STAT 536 Midterm: Calibrating the Community Scale Air Quality Model for use in Ozone Predictions

Kaylee Hodgson

Brigham Young University

Abstract. The Environmental Protection Agency (EPA) monitors closely the ozone levels in the United States, due to the rising levels of O3 (ozone) in the atmosphere and the subsequent health costs. Currently the EPA only obtains scattered measurements across the United States, and the main prediction tool used for all areas is the Community Multi-scale Air Quality Model (CMAQ), a mathematical model that predicts ozone levels based on area characteristics. However, scientists are aware that the predictions from the CMAQ model often are not accurate to the station measurements. In this analysis, I employ spatial regression analysis in an attempt to identify the relationship between the CMAQ predictions and the EPA station measurements of ozone levels in the mid and eastern Unted States. I additionally use the model, utilizing Guassian process regression, to predict the ozone levels for other locations. The purpose of this analysis is to identify the relationship between CMAQ and EPA measurements in an effort to find a better way to explain and predict ozone levels in the eastern United States, than the current method used.

1 Introduction and Motivation

Ground ozone levels have increasingly become a policy concern in the United States. Ground-level ozone is the main component of smog, mainly caused by pollution from industrial activities and vehicle emissions, that "bakes" in the sunlight and forms the ozone layer. Breathing in high concentations of O3 (ozone) causes many health issues, including chest pain, bronchitis, emphysema, and asthma. The Environmental Protection Agency (EPA) monitors the ozone in the United State by setting up stations to measure the ozone levels across the country. Because there is a high cost incurred by collecting these measurements, the EPA cannot measure the level at every location in the United States.

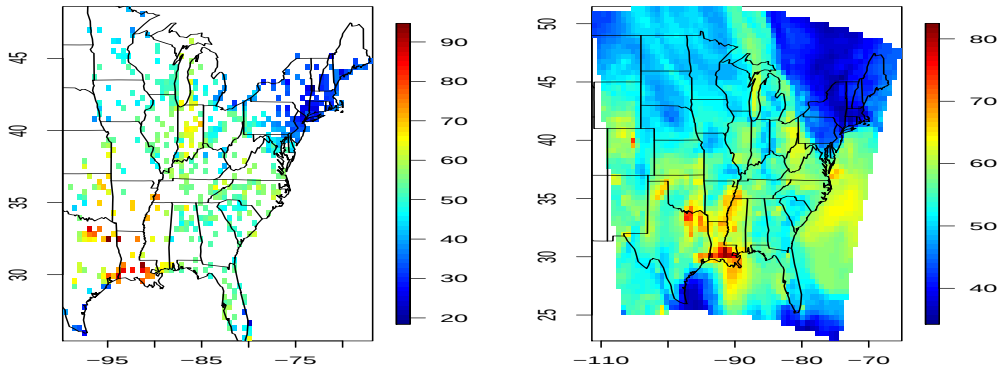
The current method used to predict (or forcast) ozone levels where measurements cannot be taken is the Community Multi-scale Air Quality Model (CMAQ), which mathematically simulates ozone levels based on different characteristics of the area, including ground characteristics,temperatures, and urban density. However, scientists have acknowledged that this model has many flaws and its predictions do not match up exactly with actual EPA measurements taken on the day the model predicted for. The goal of this project is to characterize the relationship between CMAQ predictions and EPA station measurements, and use this model of the relationship to provide better predictions of ozone levels in areas where measurements have not been taken.

I use a spatially-correlated regression model to evaluate this relationship between the CMAQ model's predictions for ozone levels and the EPA's station measurements, because this model matches the characteristics and nuances of the data (which are explained more in following sections) and provides a useful tool in characterizing how well CMAQ predictions explain EPA measurements. After evaluating this relationship, I use Guassian process regression to predict ozone levels in areas where the EPA does not collect station measurements, in the mid and eastern United States.

2 Data Summary and Exploration

The data used for this analysis includes the CMAQ predictions for 66,960 locations in the mid and eastern United States. I also use the EPA station measurements for 800 locations in the mid and eastern United States collected on the same date as the CMAQ predictions were made. Figure 1 gives the maps of both the EPA station measurements (left) and the CMAQ predictions (right). One important thing to note from

Fig. 1. EPA Station Ozone Measurements and CMAQ Ozone Predictions



the EPA station measurements is that their locations are not evenly spread out, an attribute that will be important to keep in mind when building the model.

I combine the two datasets, and match each of the 800 EPA station measurements with the 10 CMAQ predictions that are closest in location to the station where the measurement was taken. I did not add any of the CMAQ predictions beyond the 10 closest because I determined that any further than that, the predictions would not add much additional information to the estimates of station measurements.

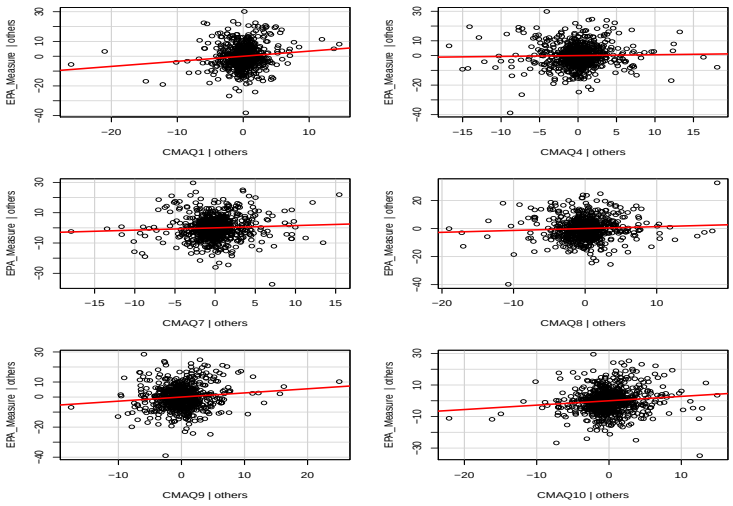
There are a couple of immediately apparent issues with the data. The first is that the range of station measurements is positive, which is not accounted for in a regression model. In order to rectify this issue, I log-transform the values of the station measurements, which allows the response variable to range between any negative or positive number. Although this transformation method is a useful tool to help the response values meet the assumptions of a regression model, we do lose interpretability of the coefficients, which in this case measure how the logarithm of the station measurements change as the values of some of the closest CMAQ predictions change. Given this transformation, the model will be more difficult to interpret, but I determine that this loss in interpretability is worth meeting the regression model assumptions, and supplement this by spending extra time in the Results section explaining how to interpret the coefficient values for the model. Additionally, once the regression model is built and the predictions for the log-transform of the ozone levels are made, the values are transformed back to give the actual predicted ozone values.

While there is not an issue with high dimensionality of the variables, the best subset analysis was run to find the model with the highest adjusted R-squared. This method was chosen because I hoped to find the model that could explain the largest amount of change and variance in the EPA station measurements, so that the results on the relationship between CMAQ predictions and station measurements would be more conslusive and informative. The best model chosen from this subset selection kept the closest CMAQ measurement, and the fourth, seventh, eighth, ninth, and tenth closest measurements. Despite the best subset analysis, I was concerned that there was an issue with collinearity of the explanatory variables chosen because the CMAQ predictions are spatially correlated. I calculate the variance inflation factors (VIF) for these explanatory variables, and find that all of the VIF values are far below 10, indicating that none of these variables are too highly correlated, so there is not an issue with collinearity in the subset model chosen.

I also verify that the data is linear. I provide added variable plots in Figure 2 for all of the variables in the model. The data appears to be fairly linear, although clumped, in these plots. I do not find any specific non-linear patterns in the data that are a concern to the linearity assumptions of a regression model.

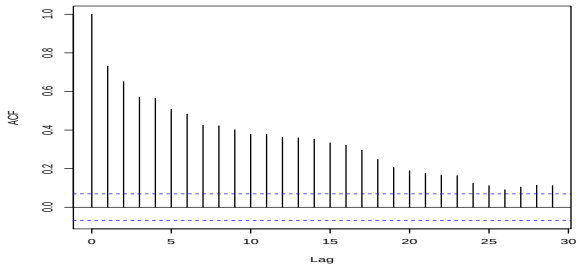
The last potential issue I evaluate in the data are that the EPA measurements appear to be spatially correlated. A simple regression model generally assumes that the response values are independent of each other. I evaluate this assumption in the dataset by building an auto-correlation function (ACF) plot in Figure 3, used to diagnose spatial correlation. The ACF plot shows that measurements that are closer to each other in location are more highly correlated, and that this correlation only gradually decreases as the distance

Fig. 2. Added Variable Plots - Test of Linearity



between the measurements increase. I therefore find that the EPA measurements are spatially correlated, and build a regression model with a correlation structure that accounts for that. Having accounted for and dealt with the potential issues in the data, I proceed to choose the best model and verify the assumptions of the particular model chosen.

Fig. 3. Auto-Correlation Function Plot - Test of Spatial Correlation of EPA Measurements



3 The Model and Its Assumptions

Because the EPA measurements are spatially correlated, I employ correlated regression analysis, accounting for the correlation between the measurements based on their proximity. There are many different options for the correlation structure in correlated regression analysis, such as the auto-regressive or moving average structures. However, both of these structures assume that the observations are equally spaced, which, as mentioned before, is not true of the EPA station measurements. I choose to use the Exponential form with a nugget for the correlation structure, to evaluate the relationship between CMAQ predictions (explanatory variables) and the EPA station measurements (reponse variable). The exponential correlation form is chosen because the distance between measurements are not equally spaced, which this type of correlation structure accounts for. The nugget is used because it accounts for sampling variability and helps stabalize the estimation. In regression analysis for correlated data, the response (EPA measurements) is still assumed to be

normally distributed, but with variance (Σ) that accounts for correlation between measurements based on proximity, and is built differently according to the correlation form chosen. The model is written below and illustrates this assumption, specifically for the exponential form:

$$Y \sim N(X\beta, \Sigma),$$

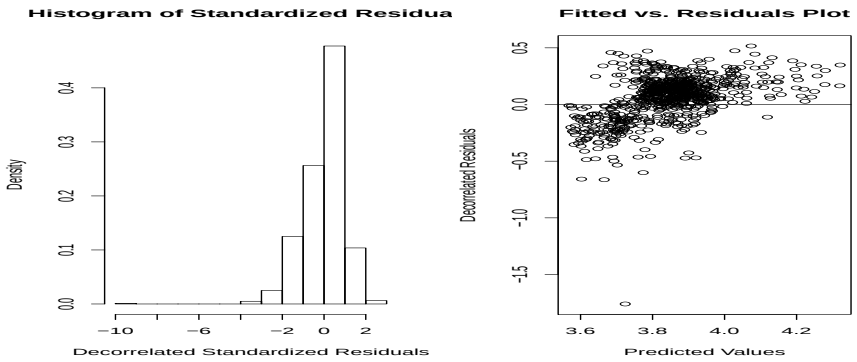
where the correlation structure is calculated as $\Sigma = \sigma^2((1-\omega)R + \omega I)$, where ω is the nugget, σ^2 is the variance of the residuals, R_{ij}=exp$\{-||s_i - s_j||/\phi\}$ ($||s_i - s_j||$=Euclidian Distance), and I is the identity matrix. Note that this structure assumes that the residuals are no longer independent of each other. The maximum likelihood estimators of $\hat{\beta}$, the coefficients, and $\hat{\sigma}^2$, the variance of the residuals, are calculated as follows:

$$\hat{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y \tag{1}$$

$$\hat{\sigma}^2 = \frac{1}{N}(Y - X\beta)'\Sigma^{-1}(Y - X\beta) \tag{2}$$

I also use Gaussian processes (for spatial data, this is called Spatial Kriging) for predictions of the areas were the station measurements have not been taken. The Gaussian process regression assumes that the data is multivariate normal, which means that the residuals should be normally distributed. The model also assumes constant variance. Because the residuals are spatially correlated, we cannot use those to verify the assumptions. I use Cholesky Decomposition to decorrelate the residuals, then use those decorrelated residuals to verify the assumptions. Figure 4 shows a histogram of the standardized decorrelated residuals and a fitted values vs. (decorrelated) residuals plot. The histogram of the standardized residuals shows that the residuals are approximately normally distributed, with the exception of the fairly large outlier on the left side. Also, the fitted vs. residuals plot indicates that we have close to constant variance, again with the exception of the lower outlier. Although the outlier is quite far from the other residuals, I decide to keep it in the model because the goal is to be able to evaluate the relationship between the CMAQ prediction and the station measurements, and outliers are sometimes observed in ozone level measurements.

Fig. 4. Gaussian Process Regression Assumptions Verification



Because the data meets the assumptions of the gaussian process regression, namely the normality and equal variance, I determine that I can proceed with the analysis. I build the exponential form regression model that accounts for correlation in location, and first evaluate the fit, then report the results.

4 Model Fit

I evaluate the fit of the model by finding the R-squared, mean bias, mean root predictive mean squared error (RMSE), mean prediction interval coverage, and mean prediction interval width. The R-square value is

found using the decorrelated residuals found to check the model in the previous section. The value found for R-squared is 0.4647. This indicates that 46.47% of the variation in the station measurements can be explained by the exponential form regression model that accounts for spatial correlation and with the covariates: the closest, and the fourth, seventh, eighth, ninth, and tenth closest CMAQ predictions. This is a respectable R-squared value, but a little lower than hoped for. This could maybe be improved on in future research by looking for other covariates that explain ozone levels. This lower R-squared value is possibly an indication that CMAQ predictions are not sufficient in explaining ozone levels in the United States, and that other factors should be considered. All of the other indicators of fit were found using 10-fold cross-validation. The values are found in Table 1.

Table 1. Model Fit

R-Squared	0.4647
Mean Bias	-0.0547
Mean RMSE	0.1744
Mean Prediction Interval Coverage	0.9495
Mean Prediction Interval Width	0.7221

Recall that all of these values are in terms of the log-transform of the station measurements. Therefore, for example, the RMSE value indicates that, on average, the predictions of the logarithm of the station measurements are off by 0.1744. I find that the bias is quite small in this model, and has a negative value, indicating that the model is generally slightly underpredicting ozone levels. This finding is not surprising given that the residual that appears to be an outlier in the model is much lower than all of the other values. I also find that the prediction interval coverage is almost exactly what is expected of a 95% prediction interval, and that the prediction interval width, on average, is small, meaning that the predictions are fairly precise and cover almost 95% of the data.

I find that, while not as good as expected, the model fits well enough to provide fairly accurate inferential conclusions and predictions. I also note that the lower R-squared value may be an indication that the CMAQ predictions are not sufficient in predicting EPA measurements, and that other variables should be taken into account in future prediction models. I keep these results in mind and discuss them further as I report the results of the analysis in the following section.

5 Results

The first goal of this analysis, explaining the relationship between the CMAQ predictions and EPA station measurements, is partially answered by the coefficient values in Table 2. The CMAQ coefficient estimates indicate that as the CMAQ prediction of the ozone levels for the location that is the ith closest location to the station measurement (where $i = 1, 4, 7, 8, 9, 10$) changes by one unit, the logarithm of the EPA station measurement changes on average by the estimate value for the corresponding coefficient.

Because there are multiple tests performed on the model, a Bonferroni correction is used to adjust the $\alpha = 0.05$ for these tests. Ten coefficients (including the variance components) are tested in the model for significance, so the significance level to reject the null hypothesis is adjusted to $\alpha* = \alpha/10 = 0.005$, which is a conservative adjustment. Coefficients with p-values lower than 0.005 in the table are considered significant to the model. Under this criteria, I find that the only significant coefficient is the CMAQ prediction that is the closest to the station measurement. I do not find this result surprising, given that the prediction closest to the measurement should be the most significant and similar to the EPA measurement, because ozone levels are not isolated, and carry over and affect other areas close by.

Table 2. Exponential Correlation Model Coefficient Estimates

Coefficient	Estimate	95% CI	p-value
(Intercept)	3.09531	(2.91517,3.27546)	0.0000
CMAQ (closest)	0.00551	(0.00255,0.00847)	0.0003
CMAQ (4th closest)	0.00155	(-0.00096,0.00407)	0.2258
CMAQ (7th closest)	0.00123	(-0.00150,0.00395)	0.3782
CMAQ (8th closest)	0.00310	(0.00071, 0.00548)	0.0109
CMAQ (9th closest)	0.00243	(-0.00006,0.00492)	0.0559
CMAQ (10th closest)	0.00009	(-0.00241,0.00258)	0.9464

The regression model also estimates the values used to calculate Σ, the variance-covariance structure discussed above. Table 3 gives the estimates and the 95% confidence intervals for these variance components. Again, this covariance structure accounts for the dependence of the residuals in this dataset, and estimating these provides a more accurate model.

Table 3. Exponential Correlation Model Variance Components Estimates

Coefficient	Estimate	95% CI
σ	0.18452	(0.14367, 0.23699)
Range	5.32770	(2.39486,11.85222)
Nugget (ω)	0.28894	(0.16500,0.45522)

The fitted values from the model are compared to the observed station measurements in Figure 5. The plots appear to indicate that the fitted values for the station measurements of ozone levels are very close to the actual observed values. The most apparent exception is the ozone measurements for the mid-eastern states, particularly Indiana and Ohio, where the model's fitted values appear consistently lower than the observed station measurement values. This finding is consistent with the negative mean bias found in the previous section.

The second goal of this study is to use the model that explains the relationship between CMAQ predictions and EPA measurements to predict of ozone levels for areas where the EPA did not collect the measurements. The predictions were found using Gaussian processes (Spatial Kriging), which assumes that predictions are also spatially correlated. The predictions for the locations provided can be found in Figure 6. The predictions for the log-transform of the ozone levels were found using the model, then these predicted values were exponentiated, to give the actual ozone level predictions. The uncertainty was estimated using prediction intervals, however, I did not include the maps with these lower and upper bounds because they appeared almost identical to the prediction map in Figure 6. However, from the fit verification, we know that the prediction intervals have 94.95% average coverage, and that the prediction intervals are, on average, quite small. The prediction intervals are both precise and provide approximately the expected coverage.

I find that CMAQ predictions near EPA measurement stations explain quite a bit of the variation in the EPA measurements, although these CMAQ predictions may not be sufficient in explaining EPA measurements. However, evaluating this relationship using a regression model provides a better understanding of the relationship between the CMAQ predictions and the EPA observations of ozone levels, and this model appears to be useful in predicting ozone levels in the United States. This is an important finding given that it is becoming increasingly important for the EPA to monitor the ozone levels in the United States in order to create the best regulations and policies to prevent the respiratory health issues caused by breathing in high levels of O3.

Fig. 5. Observed Ozone Measreuments vs. Fitted Ozone Values

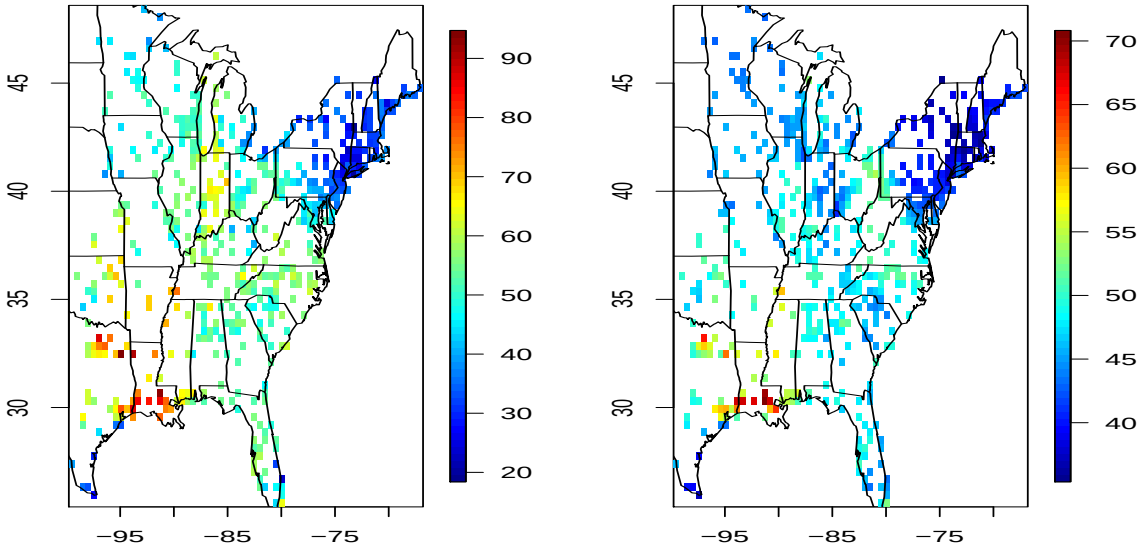
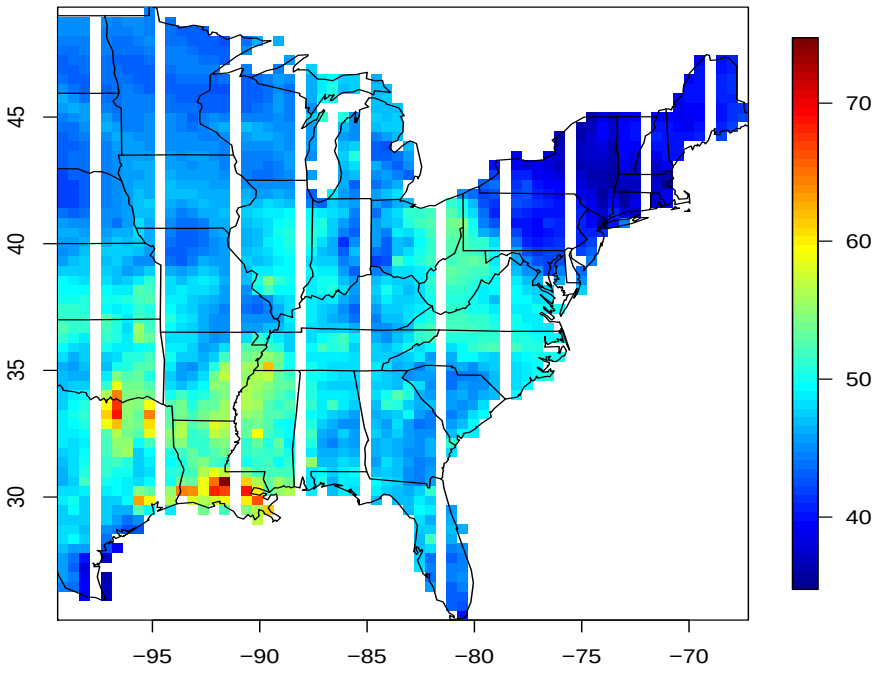


Fig. 6. Predictions for Areas not measured by the EPA



6 Conclusion

While this model can be improved on, the findings in this study provide important information regarding the relationship between CMAQ predictions and EPA station measurements and provide a more accurate method to predict the ozone levels in the eastern United States. This is a useful tool for the EPA in their quest to find better prediction models for the ozone levels in the United States, especially as ozone level prediction becomes a more urgent policy concern with the growing levels of O3 in the environment that cause greater levels of respiratory health issues.

Despite the usefulness of this model, it does not account for every CMAQ prediction that could affect a station measurement, only the ten closest predictions, and no other explanatory variables are considered in the model. Future researchers should evaluate whether the CMAQ predictions that are further than the 10th closest to the station measurement significantly affect the model, and look for other variables that could explain ozone levels, in an attempt to explain more of the variability in the EPA's ozone measurements (improve the R-squared value). The other issue with this model is that the data used is only taken from a single day of CMAQ predictions and EPA measurements, which potentially makes the findings not generalizable. Future researchers should also expand the dataset to evaluate multiple days across different seasons and years to evaluate the generalizability of this model. Despite these shortcomings, this model provides a glimpse into the relationship between CMAQ predictions and EPA measurements, and provides a useful tool to predict ozone levels in the United States.