Exam 1: Spatial Data Analysis on Ozone Levels in the Eastern United States

Matt Oehler

March 19, 2018

Abstract

Many things such as industrial processes, automobiles, oil refineries, gas stations, and manufacturing, produce atmospheric byproducts. Under certain conditions these atmospheric by products can combine to form ground level ozone, which is a principal component in smog. Breathing air with high amounts of ozone can cause people to experience chest pain, emphysema, asthma, bronchitis, or other health issues. In this study, we will use correlated data modeling methods to assess the relationship between the Community Multi-scale Air Quality Model measurements for ozone, and measurements of ozone that were collected from weather stations across the eastern United States.

1 Introduction

Ground level ozone formation is a phenomenon that occurs as a result of many man-made industrial processes and facilities. These processes/facilities include operating automobiles, oil refineries, gas stations, and manufacturing. The atmospheric byproducts of these processes vary, but include types of nitrogen oxides and other volatile organic compounds, which under certain weather or atmospheric conditions 'bake' together and form ground level ozone, which is a principal component of smog. Breathing in high concentrations of ground level ozone can cause people to experience health issues such as asthma, bronchitis, emphysema, chest pain, etc.

In this study we explore how to model the concentration of ground level ozone in areas across the eastern United States, based on measurements from the Community Multi-scale Air Quality Model (CMAQ). CMAQ mathematically simulates ozone formation based on characteristics of the ground, temperature, urban density, etc. CMAQ is a great tool, but it still does not exactly reproduce ozone measurements in areas as recorded at weather stations. By modeling the relationship between CMAQ simulated values and ozone measurements at 800 weather stations across the United States, we will be able to a) understand the relationship between CMAQ model simulated values and actual measurements of ground level ozone and b) predict the ground level ozone concentration in areas where there is not a weather station. Groups such as the EPA would be able to better moderate the levels of ground level ozone that are produced in the United States.

2 Data

For this study, we will used 66,960 simulated CMAQ values for regions across the eastern United States. The quilt-plot in figure 1 displays these data clearly. We also used ground level ozone measurements from 800 different weather stations across the same region, and these measurements can be seen in figure 2. Since we don't have as many station measurements as we do CMAQ values (do to the limited number of weather stations) we merged the data sets together by pairing station measurements with the CMAQ value that was closest in terms of geographical distance (based on latitude and longitude coordinates). This results in data set of CMAQ values and station measurements for 800 distinct locations. Later on, when attempting to predict ground level ozone at various locations we will similarly match CMAQ values with the list of latitude and longitude coordinates for the locations at which we wish to predict ground level ozone as it would be recorded by a weather station.

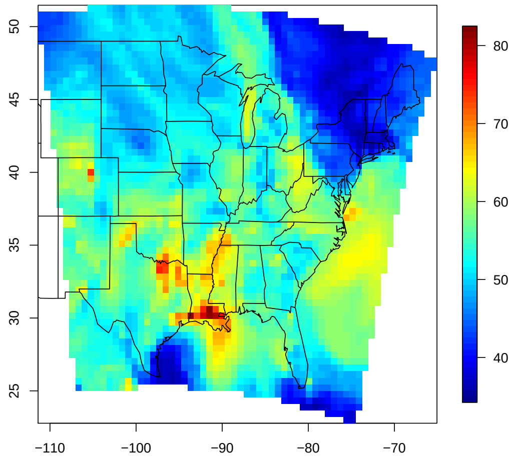


Figure 1: Quilt-plot of ozone levels based on CMAQ model

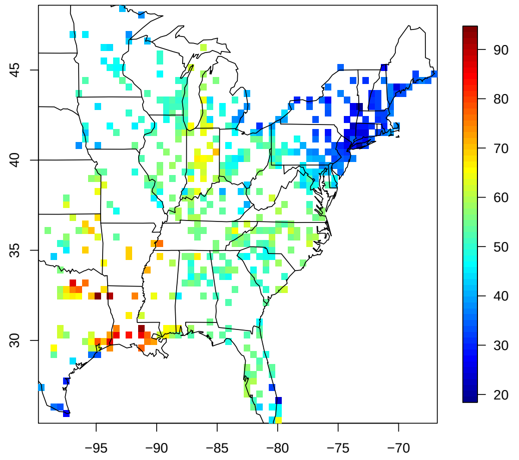


Figure 2: Quilt-plot of ozone levels as recorded at weather stations

3 Method

We can see from figures 1 and 2 that simulated CMAQ values and ozone measurements are correlated in space, meaning that observations that are a short geographical distance apart are closely correlated. As a result, we can't use simple linear regression methods because then we would not be able to properly quantify the uncertainty associated with the derived model coefficients. To properly account for the correlation that is inherent in the data we will use Gaussian process regression.

A Gaussian process is a finite collection of random variables that follow a multivariate normal distribution, as shown below in equation 1. Additionally, each of the terms in the model are explicitly defined below. This model allows us to choose a structure for the covariance that we believe will properly account for the correlation in the data.

$$\mathbf{Y} \sim \mathbf{N}\big(\mathbf{X}\boldsymbol{\beta}, \sigma^2((1-\omega)\mathbf{R} + \omega\mathbf{I})\big) \tag{1}$$

$$\begin{aligned}
\mathbf{Y} &= \text{response vector} \quad (nx1)\\
\mathbf{X} &= \text{model matrix} \quad (nxp)\\
\boldsymbol{\beta} &= \text{model coefficients} \quad (px1)\\
\sigma^2 &= \text{variance term} \quad (scalar)\\
\mathbf{R} &= \text{structure of correlation matrix} \quad (nxn)\\
\omega &= \text{nugget term that allows for sampling variability} \quad (scalar)
\end{aligned}$$

When doing Gaussian process regression, there are various covariance structures that can be used. We chose to use an exponential covariance structure for our model, which is shown in equation 2. Using an exponential covariance structure seems practical in this case because the observations in the data set are not equidistant from each other. The exponential structure allows for non-equal spacing when accounting for correlation between points. Equation 2 shows how we account for the correlation between the residuals of two points at locations s_1 and s_2, where $||s_1 - s_2||$ is the Euclidean distance between the two points and ϕ is the range parameter that controls for how quickly correlation decreases with distance. We chose to include a nugget term ω in our model as well because it seems plausible that there could be sampling variability (different ozone measurements at the same location) in this scenario.

$$\text{Corr}(\epsilon_{s_1}, \epsilon_{s_2}) = \exp\left\{ -\frac{||s_1 - s_2||}{\phi} \right\} \tag{2}$$

It is also important to note that when performing Gaussian process regression, there are a few assumptions that need to hold. We assume that the data are multivariate-normally distributed and that there is constant variance. Since we are accounting for the correlation in the data, we do not assume independence, but rather dependence. These assumptions will be further addressed and verified in following sections.

4 Results

4.1 Model Fit

We used the model in equation 3 to get the results contained in this section. This model is simply a model of the station measurement, Y, based on the CMAQ measurement while still accounting for the correlation in the residual term based on the latitude and longitude coordinates of each observation using the structure described in the previous section. The terms β_0 and β_1 correspond to the intercept and slope terms respectively.

$$Y = \beta_0 + \beta_1 \times (CMAQ) + \epsilon \tag{3}$$

After fitting the model, we resulted in coefficient estimates and 95% confidence intervals, which are displayed in table 1. At first glance we can see that the intervals are relatively wide, which likely implies that this model doesn't fit the data very well,but we will use cross-validation later to get a better assessment of model fit. We estimated the intercept term to be 27.39, which if interpreted would be the average weather station ozone measurement for an area with a simulated CMAQ value of 0. The CMAQ coefficient, β_1, is 0.37 meaning that on average each increase of 1 for a simulated CMAQ value, a weather station at that same area would increase by 0.37.

	2.5%	Estimate	97.5%
Intercept	19.02	27.39	35.76
CMAQ	0.28	0.37	0.47
Range (ϕ)	2.59	5.45	11.47
Nugget (ω)	0.09	0.17	0.30

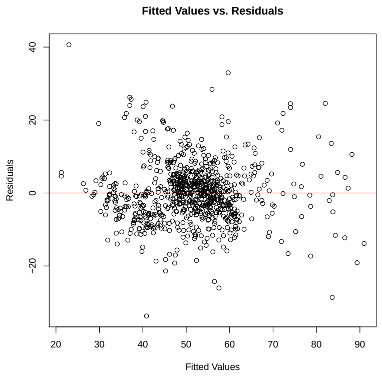
Table 1: Model coefficients and confidence intervals

Now that we have fit a model, we must also verify that the assumptions of Gaussian process hold. This is difficult to do with the correlated model structure. However, using the lower Cholesky decomposition of the covariance matrix, $\sigma^2((1-\omega)\mathbf{R} + \omega\mathbf{I})$, we can decorrelate the model and then use standard linear regression to verify the assumptions. Multiplying the response vector of station measurements by the inverse of the lower Cholesky decomposition results in values that are independent of each other (essentially we divided out the correlation). We then model these independent measurements based on the CMAQ values (which are also multiplied by the inverse of the lower Cholesky decomposition) and assess the assumptions using the results of the decorrelated model which are shown in figures 3 and 4. The plot of residuals shown in figure 3 shows that the variance is approximately constant throughout the data. The aforementioned normality assumption is confirmed as well since figure 4 shows that the residuals are approximately normally distributed. We also used this decorrelated model to calculate an R^2 of 0.623, meaning that about 62% of the variation in station measurements is explained away by the CMAQ values. It's good that we can explain more than half of the variability of station measurements with our model, but there is still a lot of room for improvement.

4.2 Performance Evaluation

Next we performed a cross-validation to see how well the model performed in terms of predicting the ground level ozone concentration given a CMAQ value. To do this we randomly removed a portion of the data, fit the model, and then used the resulting model to predict the removed portion of the data, and compared those predictions with the true, observed values. This process was repeated for 200 iterations, and the results are compiled together in table 2

From the table we can see that the model predicts with an average bias of -3.98 meaning that on average, the predictions are lower than what they should be by 3.98. The calculated RPMSE



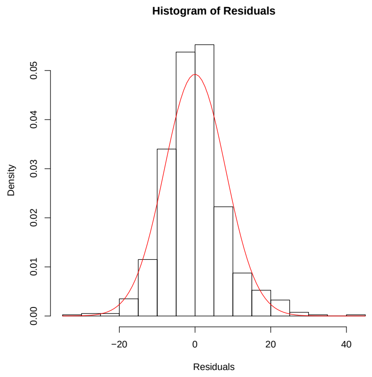


Figure 3: Plot of residuals vs fitted values

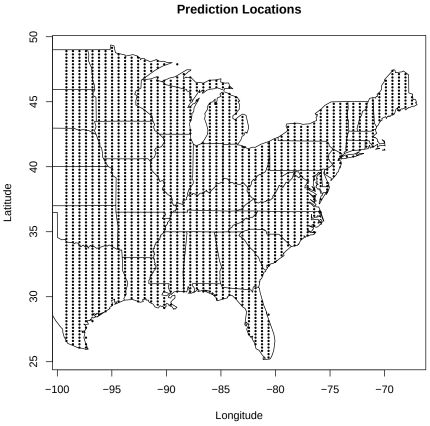
Figure 4: Histogram of residuals

	Prediction Performance
Bias	-3.98
MSE	127.55
RPMSE	11.26

Table 2: Performance metrics for prediction

tells us that on average our predictions are off by 11.26. Given that the scale for ground level ozone has a range of about 20 to 100, this means that our model predicts fairly well, but again there is certainly room for improvement.

Since we are satisfied with our models performance, we use it to predict the values of several more (about 2500) locations across the eastern United States. The locations for which we wish to predict are shown in figure 5 and the quilt plot of those locations with their corresponding predicted value for ground level ozone is shown in figure 6.



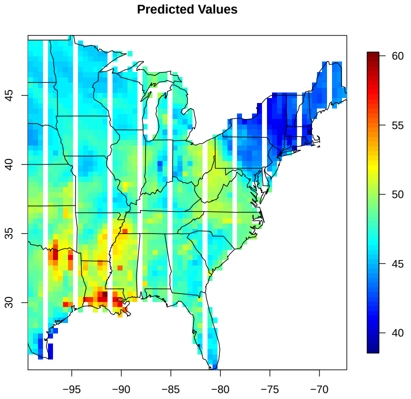


Figure 5: Plot of locations where we will predict ozone concentration

Figure 6: Quilt-plot of predicted ground level ozone

5 Conclusion

In conclusion we were able to accomplish both of the primary goals of this analysis: We were able to find coefficient estimates that describe the relationship between CMAQ values and weather station measurements for ground level ozone, and we showed that we can use these coefficient estimates to predict the ground level ozone concentration with reasonable accuracy. In this study we used a simple model with the Gaussian process framework. This analysis could be improved by trying and comparing multiple plausible covariance structures, or possibly by using various types of basis function expansions to see if there is a practical model that fits the data better and predicts more accurately.