Ozone Measurements: A Spatial Analysis

Jackson Curtis

March 19, 2018

Abstract

This paper analyzes the effectiveness of the Community Multi-Scale Air Quality Model (CMAQ) at predicting max Ozone levels over areas of the United States. I use a Gaussian process model with spatial correlation to analyze instrument measured and CMAQ-predicted ozone levels.

1 Introduction

Ground level ozone is created when pollutants from various human activities (factories, automobiles, etc) undergo a chemical reaction initiated by sunlight. Ozone is the main component of smog and can contribute to respiratory problems such as asthma, bronchitis, and emphysema.

Because of these problems, much research has gone into forecasting ozone levels based on climate patterns, geography, and human activity. The Community Multi-Scale Air Quality Model (CMAQ) is one tool that scientists have built that predicts ozone levels at a particular location based on inputs such as elevation, weather, human activity and more. Figure 1 shows these predictions over the eastern United States on May 22nd, 2005.

We can see the limitations of CMAQ however, by comparing it to stations that measure the ozone directly. Figure 3 shows the actual values obtained on May 22nd from 800 stations scattered around the country. Figure 4 shows the difference between the observed and modeled ozone levels. Clearly, CMAQ isn't perfect. Not only can it be off by as much as 50% in some locations, but high and low estimates cluster by location, with the Midwest predictions being too low and the Northeast being too high. Figure 2 shows that in general for this day in May, the majority of the predictions were too high (by an average of 2 units).

This leaves researchers who want to know the precise ozone measurements at all possible locations with a problem. Clearly, sticking in the CMAQ value will not account for the spatial variability that we observe from the stations. However, attempting to use the station measurements for all locations might have serious problems. For example, if you tried to get the ozone level by averaging two nearby stations, it might be very wrong if a mountain range separates the stations. CMAQ incorporates information about geography, so a way of combining the information from the stations and CMAQ would help researchers.

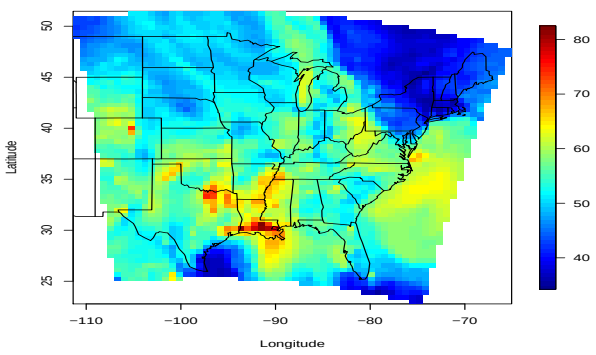

Figure 1: CMAQ Ozone predictions for the eastern U.S.

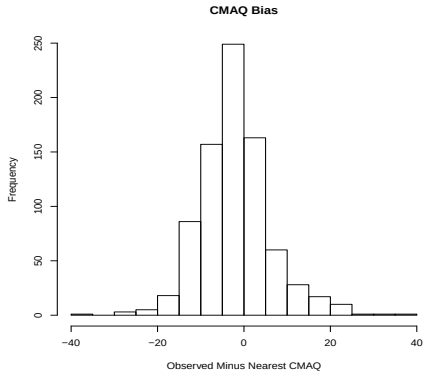

Figure 2: More locations had predictions that were too high

Our goal will be to build a model that lets us understand the relationship between CMAQ and actual ozone levels and predict the values of ozone at unmeasured locations by using what we know about nearby measurements and CMAQ values.

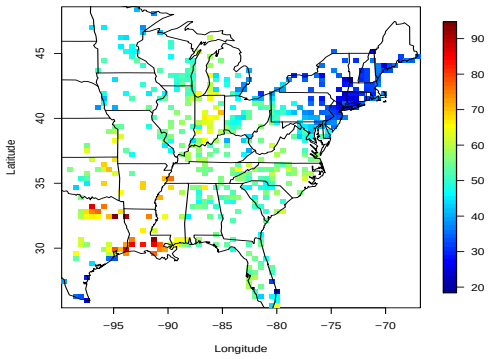

Figure 3: Actual measurements taken at 800 stations

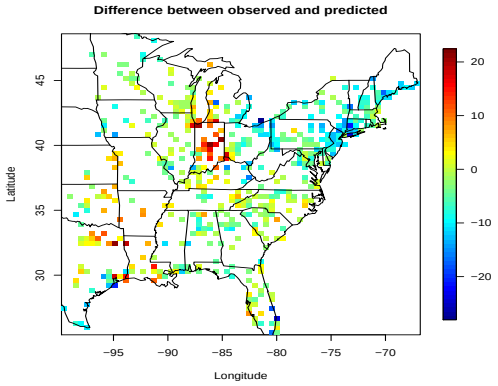

Figure 4: Differences between ozone measurements and CMAQ predictions

2 Choosing a Model

For our model we would like to use observed station values as our response variable and CMAQ variables as predictor variables. An important feature of this model is that none of our measurements are independent of each other. Because each measurement happens in a geographic area, we expect areas close to one another to have similar measurements.

To deal with this, our model will be built using a Gaussian process. A Gaussian process assumes that all possible values (observed and unobserved) come from a multivariate normal distribution where the covariances of that distribution do not have to be zero. This model can be written as:

$$\boldsymbol{Y} \sim \boldsymbol{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{1}$$

where Y is a random vector of ozone measurements, μ is a vector of means for each Y_i and Σ is the variance-covariance matrix for Y.

By specifying a mean function and a correlation function, we can estimate the parameters in the mean and correlation functions and use that model to do inference and make predictions. For the mean function we can use $\mu = X\beta$ where X is our design matrix and β is our vector of fixed effects. Intuitively, for our correlation function we want something that will produce high correlation when values are near each other and low correlation when they are far apart. You can see why this would be desirable when looking at Figure 3, where observations in the same state are almost all the same, but start to differ as distance increases. Consider the correlation function:

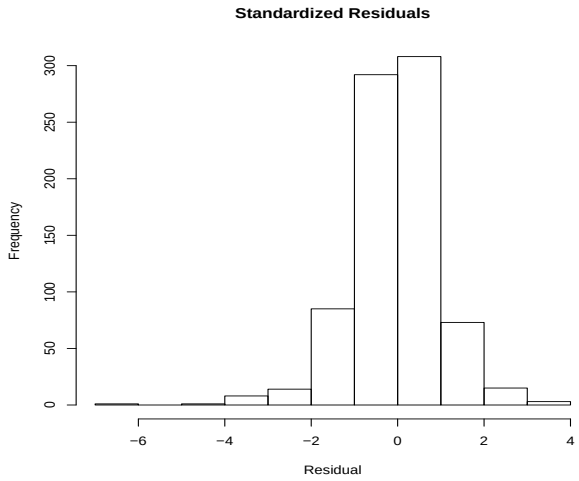

Figure 5: Our standardized residuals show large outliers

$$Corr(\epsilon_{l_1}, \epsilon_{l_2}) = exp\left\{\frac{-|dist(l_1, l_2)|}{\phi}\right\} \tag{2}$$

where ϵ_{l_1} represents the error at location one, and the dist function gives Euclidean distance between locations l_1 and l_2. Clearly, the correlation decreases as distance increases, but how much it decreases depends on our range parameters ϕ. Using this correlation function, we can calculate Σ as follows:

$$\Sigma = \sigma^2((1-\omega)R + \omega I) \tag{3}$$

where R is the matrix of all pairwise evaluations of our correlation function, σ^2 is our residual variance, and ω is a nugget. The nugget is another parameter that will be estimated from the data that accounts for sampling variability of samples in the same location. We will find β, σ^2, ϕ, and ω using maximum likelihood and we can get predictions on unseen locations by calculating the conditional means and variances from their X-matrix and location information.

This model is great for spatial data like ours because it allows us to condition a prediction on all the known points near the prediction area. This model assumes all points equidistant from each other have equal variance, which seems like a safe prediction for land based measurements, but oceans and mountain ranges could cause concern. Additionally, we will use Euclidean distance which is a good approximation over short distances, but not completely accurate because of the curvature of the earth.

3 Building a Model

We have many options on how to build predictors for our model. We could use only the CMAQ value closest to the observed value as our predictor, or we could use the closest 5, 10, or 100. Choosing more than one will cause extreme collinearity because points next to each other have almost identical CMAQ predictions. To adjust, we can use principal component analysis to orthogonalize our predictors. I explored using the 100 closest points to my observed value. I calculate the principal components of these 100 points and then built models using the first 1, 2, 5, 10, and 20 components. I used those models to get out of sample predictions and found MSE was minimized between 5 and 10, so I used eight principal components in my final analysis.

We can validate some of our assumptions by looking at a transformation of our model based on the Cholesky decomposition. If L is the lower-triangle Cholesky decomposition:

$$L^{-1}Y \sim N(L^{-1}X\beta, \sigma^2 I) \tag{4}$$

Because this is an uncorrelated linear model, we can assess our assumptions. Figure 6 shows that our linearity assumption is met, and an examination of fitted values vs. residuals shows no

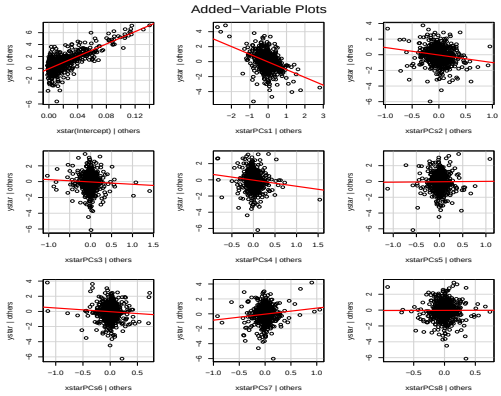

Figure 6: There are no apparent deviations of linearity

R^2	0.693
Out of Sample RMSE	4.919
Prediction Interval Coverage*	0.9875
Mean Interval Width	33.565

Table 1: Fit statistics for our model

heteroskedasticity problems. However, as shown in Figure 5, there are clearly several outliers that are concerning. Exploring these residuals, both extreme outliers on the left are places surrounded on all sides by large bodies of water (Long Beach and the Florida coast). We will proceed noting that the water may be causing issues with prediction, and predictions in areas with large bodies of water should be suspect. Also, because we have more extreme values than we would hope, it is doubtful that our intervals will have our desired α level.

Table 1 summarizes the fit of our model. From the transformed model we can show that we are explaining about 69% of the variance in our model. The other statistics come from an 8-fold cross validation where 100 observations were predicted from a model built on the other 700. Our out of sample RMSE is just under 5, which is a big improvement over just using the CMAQ scores (which produces a RMSE of 8.9). Almost 99% of our prediction intervals contained the true value, but it should be noted that because a Gaussian Process treats our entire sample as a single observation from a mulitvariate normal, measuring prediction interval coverage over a single day of observations doesn't hold because of lack of independence among the sampled points, so it is not surprising to see such a high level. A prediction region over an entire sample would need to be specified (and then many samples examined to verify) to get a specific α level. Our prediction intervals had a mean width of 33.6, meaning we were usually within 16 to either side of our predicted ozone level.

4 Results

	$\hat{\beta}$	Confidence Interval
1	0.34	(0.223, 0.458)
2	0.42	(0.32, 0.519)
3	0.32	(0.224, 0.408)
4	0.35	(0.253, 0.443)
5	0.18	(0.065, 0.291)
6	0.32	(0.22, 0.413)
7	0.30	(0.212, 0.393)
8	0.32	(0.24, 0.391)
9	0.39	(0.279, 0.504)
10	0.37	(0.284, 0.461)
$\sigma = 7.13$	$\phi = 2.14$	$\omega = 0.30$

Table 2: Back-transformed coefficient estimates

Table 2 shows our range, nugget, and variance estimates, as well as our first ten betas. Our beta estimates have a complex interpretation. The first beta describes the expected change in ozone as the CMAQ prediction nearest to the ozone measurement location increases by one standard deviation. Similarly, the second beta describes the change for the second nearest CMAQ location. One difficulty in using this interpretation is that it assumes that change will happen if all else is held constant. Nearby locations are so extremely correlated that they almost never change independently.

Table 2 only reports the first 10 of the 100 beta estimates. 28 of the first 30 coefficients are significant, and then the proportion drops off. Also, the vast majority are positive, which is a positive sign that CMAQ is doing what we hoped: incorporating and adjusting for information that our model didn't have.

Unfortunately the model doesn't provide simple interpretations of the coefficients, but one way we can examine the relationship between CMAQ and reality is to compare a model fit using only an intercept and spatial correlation to our full model. By every measure CMAQ improves the fit of our model. A likelihood ratio test produces a p-value less than 0.0001 and the residual variance drops from 160 to 50.

Using our model and what we know about the correlation with our observed variables, we can create predictions at any location in the range of our data. Figure 7 shows our predictions, and 8 shows the uncertainty at each of the predictions. Comparing Figures 3 and 7 show that our predictions recreate our observed values well. Figure 8 shows that the East has high certainty and then the West and the coastal states have higher uncertainty.

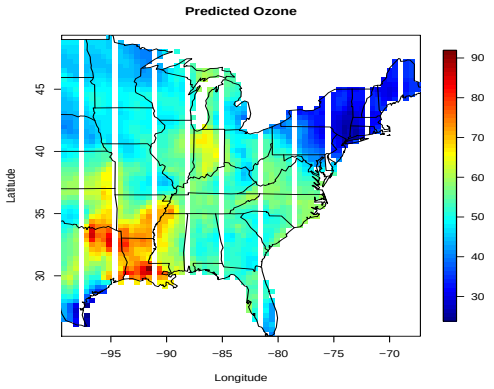

Figure 7: Predictions for 2,685 locations

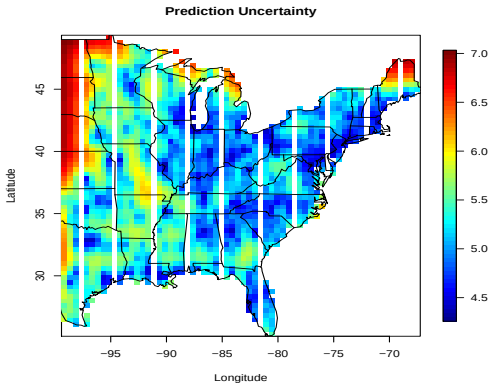

Figure 8: Standard errors of predictions

5 Conclusion

This analysis provided a clean way to merge CMAQ predictions and observed values to improve interpolation of values where direct measurements were taken. The analysis showed definitively that this model is better than either using the CMAQ values alone or interpolating based only off observed values.

It's important to remember the conclusions that can be drawn about the CMAQ model. Although we showed that the CMAQ was consistently too high for this single day, because we are modeling using the Gaussian process (and the non-independent, multivariate normal distribution), we should not assume that it would be consistently biased over time. That conclusion would require analysis of many days of data.

As noted before, our coefficients don't provide the neatest interpretation of the CMAQ values. In addition, an analysis of our residuals showed that we are probably not adjusting for the presence of water correctly, and future work should look at building water locations into the model. Future models could also add a temporal dimension and include much more data over time. This would provide a better understanding of how CMAQ's overall bias.