Ground Level Ozone Analysis

Jacob Merrell

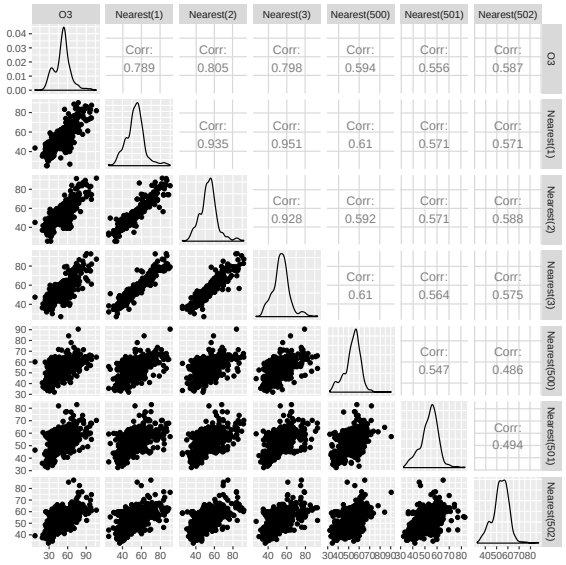
Brigham Young University

Abstract. O3 is the main component of smog. High amounts of O3 can cause several health problems. CMAQ is a method to estimate the O3 levels in an area. This study explores the relationship between CMAQ measurements and O3 levels. We also construct a statistical model that deals with the spatial correlation between CMAQ values at different locations, and provides predictive power. Predictions of O3 levels are calculated across most of the United States. The predictions show which geographical areas have low levels of pollution and those that display dangerous levels of pollution.

1 Introduction

Ground-level ozone (O3) is caused by the outputs of cars, industrial processes, refineries, manufacturing and more. These outputs combine together and "bake" in the sunlight creating O3. Smog is mainly comprised of O3 and is detrimental to health. Smog can cause chest pain, bronchitis, asthma, and more. The Community Multi-scale Air Quality Model (CMAQ) estimates the O3 levels based on temperatures, urban density, etc. However, the CMAQ estimates don't exactly match the actual O3 measurements. The purpose of this analysis is to: (1) understand the relationship between CMAQ and O3 (2) predict the O3 levels using CMAQ measurements. Regression techniques to handle spatially correlated data will be used to achieve the goals of the analysis.

2 Exploratory Data Analysis



The graph above shows the relationship between O3 levels and the explanatory variables calculated for the X matrix. The columns labeled "Nearest(i)" represent data for all the ith geographically nearest CMAQ measurements to the O3 measurement locations. The "Nearest(1)" column shows a 0.789 correlation with actual O3 measurements. There is a strong positive relationship between CMAQ and O3; as O3 rises, the CMAQ measurement is expected to increase as well. It should be noted that the CMAQ of the nearest location is highly correlated with the CMAQ of the second closest location. As the points get farther away from the O3 measurement location, collinearity decreases. Collinearity in the explanatory variables can inflate the standard errors and create wider prediction intervals. Even though there is collinearity in the explanatory variables, the prediction intervals, as will be seen further in the study, cover almost %95; the intervals in this case are not too wide. There are no non-linear relationships in the data.

3 Model Selection

The model used for this analysis is

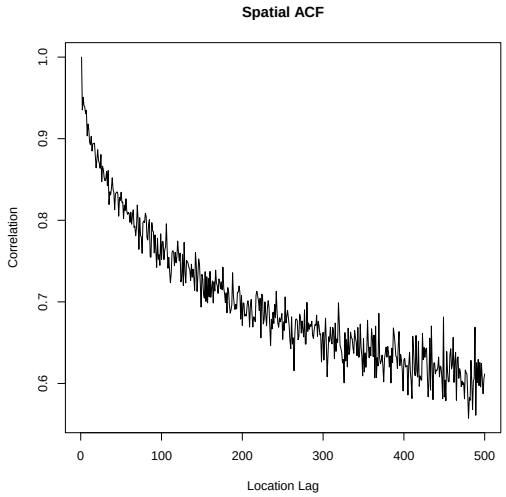
$$Y = \begin{pmatrix} Y_{(s_1)} \\ \vdots \\ Y_{(s_N)} \end{pmatrix} \sim N(X\beta, \sigma^2((1-w)R + wI)) \tag{1}$$

In the model $Y_{(s_i)}$ denotes the O3 measurement at location i, and the X matrix contains the CMAQ values for the locations geographically nearest to the O3 measurement at location i; euclidean distance calculated

from the latitudes and longitudes determines which locations are nearest. The β vector contains the model coefficients, σ^2 is the variance of the residuals, R is the correlation matrix, and w is the nugget. The nugget accounts for the fact that two sample points at the same location may show different measurements, so the nugget adds randomness; spatial statistics almost always uses a nugget variance. The correlation matrix in this model has the following exponential structure: $R_{ij} = exp\{-\frac{\|s_i - s_j\|}{\phi}\}$

In the exponential correlation structure above, $\|s_i - s_j\|$ is the euclidean distance between points, and ϕ is the range parameter estimated by gls function in R used to run the model. The auto-regressive and moving average correlation structures don't make much sense to use in the context of this problem since the locations are not evenly spaced. The correlation structure shown above allows for unevenly spaced distances.

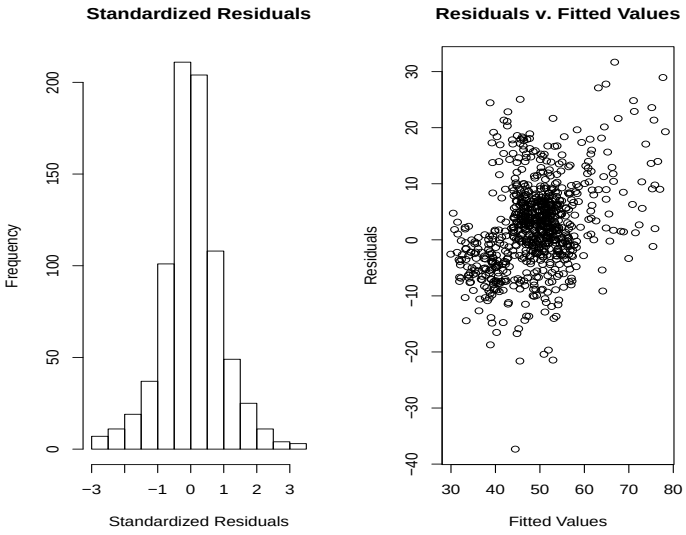
The model assumptions are normality of the standardized residuals, homoscedacisity, and that the data are multivariate normally distributed. Normality of the residuals can be observed by plotting the residuals to verify if they seem normally distributed. Homoscedacisity is verified through a graph of the fitted values v. the residuals. A constant variability or jitter of the residual values about 0 should be similar across all fitted values. The model follows multivariate normal distribution.



The graph above shows the spatial auto-correlation. Locations that are near to each other have highly correlated CMAQ values. As the locations become farther apart, their correlation decreases. The model will allow us to account for the spatial correlation from one location to another and make predictions about the level of O3 at each location. The closest CMAQ measurement to each O3 measurement will help us understand the relationship between CMAQ and O3.

4 Model Justification

The support of the data in this study are positive numbers; there can't be a negative O3 or CMAQ measurement. However, there are no O3 nor CMAQ measurements that come close to 0. Since the predictions are a weighted average of a group of CMAQ values all far greater than 0, then predictions are highly likely to be positive as well. We calculated the CMAQ of the 1,000 closest locations to each of the O3 measurement locations, and these are the explanatory variables being explored. To determine how many locations should be included in the model, we used forward selection methods; exploring all possible subsets of models is computationally inefficient. The forward selection showed that the CMAQ of the 17 closest locations to each O3 station produced the model with the lowest AIC; AIC was the criterion chosen since the main focus of this study is prediction. Since the data are correlated, a decorrelated regression model allows us to check the model assumptions. Let L = lower cholesky decomposition of the correlation matrix R, then we create a regression model of $L^{-1}Y$ on $L^{-1}X$. The residuals and fitted values from the decorrelated regression model can be used to verify whether or not the assumptions hold. The assumptions for the model are explored in the graphs below



The histogram of the standardized residuals shows a normal distribution. The fitted v. residual plot shows the variance of the residuals about the 0 line is about the same. All assumptions hold for this study.

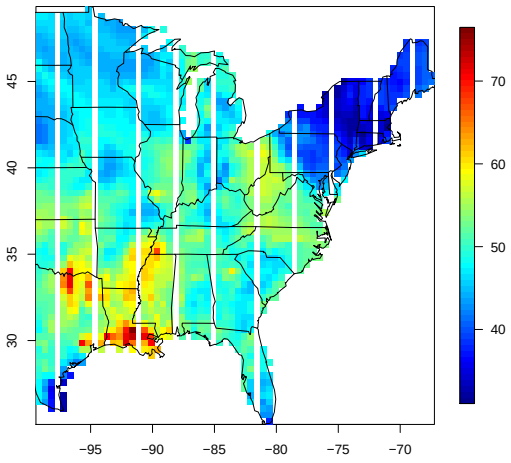
5 Performance Evaluation

The model had an adjusted R-squared of 0.6129. This mean 61.29% of variation in O3 is explained by the model. To test the predictive power of the model, we used test and training data to cross validate. The test

and training datasets were chosen to preserve the spatial correlation structure. Instead of choosing random locations as the test dataset, we split the geographical area for which we had data into 10 segments; each segment would be predicted using only the data from the other nine. The geographical region was split vertically and horizontally into 10 regions. The training data was used to predict the test data. The RMSE, bias, and 95% prediction interval coverage we calculated. On average, the RMSE was 8.79; this means the predictions were off by 8.79 O3 units on average. The bias showed that we were underestimating the O3 level by 2.36 on average, and almost 95% (94.25%) of all O3 values were contained within the prediction intervals. As mentioned before, the collinearity in the explanatory variables is not an issue in this study since it doesn't have a large effect on the coverage and size of the prediction intervals. Even though the coverage was very good, the average prediction window was 35.29. This means most of the predictions for O3 at a given location would have been plus or minus 35.29 away from the point estimate.

6 Results

The relationship between O3 level and CMAQ is strong, linear, and positive. The RMSE when using only the closest CMAQ measurement location to predict O3 is 8.38. However, when we use the 17 nearest locations, as suggested by forward selection, the RMSE is 7.83. Even though CMAQ gives us a good idea of what the O3 level is, it is better to use multiple CMAQ measurements to predict each O3 level. We were supplied with a list of locations for which we predicted O3 levels. We calculated the CMAQ measurements of the 17 locations nearest to the locations provided. The CMAQ values and model coefficients allowed us to predict the O3 levels. The graph below shows the predicted O3 levels.



Notice that the O3 levels are highest around Louisiana and parts of Texas. Also, the northeast region of the United states has low levels of O3.

7 Conclusion

CMAQ is a useful tool to help approximate the levels of O3 in a region. However, using multiple CMAQ measurements from neighboring locations provides more accurate predictions. However, the prediction intervals are rather wide compared to the magnitude of the CMAQ values. Though the bias of the predictions is small, the actual O3 values for a specific location may vary significantly. Observing broad geographical areas can be useful in identifying problem areas. Locations such as Louisiana and Texas should consult with policy makers in the northeast to find out if any changes could be made to reduce pollution.

References

1. R: A Language and Environment for Statistical Computing.R Core Team. R Foundation for Statistical Computing.Vienna, Austria. (2017).url = https://www.R-project.org/