Predicting Ozone Levels Using Weather Simulation Models

Stat 536 Midterm

W. Zachary Horton

Brigham Young University

Abstract. The EPA monitors ground-level Ozone, a pollutant that is harmful to respiratory function. Estimating Ozone pollution levels can be difficult because the nearest weather station will give an innacurate representation if it is far from the location of interest. A simulation based model called CMAQ can be used to calculate Ozone levels at any level of granularity, but has been found to be innacurate. We seek to explore the relationship between CMAQ and the true Ozone concentration by using a linear model with spatial correlation. We assess the predictive power and accuracy of the model and discuss its use in predicting Ozone concentrations.

1 Introduction

Ground-level Ozone (O3) is a pollutant formed by the combination of Nitrogen-Oxides and Volitle Organic Compounds, both of which are produced when burning fossil fuels. Exposure to high concentrations of O3 can trigger a variety of respiratory issues such as asthma and bronchitis. Long-term exposure has even been linked to significantly reduced life expectancy. The Environmental Protection Agency has regulaions in place that help reduce O3 pollution levels, however they can be costly and unnecessary if O3 concentration is low. Weather stations provide information on O3 levels, but these stations are sparsely scattered across the country. A mathematical O3 simulation model called the Community Multi-scale Air Quality Model (CMAQ) has been developed to predict O3 levels in all areas using easily obtained local variables, but researchers have found that this model is not accurate enough when compared to actual measurements from weather stations. The goal of this analysis is to develop a method of estimating O3 concentration in a given area that combines both the CMAQ predictions and regional O3 readings in order to better appropriate resources to more highly polluted areas.

2 Data

We use data measured in an eight hour window on May 22, 2005 in locations all over the midwestern and eastern portions of the USA. There are 800 weather stations that reported average O3 concentration and almost 67,000 CMAQ predictions that cover the same region. Each location is specified by a longitude and lattitude. Table 1 shows numerical summaries of the weather station measurements and the CMAQ estimations. Note that both variables must be positive.

Table 1. Summary of Ozone Measurement Data

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Station	7.13	43.75	52.50	51.45	57.78	106.63
CMAQ	25.50	49.02	53.53	53.01	58.00	97.74

Figures 1 and 2 below show the weather station O3 measurements and CMAQ predictions. Notice that spatial correlation exists within the data.

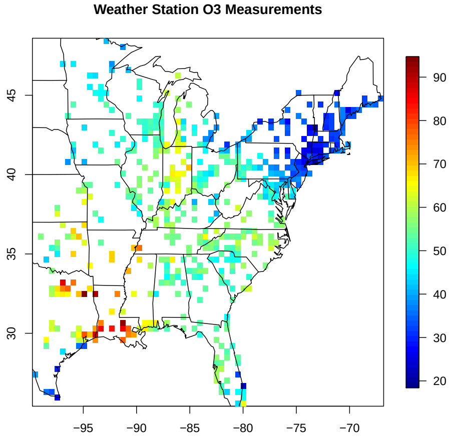

Fig. 1. Heat map of weather station measurements. White space means no measurements.

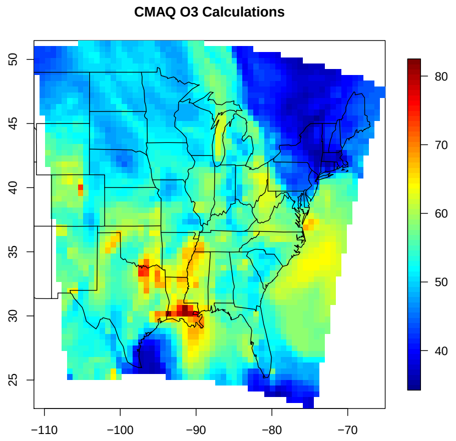

Fig. 2. Heat map of CMAQ calculations. More dense than the weather stations.

Figure 3 below shows how the O3 weather station measurements compare to the estimated CMAQ values in the locations where both have been measured. Figure 4 shows similarly with the average of the four closest CMAQ values instead of just the closest. Notice they have a linear relationship with the true O3 measurements.

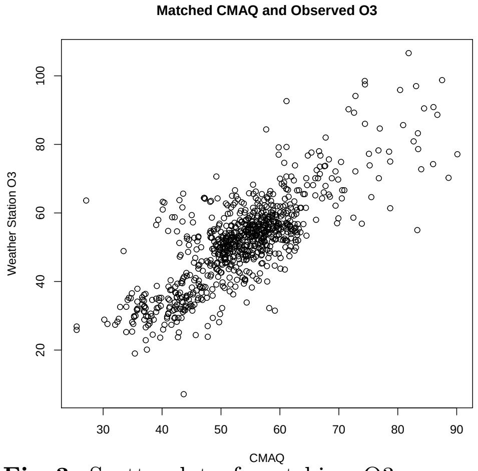

Fig. 3. Scatterplot of matching O3 measurements and CMAQ predictions

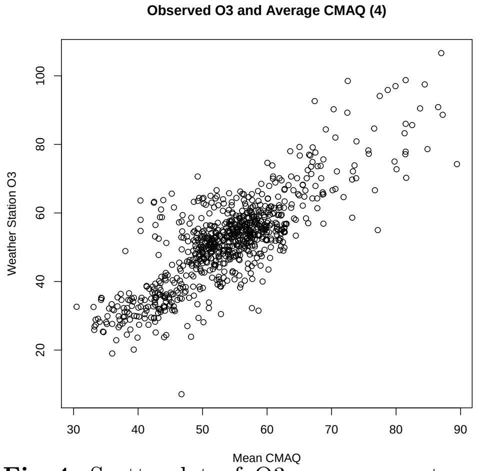

Fig. 4. Scatterplot of O3 measurements and averages of nearest four CMAQ values.

3 Model and Methods

In order to assess the relationship between measured O3 levels and CMAQ predictions, we fit a linear model with spatial correlation that used CMAQ as an explanatory variable and measured O3 as the response. The model can be written as:

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \varSigma) \tag{3.1}$$

where $\mathbf{y}$ is the nx1 vector of measured O3 values, $\mathbf{X}$ is the nxp design matrix that corresponds to the px1 parameter vector $\boldsymbol{\beta}$, and $\varSigma$ is the covariance matrix. We used measured O3 as the response because one goal of this study is to predict actual O3 levels using a given CMAQ value. We chose to fit a linear relationship between weather station O3 and CMAQ, meaning that $\boldsymbol{\beta} = [\beta_0 \beta_C]'$ where β_0 is an intercept and β_C is the slope. The assumptions of this model consist of linearity between the explanatory and response variable, normality about the fitted line, and equal variance about the fitted line. Although typical linear regression also assumes independent data, we chose to model using an exponential spatial correlation structure. The corresponding design of $\varSigma = \sigma^2\mathbf{R}$ is $exp\left\{\frac{-||\mathbf{x}_i - \mathbf{x}_j||}{\phi}\right\}$ at the i^{th} row and j^{th} column, all with a scaling variance term σ^2. The parameter ϕ is called the range and is related to the distance that spatial correlation makes an effect. We will also consider adding a nugget effect ω since it often improves spatial analysis. The nugget changes covariance to be $\varSigma = \sigma^2\left((1-\omega)\mathbf{R} + \omega\mathbf{I}\right)$ where $\mathbf{I}$ is the identity matrix.

There are two nuances in the data that this model needs to address. The first is the spatial correlation in the O3 observations. By allowing for an exponential structure, this has been accounted for. We chose an exponential spatial structure because it accounts for relative distance between points, irrespective of other coordinate features. The other nuance is the sctict positive nature of the data. Although the normality assumption allows for negative values, the data seen in table 1 show that there is no data near zero, thus the model will virtually never assign any weight to negative values, making this a nonissue.

This model allows us to estimate the relationship between observed O3 and CMAQ by providing an estimated intercept and slope that describe how O3 changes as different CMAQ values are seen. The model also provides a correlation-adjusted variance term which can be used to assess uncertainty in the model through intervals. These estimated parameter values form a line and intervals which can be used to predict O3 values in regions where there are no nearby weather stations.

In this study we consider two different ways of including CMAQ in the model. The first is to match observed O3 with the nearest CMAQ coordinate. The second is to take the average of the four nearest CMAQ values. We call these the matched CMAQ and the local mean CMAQ. The averaging may be useful in controlling for rogue predictions in a single CMAQ value. We will compare the performance of each variable as used in seperate models.

4 Model Justification and Performance

Using the AIC, we found that the best model used the local mean CMAQ and included a nugget in the covariance structure. The models that did not include a nugget or used the matched CMAQ all had higher AIC values. We also fit a model that did not include correlation and performed a likelihood ratio test. The p-value for the significance of the spatial correlation structure was less than .0001, so we kept the correlation structure. For the duration of the report we will use this model.

We now assess the validity of the assumptions. Figure 4 plots the local mean CMAQ against the observed response O3. This shows that linearity is met. Figure 5 is a histogram of the decorrelated residuals, showing that the normality assumption is met. Decorrelated residuals can be obtained by taking the inverse of the Cholesky decomposition of Σ and multiplying it by $\mathbf{X}$ to obtain a new design matrix. The resulting residuals from this decorrelated model can be used to assess both normality and equal variances. Figure 6 is the decorrelated residuals versus fitted values plot. With the exception of the outlier at the bottom, there doesn't appear to be a growing or shrinking band about the center, indicating that the equal variance assumption is tentatively met.

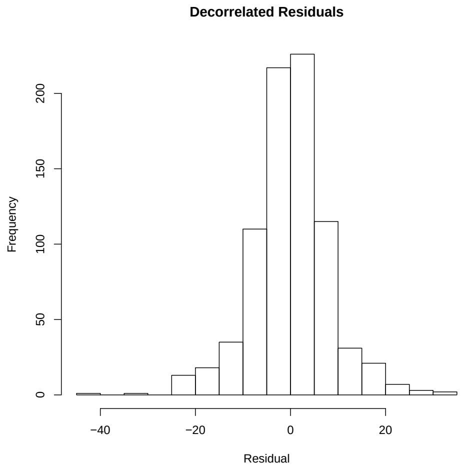

Fig. 5. Histogram of decorrelated residuals. Shows normality assumption is met.

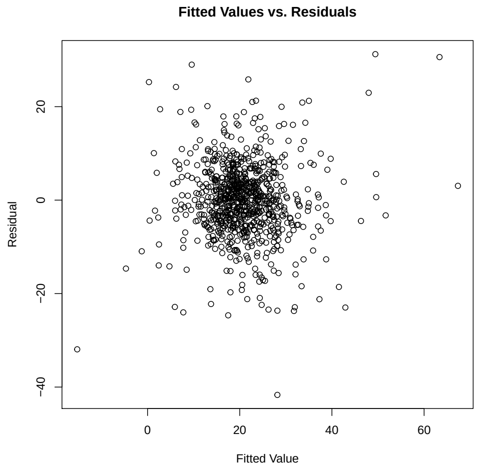

Fig. 6. Fitted value by residual plot. Shows equal variance assumption is met

To assess how well the model fits the data, we report R^2 percent of variation and RMSE. R^2 can be obtained using the decorrelated model. This model calculates $R^2 = 0.8808$, meaning that 88% of the variation in observed O3 can be explained by the local mean CMAQ and the spatial correlation structure. That high percent indicates the model fits the data pattern well. To obtain root mean squared error (RMSE) we perform a cross-validation study where we randomly remove 10 O3 observations, fit the model, and then compute prediction MSE. Because the correlation is based on distance, removing points will preserve the correlation structure. We repeat this 10 times, find an overall average MSE, and then square root. We found that RMSE = 2.8, meaning that predictions are off by 2.8 pollution units on average. This number is very low compared to the range of the data, indicating that predictions are quite accurate.

5 Results

Table 2 below shows the model output. Note that the 95% confidence intervals have been created using a t-distirbution with 795 degrees of freedom.

Due to the correlation structure, the coefficient line does not run perfectly through the data. When predicting new data, the correlation structure is accounted for and will move predictions off the line relative to which data points are close and correlated. Figure 7 shows both the coefficient line and the decorrelated regression line, which gives a sense of how well the model fits.

What these results mean is that there is a linear relationship between local CMAQ mean and actual O3 concentration. These results also show that O3 is certainly correlated spatially. Using the model coefficients

Table 2. Spatial Linear Model Output

Parameter	Estimate	95% Interval
Intercept	12.538	(5.45,19.62)
Local Mean CMAQ	0.679	(0.574,1.728)
Range ϕ	3.853	
Nugget ω	0.235	
σ^2	71.38	

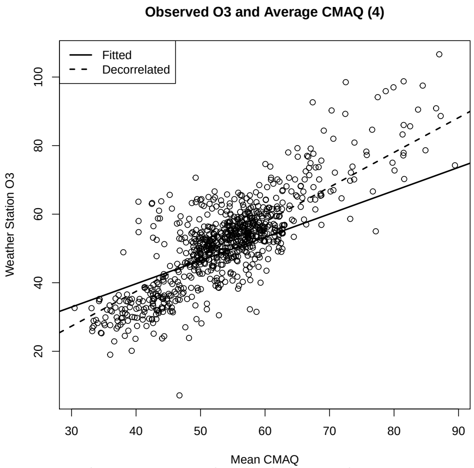

Fig. 7. Fitted Model on the Data

and the covariance terms, one can predict O3 with a high level of accuracy. In general, the slope parameter can be interpreted as an average increase in O3 by 0.679 for every one unit increase in local mean CMAQ, but then modified according to the correlation. Due to uncertainty, with 95% confidence that average increase could reasonably be anywhere within (5.45,19.62), again before accounting for correlation.

5.1 Prediction Application

To see the predictive power of the model, we predicted O3 levels in over 2,500 new locations. Figure 8 shows a heat map of the predicted values. Figure 9 is a heat map that shows the standard error of those predictions. Notice how the values are low except for in areas where the predicted values are in the extreme.

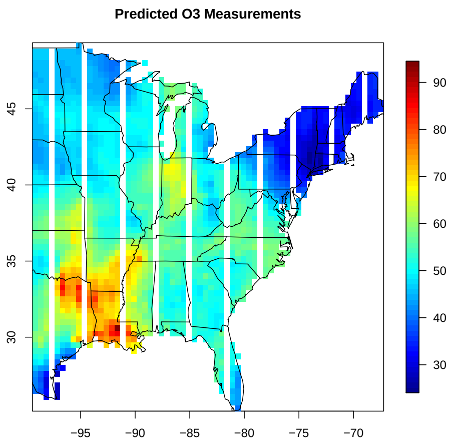

Fig. 8. Heat map of predicted O3 at new locations.

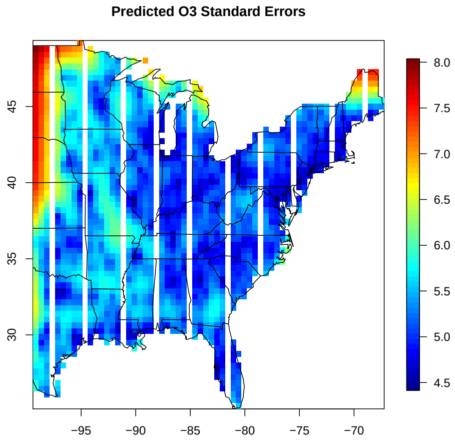

Fig. 9. Heat map of prediction standard errors showing uncertainty.

Based on these results, it seems that the Louisiana region has the highest concentration of O3. Notice that these results closely resemble the data in figure 1. It is interesting to see that the uncertainty seems constant relative to the magnitude of O3. Rather, it seems that uncertainty is only increased near the borders of the CMAQ data.

6 Conclusion

The goals of this study were to quantify the relationship between CMAQ and observed O3 levels and to predict O3 in new locations. We found that a linear regression that accounts for spatial correlation models the data well. We found that mean local CMAQ served as a well fitting variable, better than simply using the matching CMAQ value. We also found that this model predicts well with a very low RMSE. A short coming of this model is seen in the assumption plots. There are a handful of outliers that were not accounted for very well in the model. Future research would be to explore the outliers as well as to explore more complicated variable choices and correlation structures. In the end, the ability to accurately estimate O3 levels could enhance our efforts against pollution by helping to pinpoint problematic areas.