Ozone Prediction with CMAQ

David Lowe

March 23, 2017

1 Introduction

As the world industrializes, pollution becomes an increasingly prevalent problem. Air quality has decreased significantly since the industrial revolution, and only continues to deteriorate. Ozone (O_3) is a composite pollutant produced from burning fossil fuels and refinery byproducts. This gas poses a health hazard when breathed in high quantities. When exposed to large amounts of ozone, humans can experience chest pain, bronchitis, emphysema, and asthma. Ozone is measured with an air quality rating; high ozone levels (100+) are more hazardous than low ozone levels (0-100). For this reason, information about ozone levels is of national interest, as it affects the health of the population, and environmental regulation legislation. Ground stations located across the country measure ozone, though they can only measure ozone at a local level.

Community Multi-scale Air Quality Model (CMAQ) is a widely available measurement of air quality, based on temperature, urban density, and other ground characteristics. Like ozone, CMAQ is measured on a scale from low to high, indicating worse air quality at higher levels. CMAQ ratings have been used to estimate ozone, however, the relationship between the two is not perfect. Though not entirely accurate, scientists hope to understand how well CMAQ can explain ozone levels, as well as predict ozone values at any given location. Therefore, the goal of this analysis is to a) estimate the relationship between CMAQ and ozone, and b) predict ozone given the CMAQ measurements near a given latitude and longitude.

1.1 Data

To make inference and prediction on ozone given CMAQ, we are given the ozone measurements from 800 ground stations (represented by longitudinal coordinates) that span across the Eastern United States, but do not cover the entire area of the region (see Figure 1). The ozone air quality ratings range from 7.13 to 106.63. CMAQ is measured at 66,960 longitudinal coordinates, and does cover the entire Eastern United States (see Figure 2). The available CMAQ measurements range from 25.50 to 97.74. Because the longitudinal coordinates between CMAQ and ozone are not exactly the same, the Euclidean distance is calculated between each ground station (ozone coordinate) and all CMAQ coordinates, and the CMAQ values are ordered by proximity to the

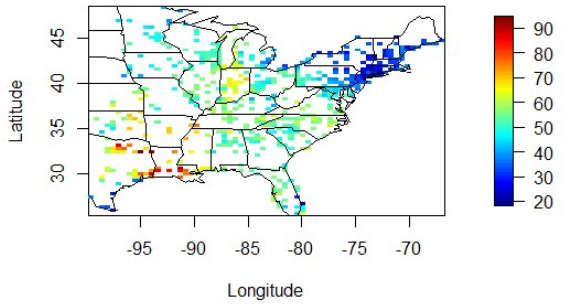

Figure 1: Plot of ozone levels for 800 ground stations across the Eastern United States. Red indicates a high ozone measurement, while blue has a low ozone measurement.

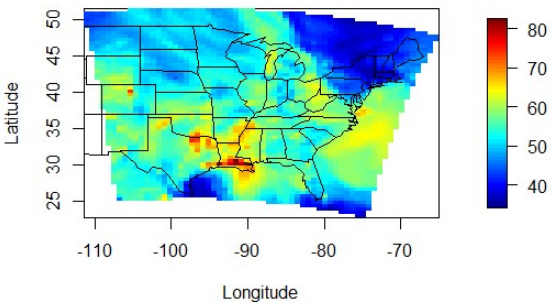

Figure 2: Plot of CMAQ levels across the Eastern United States. Red represents a high CMAQ measurement, while blue indicates a low CMAQ measurement.

800 stations. The final dataset contains the longitudinal coordinates of the ground stations, the observed ozone level, and the CMAQ measurements, ordered from closest to furthest from the ground station.

As can be seen in Figures 1 and 2, ozone and CMAQ are spatially correlated. That is, longitudinal coordinates that are close together are likely to have similar ozone levels. This spatial correlation structure must be built into the model if we are to accurately report the variability in the predictions. Leaving correlation out can overstate the fit of the model, which leads to ozone prediction intervals that are too narrow. Since CMAQ is also spatially correlated, CMAQ measurements are highly collinear. Including multiple collinear variables will inflate the confidence intervals around the estimated effect of CMAQ on ozone, confounding the effect of each CMAQ measurement with the others. Each of these characteristics must be addressed by the model selection and implementation process, and will be discussed in further detail in future sections.

2 Model & Methods

2.1 Model

The goal of the analysis is to understand the relationship between CMAQ and ozone, and to use nearby CMAQ measurements to predict ozone. From the 800 available ozone measurement stations we hope to predict ozone over the entire span of the Eastern United States, given CMAQ measurements. Both of these measurements are highly correlated over space. A correlated regression model suits the goal of this analysis: inference and prediction of ozone, given CMAQ. The Gaussian Process provides a way to include spatial correlation between ozone measurements, and predict ozone levels at any given longitudinal coordinate. Proximate CMAQ scores can be represented in the model with principal components, thus accounting for collinearity between CMAQ values at different distances. Principal components also compress the information contained in several CMAQ levels into to a single, linear predictor variable. The correlated regression model we will be using to understand the relationship between CMAQ and ozone is written out in Equation 1.

$$O_3 = \beta_0 + Z\theta + \varepsilon \tag{1}$$

$$\varepsilon \sim \mathcal{N}(0, \Sigma)$$

$$\Sigma = \sigma^2((1-\omega)\mathbf{R} + \omega\mathbf{I}) \qquad \mathbf{R}_{ij} = \exp\left\{\frac{-|t_2 - t_1|}{\phi}\right\} \qquad \exp\left\{\frac{-|t_2 - t_1|}{\phi}\right\} \in (0, 1)$$

Equation 1 explains ozone levels using an intercept, β_0; a principal component made up of CMAQ measurements spatially close to ozone observations, $\mathbf{Z}$, along with its effect, θ; and a correlated error term, ε. β_0 represents the ozone level when all CMAQ measurements adjacent to the ozone measurement (contained in the principal component, $\mathbf{Z}$) equal 0. The ordered CMAQ measurements are contained in the principal component, $\mathbf{Z}$. The principal component is an orthogonal component (or linear combination) of the CMAQ measurements. This orthogonal transformation makes a linearly uncorrelated set of values that contain the majority of information contained in the CMAQ values, thus taking away any collinearity between CMAQ measurements at different distances from the observed ozone level. The coefficient, or effect, of the principal component, is represented by θ. While there is little interpretability in context of the principal component, θ can be converted back to the coefficients of CMAQ at the CMAQ distances contained in the principal component.

2.2 Correlation Structure

The last piece of the model, ε, provides the spatial correlation structure needed to correctly capture the variability between ozone measurements. The errors are distributed normally with an exponential correlation function, shown by the equation contained in $\mathbf{R}_{ij}$. The $|t_2 - t_1|$ represents the distance between two longitudinal points, and allows for distances that are not equally spaced across the span of the country. The ϕ parameter is the range coefficient. For a fixed distance, if ϕ is large, the correlation increases. For example, if ϕ were to be ∞, the correlation function goes to 1 ($e^{-|t_2-t_1|/\infty} = e^0 = 1$). The exponential correlation function will have a small correlation for locations further apart, and a large correlation for locations closer together.

Σ represents the full covariance structure, including a nugget, ω. The nugget allows for sampling variability when the measurements are at the exact same location, and helps stabilize estimation. The σ^2 term represents the variance of the errors, or difference between the predicted values of ozone and the observed values of ozone at each of the 800 locations. $\mathbf{I}$ represents the identity matrix. Overall, the covariance structure contained in the errors, ε, allow us to account for the spatial correlation of the ozone measurements.

2.3 Model Assumptions

This model assumes a Gaussian Process, suggesting that any finite collection of ozone measurements is distributed as a Multivariate Normal (MVN) distribution. Ozone levels at any given location are then univariate normally distributed. Equal variance is also assumed at any location of ozone measurements. The Gaussian Process also assumes linearity, as we are using an intercept and the principal component of the nearest CMAQ values.

If these assumptions are met, the correlated regression model we have now defined meets the initial goals of the analysis. By using the intercept and principal component of the nearest CMAQ measurements, we can understand the relationship between CMAQ and ozone measurements. Principal components also eliminate the correlation (collinearity) between the various distances of CMAQ measurements. Spatial correlation between ozone levels is accounted for by using the exponential correlation function. The Gaussian Process assumes that the unobserved ozone levels are also distributed MVN, which provides a means to predict the ozone level at any given longitudinal coordinate, by conditioning the missing ozone levels on all observed values. Each goal is met by this model, and we can proceed to verify the model's assumptions, and test the model's performance in the aforementioned goals.

3 Model Justification & Performance Evaluation

3.1 Variable Selection

The principal component, $\mathbf{Z}$, compresses information from the values of CMAQ to be explained by a single value, and reduce the collinearity between CMAQ measurements. By using principal components we could include as many CMAQ values as we'd like. However, to retain interpretability and simplicity in the model, it is important to limit the number of CMAQ measurements included.

The CMAQ measurements used in the principal component were chosen using a variable selection method. To select how many CMAQ measurements to include, the values of CMAQ closest to the ground stations were added to a linear model (without a correlation structure) to explain ozone. Subsequently, the next closest CMAQ levels were then added one by one. After including the six closest CMAQ values, the Bayesian Information Criteria (BIC) was minimized. The linear model has an R^2 of 0.6873, indicating that the linear model fit with the six closest CMAQ measurements accounts for 68.73% of the variability in ozone. Table 1 gives the average distance between each of the six closest CMAQ measurements and the ground station where ozone was measured. By using only the first six CMAQ measurements, we are effectively saying that the CMAQ measurements within 10 miles of the location we are interested in are most useful for predicting, and understanding ozone.

The six selected CMAQ values are still highly collinear, and will inflate the variance of the effect of each CMAQ measurement on ozone. Figure 3 shows that the relationship between the CMAQ measurements is very highly correlated. A helpful characteristic of principle component analysis is that it eliminates collinearity in the variables it contains. The first principal component explains 94.23% of the information contained in the actual six CMAQ measurements. Therefore, the single principal component will be sufficient to explain the relationship between CMAQ and ozone levels, and will eliminate any collinearity between CMAQ values.

Table 1: Average distance between ground station where ozone is measured, and the six selected CMAQ measurements.

	$CMAQ_1$	$CMAQ_2$	$CMAQ_3$	$CMAQ_4$	$CMAQ_5$	$CMAQ_6$
Average Distance (in miles)	2.78	5.29	6.79	7.66	9.26	9.81

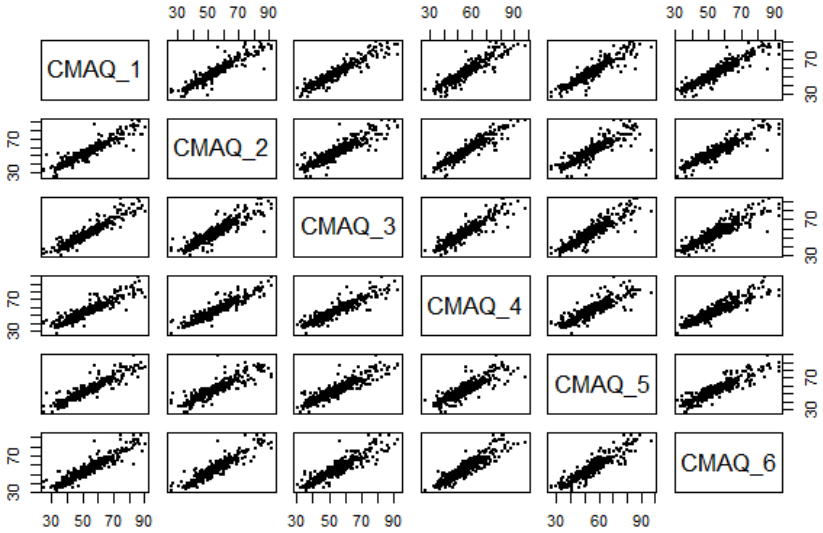

Figure 3: Collinearity between each of the closest six CMAQ measurements. A strong linear trend indicates that the variables are collinear.

3.2 Assumptions

The Gaussian Process has only a few assumptions, and many are hard to verify. The first assumption is that ozone measurements at each location is normally distributed, and therefore, any finite collection of ozone measurements from different locations is MVN. Our data contains only one ozone measurement for each location, making it impossible to verify the normality of ozone at each location. We assume normality in ozone levels at each location, as we have no reason to believe otherwise. Along with the assumption of normality at each location, the Gaussian Process assumes equal variance at each location as well. Again, with only one measurement of ozone at each location, equal variance is assumed, but can't be proven. The assumption of linearity is verified in Figure 4. CMAQ has a clear linear relationship with ozone, making linearity a good assumption. The linearity assumption is retained when using a principal component, which produces a linear combination of CMAQ measurements.

The model assumes that ozone levels are not independent, rather that they follow an exponential correlation structure. This correlation structure assumes that points (locations) closer together are

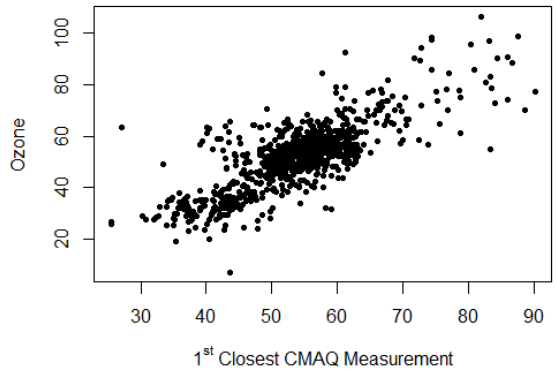

Figure 4: The relationship between ozone levels and CMAQ measurements have a clear linear trend.

more likely to have similar ozone levels than points further apart. Figure 1 suggests that ozone levels closer together do act similarly, and therefore the exponential correlation structure assumption is a good one. The Gaussian process is now justified as our choice for modeling ozone levels with CMAQ measurements.

3.3 Model Performance

With the assumptions of the Gaussian Process met, the model should be tested to see how well it performs at predicting ozone levels, using CMAQ measurements. The model's ability to predict is tested through cross-validation. A random sample of 80% of the ozone data is used to build the model, which then attempts to predict the remaining 20% of the ozone values, given their locations and surrounding CMAQ values. This process is repeated 500 times to capture the variability in performance of the model.

The overall bias of the predictions is negligible, as can be seen in Figure 5, and the average predicted root mean-squared error is 5.32. This measurement of variability between predicted values and true values is relatively small, given that the observed ozone levels range from 7.13 to 106.63. From this predicted root mean-squared error, we can see why the 95% prediction interval widths have a mean of 21.16. This means that we are 95% confident that the true ozone level is within 10 from the predicted value on the ozone air quality scale. The cross-validation showed that the model's prediction intervals contain about 94% of the true ozone levels. These results demonstrate that the model is effective at predicting ozone levels accurately. These diagnostics would also suggest that the model fits the data fairly well, though there is no clear way to represent the model fit in this case.

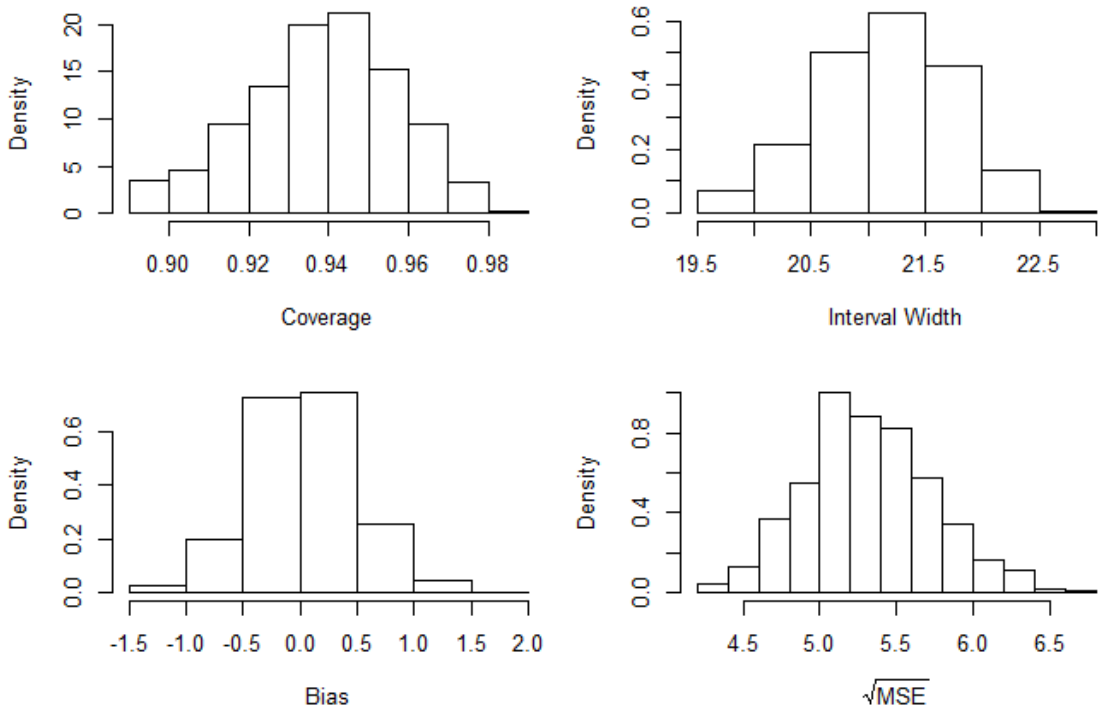

Figure 5: Diagnostics from the cross-validation. Coverage indicates the average percent of prediction intervals that contained the true ozone measurement. Interval width shows the distribution of mean prediction interval widths. Bias indicates the distance between the predicted ozone levels and the true ozone levels. Root predicted mean-squared error indicates how far off the predicted values are on average from the true values. Interval width, bias, and root predicted mean-squared error all can be interpreted on the same scale as ozone measurements.

Table 2: Estimated Model Coefficients with 95% Confidence Intervals

Coefficient	Estimate	Lower 2.5%	Upper 97.5%
β_0	9.5493	2.3997	16.6988
β_{CMAQ_1}	0.1266	0.1075	0.1458
β_{CMAQ_2}	0.1237	0.1050	0.1424
β_{CMAQ_3}	0.1226	0.1040	0.1411
β_{CMAQ_4}	0.1217	0.1033	0.1401
β_{CMAQ_5}	0.1222	0.1037	0.1407
β_{CMAQ_6}	0.1223	0.1038	0.1408

4 Results

Now confident in the fit and prediction performance of the model, we will discuss the results of the model. Once the principal component is expanded out to the original six closest CMAQ measurements, the associated coefficient estimates can be better interpreted. Table 2 gives the estimated relationship between each of the six closest CMAQ measurements on ozone levels, as well as the intercept, β_0. In context, the intercept, β_0, tells us that if all six CMAQ measurements were equal to zero, then the ozone level would be 9.5493. For the first closest CMAQ measurement, the estimated coefficient (estimated effect) can be interpreted as follows: as the closest CMAQ measurement increases by one, and holding all other CMAQ values constant, the ozone level will increase by 0.1266 on average. The confidence interval for the same CMAQ coefficient expresses that we are 95% confident that the true effect of the closest CMAQ value to the ozone measurement location is between 0.1075 and 0.1458. All other coefficient estimates and confidence intervals can be interpreted in a similar fashion.

It is interesting, though not surprising, that each of the coefficient estimates are so close in value. Earlier we showed the high collinearity between CMAQ values in close spatial proximity. It is not unusual then that CMAQ values close to an ozone ground station have similar relationships with the ozone measurement. Each estimate gives an idea of the relationship between CMAQ and ozone.

The model has produced estimated relationships between CMAQ and ozone calculated, thus quantifying the relationship between CMAQ and ozone. Under the assumption of the Gaussian Process, scientists can use this model to understand the relationship between CMAQ values and ozone levels within 10 miles from any given location. This meets the goal to better understand the relationship between CMAQ and ozone.

The second goal, ozone prediction, has been validated earlier. Scientists can predict ozone levels at any longitudinal coordinate by finding the six CMAQ measurements nearest to the coordinate. To demonstrate this, longitudinal coordinates are given that span the entire space of the Eastern United States. Using the six closest CMAQ values, the ozone levels are predicted at each point. Figure 6 visualizes the predicted ozone levels across the entire Eastern United States. The results are believable when compared with the initial plot of ozone in Figure 1. The lower and upper prediction interval bounds are also plotted in Figures 7 and 8 respectively. With the cross-validation results indicating that 94% of the prediction intervals contain the true value, we can be confident in the ozone predictions displayed.

5 Conclusions

The Gaussian Process used to model ozone has proved useful in understanding the relationship between CMAQ and ozone, and has provided a means for predicting ozone with CMAQ. The model also accounts for the spatial correlation between ozone measurements by including an exponential correlation structure in the model. Scientists can predict ozone levels for locations without ground stations by using the closest six CMAQ measurements (spanning a 10 mile radius from the prediction location) in the model, or by using charts such as those in Figures 6, 7, and 8.

While this model can predict ozone levels for any longitudinal point, it can also predict future ozone levels under different scenarios of CMAQ. Such predictions are made possible by the assumptions of the Gaussian Process, and can be used in environmental impact reports. This ultimately can help with city planning and governmental regulations in future developments.

CMAQ gives a fairly accurate representation of ozone on it's own. This result answers the goals of the study; however, the predictions of ozone still have a fairly wide intervals, indicating that there is more information that can be used to understand ozone. Other variables such as environmental landscape and climate readings may help explain ozone levels, and tighten the prediction intervals. Such a model can mitigate future air quality deteriorations, and instigate cleaner solutions in areas such as city development and transportation.

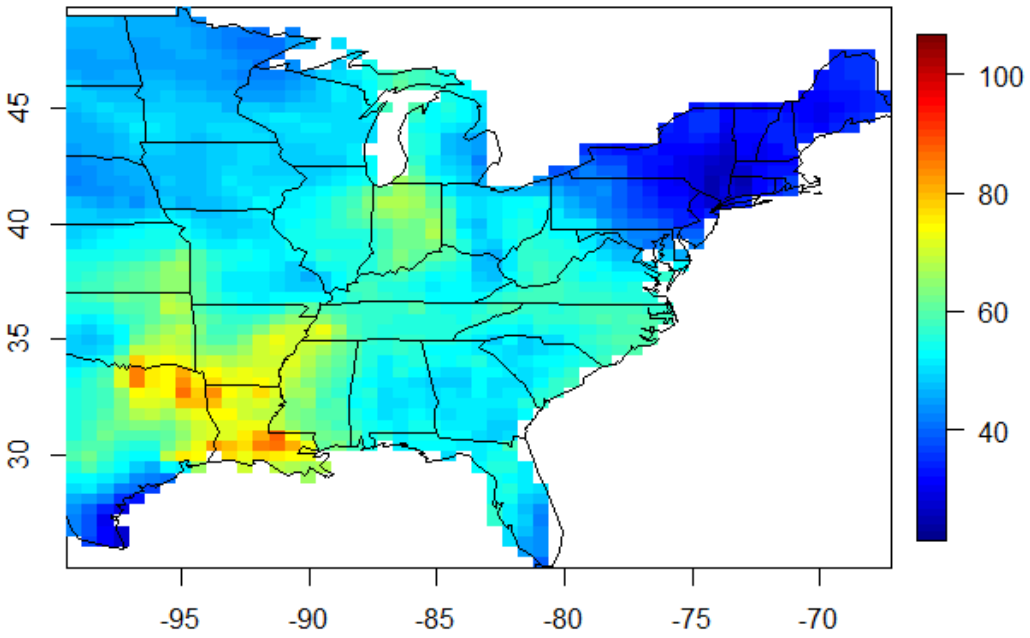

Figure 6: Predicted values of ozone for the Eastern United States.

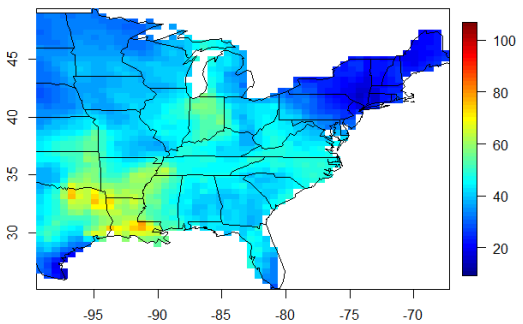

Figure 7: Lower bound of the 95% prediction interval for the predicted values of ozone in the Eastern United States.

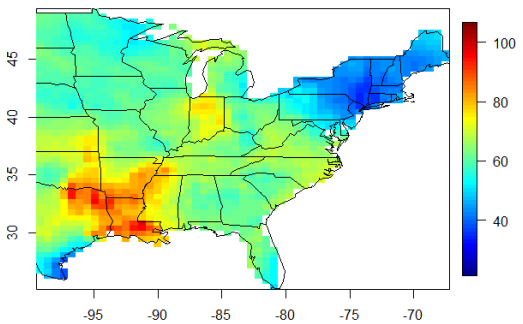

Figure 8: Upper bound of the 95% prediction interval for the predicted values of ozone in the Eastern United States.