Predicting U.S. Ozone Levels Using CMAQ and EPA Station Data

Stephen Merrill

March 10, 2016

Introduction

The increasing quantity of ground-level ozone (O3) is a major concern to environmentalists. O3 is the main component of smog and breathing high concentration can cause medical conditions such as asthma. Because of this, ozone levels are monitored by the EPA. In this study, they have provided 800 station measurements of maximum ozone in an eight hour period on May 22, 2005. Another attempt to monitor ozone is the Community Multi Scale Air-Quality Model (CMAQ), which uses an algorithm to project ozone levels. However, these projections are not in line with the actual observed ozone levels from the EPA. There is some relationship there, but neither method provides accurate coverage of the full range of locations scientists are interested in. They would like to know the relationship between CMAQ and the station observations and be able to predict ground-level 03 at more locations. To this end, we will use both sets of data in order to build a spatial model of the US that satisfies these research goals.

Data Exploration

Figure 1 below shows the difference between the two ozone measurements. Although CMAQ predictions cover more area, they are inaccurate in that they do not accurately represent the observed station levels. The analysis is complicated by the clear non-linear relationship between the ozone levels and location on the map. This relationship between ozone and latitude and longitude can be seen further in Figure 2. The amount of data we have to use further complicates the analysis. There are only 800 EPA station observations, but the CMAQ computers have made predictions at 66960 different locations. Finally, because this is spatial data, it is of note that the data is not independent. The ozone level at one location depends on the levels elsewhere.

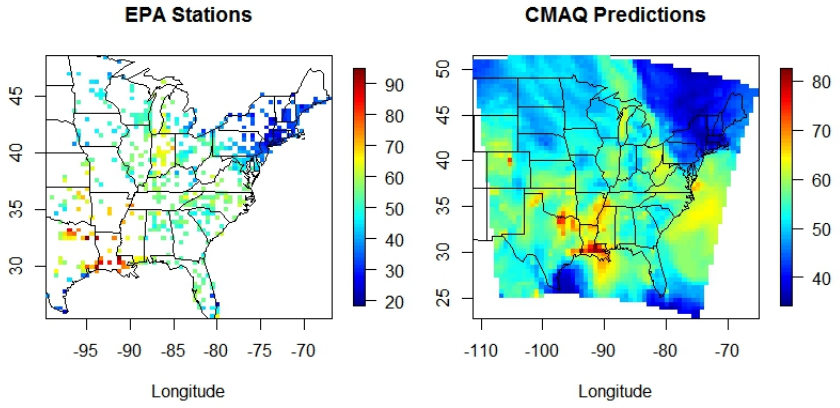

Figure 1: We would like to combine the accuracy of the stations with the spatial coverage of the CMAQ projections.

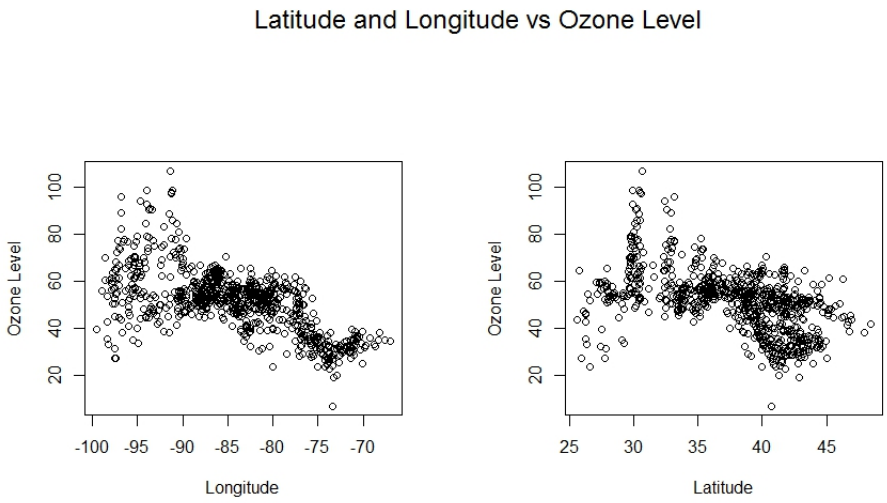

Figure 2: Latitude and longitude have a non-linear relationship with Ozone level.

Methods

Spatial statistics is a viable method for analyzing data with linear and/or non-linear relationships with observations that are correlated over space. In our analysis, using a spatial model allows for a combination of the non-linear latitude and longitude relationships and linear CMAQ distance relationships. The process can be modeled as follows:

If we observe $y(s_1), \ldots, y(s_N)$ and the covariates $x(s_1), \ldots, x(s_N)$ at N distinct spatial locations $s_1, \ldots, s_N$ in some spatial region D then

$$\boldsymbol{Y} = \begin{bmatrix} y(s_1) \\ \vdots \\ y(s_N) \end{bmatrix} \sim N(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{R} + \tau^2 \boldsymbol{I}_N) \tag{1}$$

Where $y(s_i)$ denotes the Ozone level at the for the i^{th} spatial location.
s_i denotes the i^{th} spatial location.
$\boldsymbol{Y}$ represents the distribution of the Ozone levels over the spatial grid.
$\boldsymbol{X}$ is the collection of linear covariates.
$\boldsymbol{\beta}$ is the effect each linear covariate has on the response.
τ^2 is defined below with the Gaussian Process.
σ^2 is defined below with the Gaussian Process.
$\boldsymbol{R}$ is defined below with the Gaussian Process. Each element of $\boldsymbol{R}$ is computed from a Mattern function with respect to the distance between spatial locations. The effect of the non-linear covariates are accoounted for here.

A spatial statistics model relies heavily on Gaussian process regression. Gaussian process regression is a viable method for analyzing data that is non-linear with correlated observations. A Gaussian process is a type of Stochastic process. A Stochastic process is a collection of random variables over a specified interval, where only a finite collection of random variables are observed. In this case, we observe 800 observations of Ozone levels which correspond to latitude and longitude. A Gaussian process is a stochastic process where any finite collection of random variables follow a multivariate normal (MVN) distribution. When we observe Ozone levels at various locations, we need to find a function such that the residuals are small and

the function is smooth. If we let the function at each location be random, then we can model the function with a distribution. Since there are multiple non-linear covariates, this is best notated as follows:

$$\boldsymbol{Y} = \begin{bmatrix} y(x_1) \\ \vdots \\ y(x_N) \end{bmatrix} \sim N(\mu \boldsymbol{I}_N, \sigma^2 \boldsymbol{R} + \tau^2 \boldsymbol{I}_N) \tag{2}$$

μ is the overall mean.
τ^2 is the variance of the distribution of Ozone given the random variables that describe the non-linear relationship between Ozone and latitude and longitude.
σ^2 is the variance of the collection of those random variables.

The Gaussian Process also needs to identify a covariance structure, R. We use a Mattern function, which satisfies the requirements of being positive definite and being strongly correlated over small distances. To this end we use a Mattern function with parameters α and ν which govern decay and smoothness, respectably. Decay can be considered the correlation between points. As α increases, the correlation decreases, and the function tends to jump more. Smoothness determines how smoothly the function moves from value to value. As ν increases, the function becomes more smooth.

We also need to incorporate the relationship between CMAQ location and Ozone level into the model. However, we must first solve the problem of high dimensionality present in the data. We will define the CMAQ covariates as a list, ordered by distance to the associated station observation. There are 66960 CMAQ covariates and only 800 station predictions. This creates problems with overfitting and potentially overreacting to small bumps in the data and creating false positive relationships with variables that aren't related. These issues are referred to as the "Curse of Dimensionality", and must be addressed.
We use a principal component regression algorithm in order to solve this problem. This technique represents the 66960 potential variables with far fewer covariates by reducing the dimension to fit the direction(s) of the data in which the observations vary the most. These directions are referred to as principle components. This method is modelled as follows:

$$Y(x_i) = \beta_0 + \sum_{i=1}^{N} \beta_i x_{ij} + \epsilon_j$$

Is transformed to:

$$Y(x_i) = \beta_0 + \sum_{i=1}^{N} z_{ij} \theta_i + \epsilon_j$$

$\epsilon_j \overset{iid}{\sim} N(0, \sigma^2)$
x_{ij} is the ordered CMAQ value for location j
β is the effect size of the associated x
z_{ij} is the ordered principle component value for location j
θ is the loading for the associated z

Now that the covariates are all defined, the above Spatial Statistics model can be used. This model will fulfill both goals of the analysis. First, it will allow for inference to be made on the β effect between the CMAQ covariate(s) on observed Ozone levels. Second, it will predict Ozone levels based on CMAQ and latitude and longitude. The EPA is interested in prediction at 2834 specified locations. This can be done by using the properties of the MVN to first obtain the conditional distribution of $\boldsymbol{Y^*} \mid \boldsymbol{Y}$. By the process described above, we know the mean and variance of the conditional distribution. In other words, we can find the mean and variance of the predicted values, given the observed Ozone levels. Additionally, we can obtain prediction intervals by taking the corresponding quantiles of the conditional distribution using the empirical rule.

The Spatial Statistics model does not carry the same assumptions as linear regression. It is not restricted to model only linear relationships, so linearity is only an assumption if that covariate is being included in the $\boldsymbol{X}$ matrix, as defined above. The method also predicts relationships when observations are not independent. This indicates that dependence will be present and the variance will be relative to the amount of data. Therefore, the only assumption that is made with this model is that the residuals are normally distributed.

Model Justification

The Spatial Statistics model allows us to model both the non-linear relationship between Ozone level and latitude and longitude and the linear relationship between Ozone level and CMAQ distance. Additionally, this method allows for us to model data where various Ozone levels are dependent on each other. To select the covariates, the PCR process described above was followed. First, distances between the EPA stations and CMAQ locations were calculated and ordered. Then a PCR model was fit to the closest 20 locations. Figure 3 shows that the lowest RMSE was obtained by the model with a single component and that the corresponding covariate has a linear relationship with the actual Ozone levels, fulfilling the previously outlined assumption.

As discussed previously, the model assumes that the residual values are normally distributed. In other words, the observed Ozone levels minus the predicted Ozone levels are normally distributed. However, this assumption is not easily verified as we only have one draw from the multivariate normal. Therefore, it is not possible to verify that a single draw comes from the MVN. We will move forward assuming that our data comes from a MVN distribution with the knowledge that the distribution is robust to misclassification. We will be aware that a violation on the normality assumption will increase the width of the prediction intervals. Additionally, our use of the normality assumption for constructing the prediction intervals would be incorrect.

The model also uses a decay and smoothness parameter. The decay parameter, α, is chosen using maximum likelihood estimation. However, the smoothing parameter, ν, is not easily chosen. It is usually selected by visually examining different values; however that is not easily done in this setting, so a default value of $\nu = 1.5$ is used.

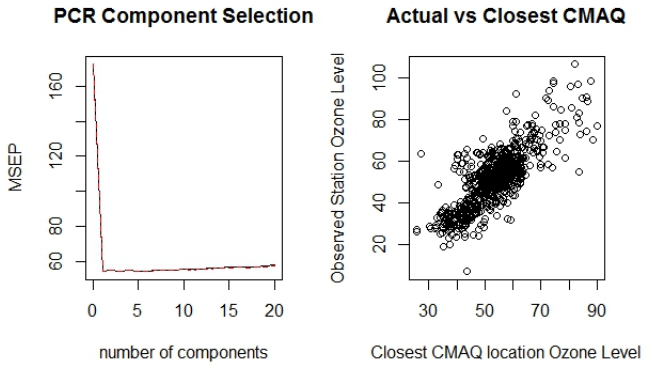

Figure 3: PCR indicates the inclusion of one additional linear variable.

The model yields an R^2 value of .8761. This can be interpreted as 87.61% of the variability in Ozone levels can be explained by the variability in all of the covariates. Additionally, there were no problems with colinearity or the fact that Ozone levels are always positive.

Performance Evaluation

This study is solely concerned with inference and prediction. We make inference on the effect of the CMAQ values on the actual Ozone levels by reporting the β values and 95% confidence intervals for Intercept and effect of the closest CMAQ value in Table 1. The β_0 for Intercept is interpreted to mean that if the closest CMAQ location value was zero, the actual Ozone level would be 11.506. The β_1 for closest CMAQ location is interpreted to mean that if the closest CMAQ location value increased by one Dobson Unit, the actual Ozone level would increase by 0.6724 Dobson Units. The confidence intervals can be interpreted in the usual fashion. If the process to create them was repeated iteratively, 95% of the intervals would contain the true β values.

Table 1: Beta Estimates

	Estimate	SE	95% CI
β_0	11.506	0.00373	(11.499, 11.513)
β_1	0.6724	7.0061e-5	(0.67229, 0.67257)

We also evaluate our model based on prediction performance. We used a Cross Validation method to simulate predictions and obtain results (see Table 2).

Table 2: Prediction Diagnostics

	Estimate
Bias	0.0109
RMSE	5.2135
Coverage	0.956
Interval Width	20.808

Of particular note in Table 2 are bias and coverage. The low bias value tells us that the overall prediction nature of our model is remarkably unbiased. This is due to a property of the Multivariate Normal Distribution, which we assumed fit our data. The coverage statistic also supports the validity of our model in that it is so close to .95, the theoretical value with our assumed error rate. This means that the correct number of actual points were contained within the intervals of error for our predicted Ozone levels.

Results

Our results include estimates of the covariance parameters, presented below in Table 3.

Table 3: Parameter Estimates

σ^2	τ^2	α	ν
66.742	23.585	0.579	1.5

These parameters are defined as: σ^2 and τ^2 are variances in Multivariate Normal Distributions that make up the model. α is the decay parameter, which governs the correlation of the regression line and ν is the smoothness parameter, which governs the smoothness of the regression line. ν cannot be estimated and is assigned based on visual inspection of different values. Both are parameters in the Mattern covariance function.

The model addresses the motivation for the study by making accurate prediction of Ozone levels in the U.S. and by making inference on the relationship between those actual levels and the projected values from the

CMAQ algorithm. The predictive accuracy was validated through a simulation. The final results, which can be seen in the following figures, display renderings of predicted Ozone levels at the desired locations that are in line with the EPA station observations.

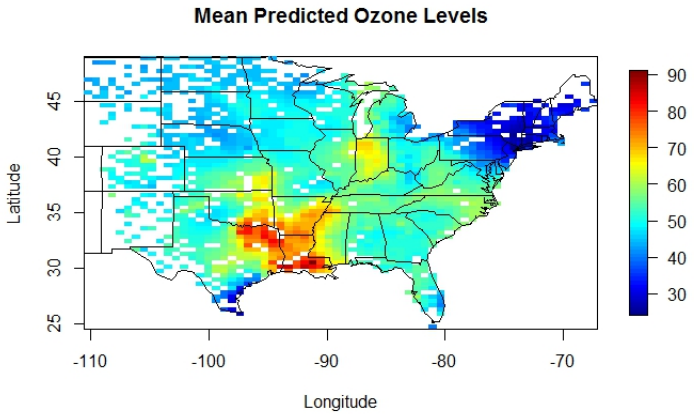

Figure 4: Predictions at 2834 locations across the U.S.

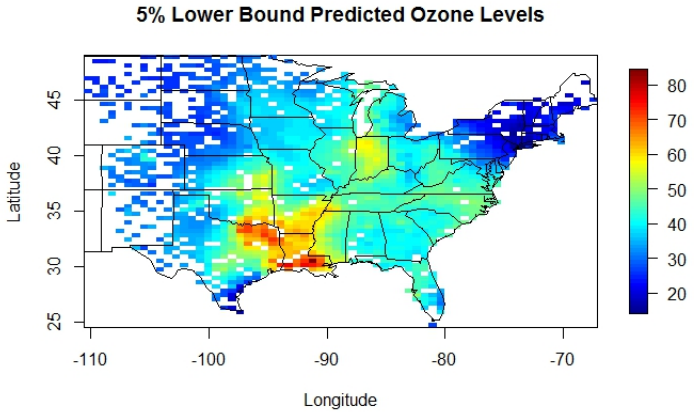

Figure 5: Predictions at 2834 locations across the U.S.

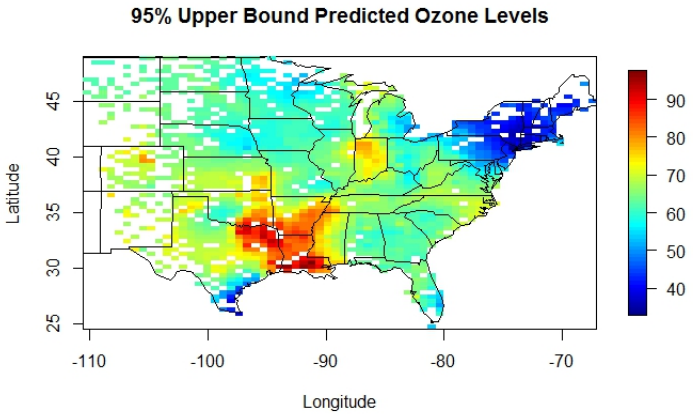

Figure 6: Predictions at 2834 locations across the U.S.

Conclusions

In order for the EPA to accurately understand ground-based Ozone levels, it is of interest to be able to predict the Ozone levels at locations without monitoring stations and understand the relationship between CMAQ projections and actual measurements. Our proposed Spatial Statistics model meets this goal by deriving effect sizes for the CMAQ relationship and then making unbiased predictions on the overall spatial relationship. Analysis of the prediction results shows a very effective approach to modeling the desired relationship.

One shortcoming of the study is the lack of a time element in the data. In order for the results to be more useful, the EPA will want to be able to project into the future. In the future, if more data were gathered over time the results would potentially be more interesting.