-06  2.5884e-07  15.8580 < 2.2e-16
## Income                  3.3317e-06  3.9642e-07   8.4044 < 2.2e-16
##
## n = 137829, p = 13, Residual SE = 1.50912, R-Squared = 0.11
```

Similarly, we built the final component models for each defining parameter with response transformation and selected factors. Here we show the results here for the NO2 AQI model, for which the final R-squared value came out to be 27.1%.

```
## NO2 Model R-squared:  0.27
```

8 - Conclusions and Future Research

Through our analysis, we conclude that the factors used in each of our models, primarily *nitrogen trifluoride, other greenhouse gases, hydrofluorocarbons, biogenic CO2, population, perfluorinated chemicals, hexafluorethane, stationary combustion, industrial power consumption, Temperature, methane, sulfur hexafluoride, and per capita income*, are significantly related to the air quality of a given area. However, despite highly significant predictors, the relatively low R squared values of the models in addition to somewhat heteroscedasticity in model residuals indicate that there are likely important covariates missing or an underlying nonlinear relationship that may be explored in further analysis.

A) Additional predictors

Future research work can include additional predictors in the model, since only up to about 27% of the variation in the response is explained by the current set of predictors. Some additional factors that can be included are - *more weather factors (wind speed, humidity and precipitation), geographical data like elevation or use city-level data rather than county-level data.* This may help explain more of the variation in the response as well as help address the serial correlation present in the current fit.

B) Additional scope

We can further investigate air quality with more data collected across the globe so that we have more informational dataset. We can also run the analysis over a larger time interval (current analysis is for a year's worth of data). We could not process more than a year of data in R, so we decided to build the model for the smalle dataset.

C) Nonlinear models

Initial investigation into general additive models seem promising, as even the unrefined model produces a higher percentage of deviance explained than our best linear model for total AQI. Further analysis is required to better understand if there is an underlying nonlinear relationship in predicting air quality index.

```
## GAM % deviance explained:  0.161007
```