Stat536 Midterm - Ozone Data

Arthur Lui

10 March 2014

1 Introduction

Ozone (O_3), a gas in the atmosphere, protects humans from the sun's UV radiation. However, ozone that is close to the ground can be dangerous to humans. Ozone is formed when pollutants react with each other in the presence of heat. It is the main component of smog. Inhaling high concentrations of O_3 triggers chest pain, bronchitis, emphysema, asthma, etc. Scientists have monitored O_3 levels by 1) direct measurement at measurement stations and 2) mathematically simulating the measurements using Community Multi-scale Air Quality Model (CMAQ). CMAQ O_3 measurements are simulated (on a fine spatial scale) based on ground characteristics, temperature, urban density, etc. So, CMAQ data is vast (see **Figure 1**), but not as accurate as direct measurements. On the other hand, direct measurements are sparse (see **Figure 2**). The Environmental Protection Agency (EPA), which monitors O_3, is, therefore, interested in understanding the relationship between CMAQ (which is inaccurate) and station measurements. They eventually hope to predict ground-level O_3 at many locations given CMAQ measurements and station measurements. Using a data set provided by Dr. Heaton, I will construct such a model using a Gaussian Process.

2 Methods & Model Used

The data received consists of longitude, latitudes, and O_3 measurements. We are interested in creating a model to predict O_3 given CMAQ and grid locations. This is a spatial problem. O_3 levels should be highly correlated to other O_3 levels close by, but less correlated to O_3 levels far away. But we would like to incorporate CMAQ as a covariate also so that we can understand the relationship between station-measured O_3 and CMAQ O_3. The model we create should also not assume linearity as we are not certain if CMAQ and station-measured O_3 are linear. The Gaussian Process Model can help us to created such a model.

2.1 Description of The Gaussian Process:

A Gaussian Process is a stochastic process where any finite collection observed random variables follow a multivariate normal distribution. In other words, for any set of $t_1, \ldots, t_N \in T$, $\boldsymbol{Y} = (Y(t_1), \ldots, Y(t_N))^T \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_Y)$.

2.1.1 Full Gaussian Process Model:

Let N = number of observations and,

$$\boldsymbol{Y}|\boldsymbol{W} = \begin{pmatrix} y(x_1) \\ \vdots \\ y(x_N) \end{pmatrix} \sim \mathcal{N}(\boldsymbol{W}, \tau^2 \boldsymbol{I}_N)$$

$$\boldsymbol{W} = \begin{pmatrix} w(x_1) \\ \vdots \\ w(x_N) \end{pmatrix} \sim \mathcal{N}(\mu \mathbf{1}_N, \sigma^2 \boldsymbol{R})$$

which marginalizes to:

$$\boldsymbol{Y} \sim \mathcal{N}(\mu \mathbf{1}_N, \sigma^2 \boldsymbol{R} + \tau^2 \boldsymbol{I}_N)$$

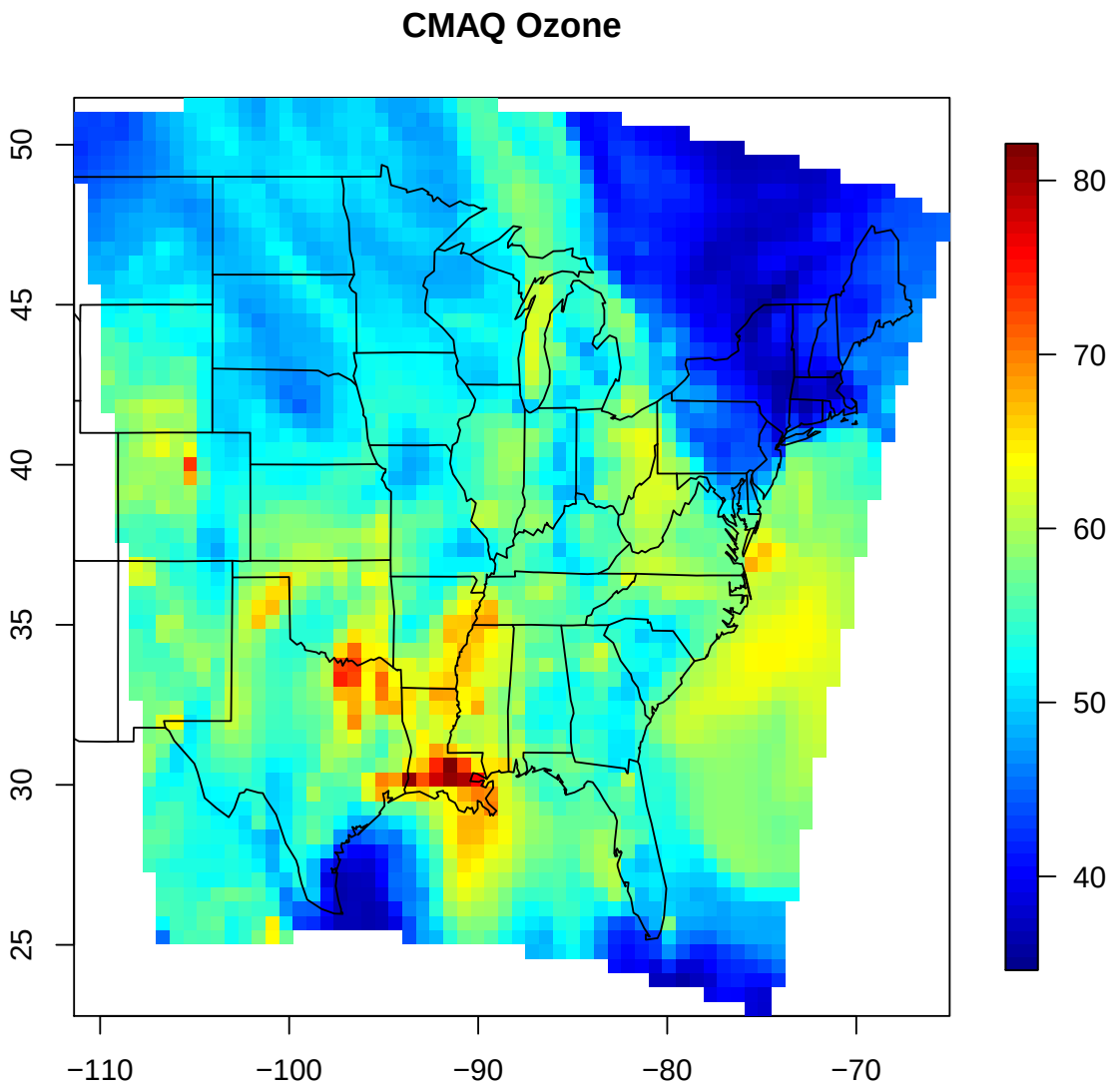

Figure 1: CMAQ measurements of O_3. CMAQ simulates O_3 measurements easily at many locations. But measurements are inaccurate.

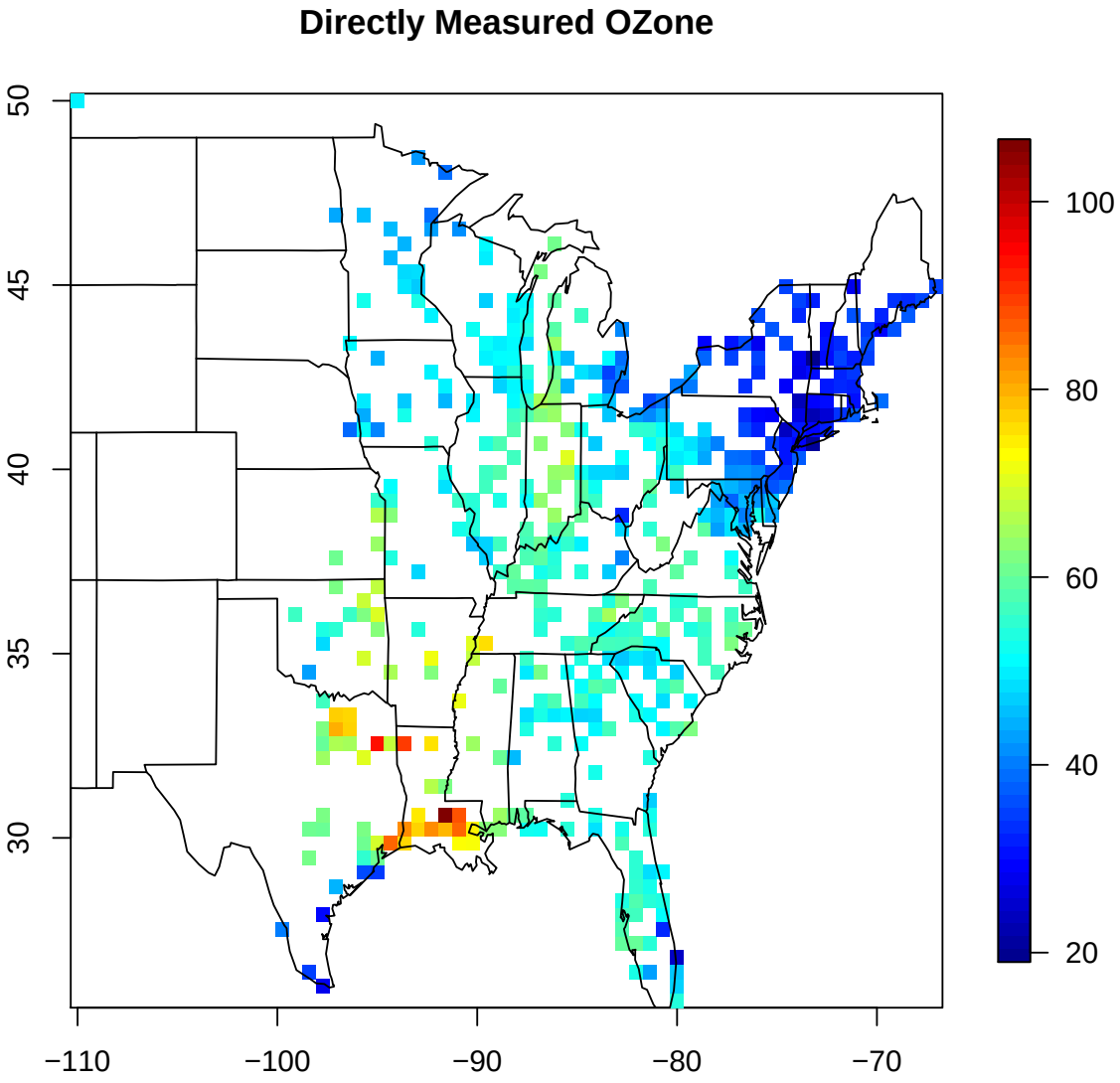

Figure 2: Station measurements of O_3. Direct measurements are sparse.

where,
τ^2 can be interpreted as the error variance;
σ^2 can be interpreted as the spatial variance;

$$\begin{aligned} \boldsymbol{R}_{ij} &= \frac{1}{2^{\nu-1}\gamma(\nu)}(2\phi\sqrt{\nu}|t_i - t_j|)^{\nu} K_{\nu}(2\phi\sqrt{\nu}|t_i - t_j|) \\ &= Matern(|x_i - x_j|, nu = \nu, alpha = \phi) \\ &= Corr(Y(x_i), Y(x_j)) \end{aligned}$$

ϕ: decay parameter. As ϕ increases, correlation (at a fixed distance) decreases.
ν: smoothness parameter. As ν increases, smoothness increases.
K: effective range. distance where correlation decays to 0.05.
We can estimate the unknown parameters $\mu, \sigma^2, \nu, \phi, \tau^2$ (using maximum likelihood or Bayes).

This is a fascinating result that reduces computation significantly. Making use of the properties of the conditional distribution of multivariate normals distributions, one can easily can predictions and prediction intervals (see Rencher & Schaalje, Theorem 4.4d, 2008).

2.2 Assumptions

3 Model Justification

3.1 Why choose a Gaussian Process?

The Gaussian Process model incorporates the spatial distance of the points at which a prediction is made to model nonlinear relationships between a response $\boldsymbol{Y}$ and covariates $\boldsymbol{X}$. Since it is the case that we want to base our predictions of ozone in this way, where our covariates are some linear combination of the nearest CMAQ levels, the Gaussian Process is very suitable for this analysis.

In spatial statistics,y(s), the (functional) response at a grid location, is highly nonlinear. 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 $\mathcal{D}$. We can set up the Spatial Statistics Model in this way:

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

where,

- $\boldsymbol{Y}$ = Station Measured Ozone
- $\boldsymbol{X}$ = a column of 1's followed by the 10 CMAQ values nearest to our prediction location
- $\boldsymbol{\beta}$ = a vector of constants to be estimated
- $\boldsymbol{R}_{ij} = \prod_{p=1}^{P}$ Matern($||\boldsymbol{s_{ip}} - \boldsymbol{s_{jp}}||$,$\nu$,$\phi$),
- $\phi, \nu, \sigma^2, and \tau^2$ can be estimated using maximum likelihood.

3.2 Are Assumptions Justified?

A plot of the histogram of the residuals shows that the residuals are normally distributed with mean 0 and covariance matrix $V = \sigma^2\boldsymbol{R} + \tau^2\boldsymbol{I}$ (see Figure 3).

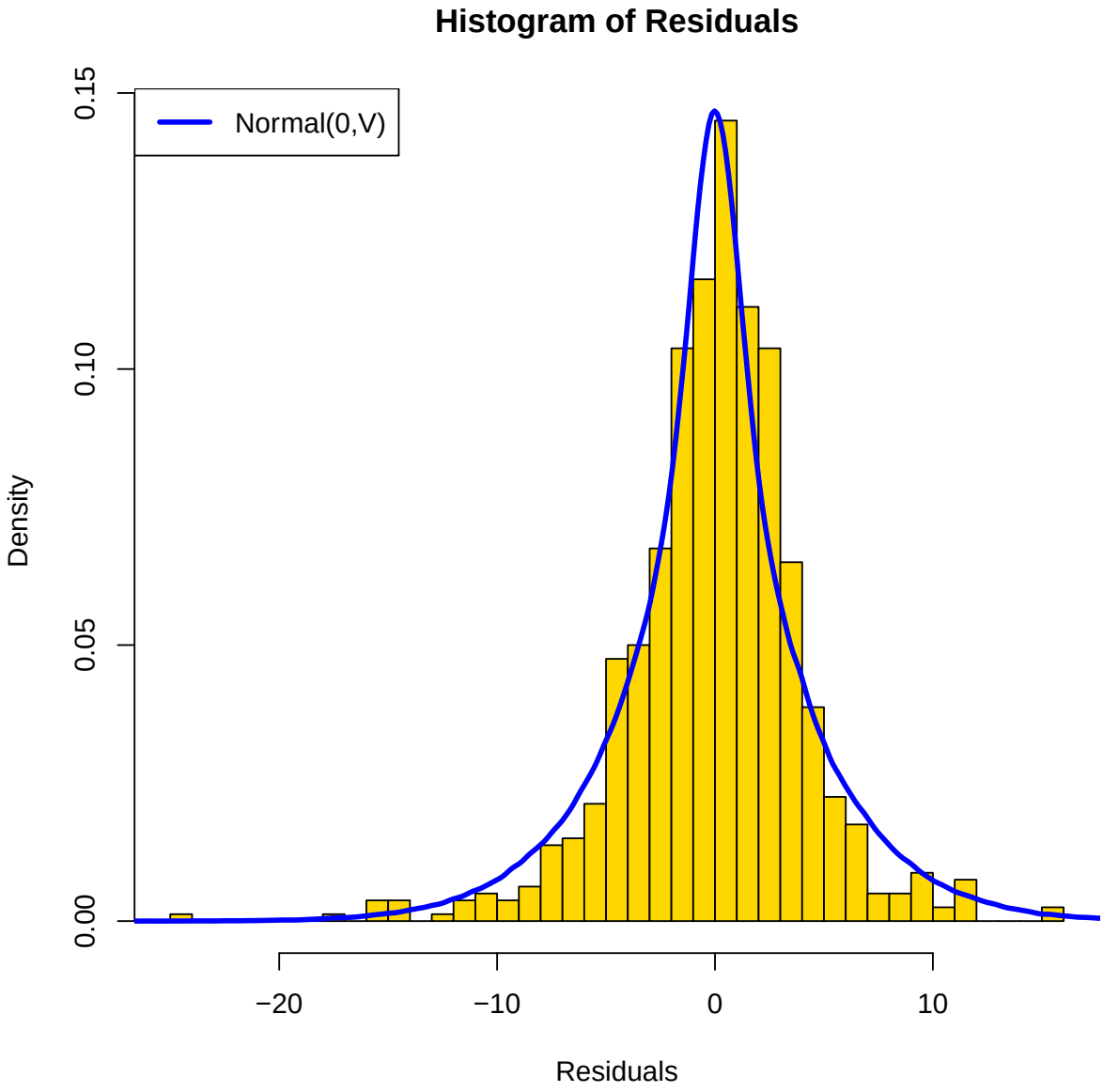

Figure 3: Plot of histogram of residuals.

4 Results

4.1 Estimates of Parameters and CI

Estimates of the parameters are listed as follows: $\hat{\tau^2} = 19.61, \hat{\sigma^2} = 1.04, \hat{\phi} = 40.88$, and ν was chosen to be 2 because the data doesn't inform us about its value and because it gives a smooth curve. **Table 1** shows the parameter estimates with their 95% confidence intervals.

Table 1: Parameter Estimates

	Estimates	CI.Lo	CI.Hi
β_0	7.93766	1.29157	14.58375
β_1	-0.11039	-0.26032	0.03954
β_2	0.20544	0.06538	0.34549
β_3	0.08400	-0.04960	0.21761
β_4	0.20130	0.06727	0.33534
β_5	-0.11577	-0.24077	0.00924
β_6	0.17512	0.05527	0.29497
β_7	0.14480	0.01922	0.27037
β_8	0.09521	-0.01989	0.21031
β_9	0.06374	-0.06132	0.18880
β_{10}	0.04647	-0.06892	0.16187

4.2 Interpretation of Parameters

β_0 is the expected O_3 level if the 10 nearest CMAQ measurements are 0. So β_0 can be thought of as the bias of CMAQ for O_3. Since the 95% confidence interval for β_0 does not include 0, CMAQ is significantly biased (by 7.94) for O_3. Also when X_i, the i^{th} nearest CMAQ location (with respect to the prediction location), increases by 1 unit, $\boldsymbol{Y} = O_3$ is expected to increase by β_i, for i $\in \{1, \ldots, 10\}$.

4.3 Coverage & MSE

Table 2 summarizes the estimate and estimates of the coverage and MSE.

Table 2: Coverage and MSE

	Estimate	CI.Lower	CI.Upper
Coverage	0.931	0.913	0.949
MSE	20979.047	11019.079	30939.014

4.4 Predictions & Uncertainties

After obtaining parameter estimates for β and the scalar parameters, predictions can be made very easily using the conditional distribution of the normal distribution, as mentioned previously.

Predicted O_3 values and their 95% confidence intervals were plotted (see **Figures 4-6**). That is, for each colored grid location, the color represents the predicted O_3 (in **Figure 4**). Also for each colored grid location, the 95% lower limit of the confidence interval was represented in **Figure 4**, and the 95% upper limit of the confidence interval was represented in **Figure 5**.

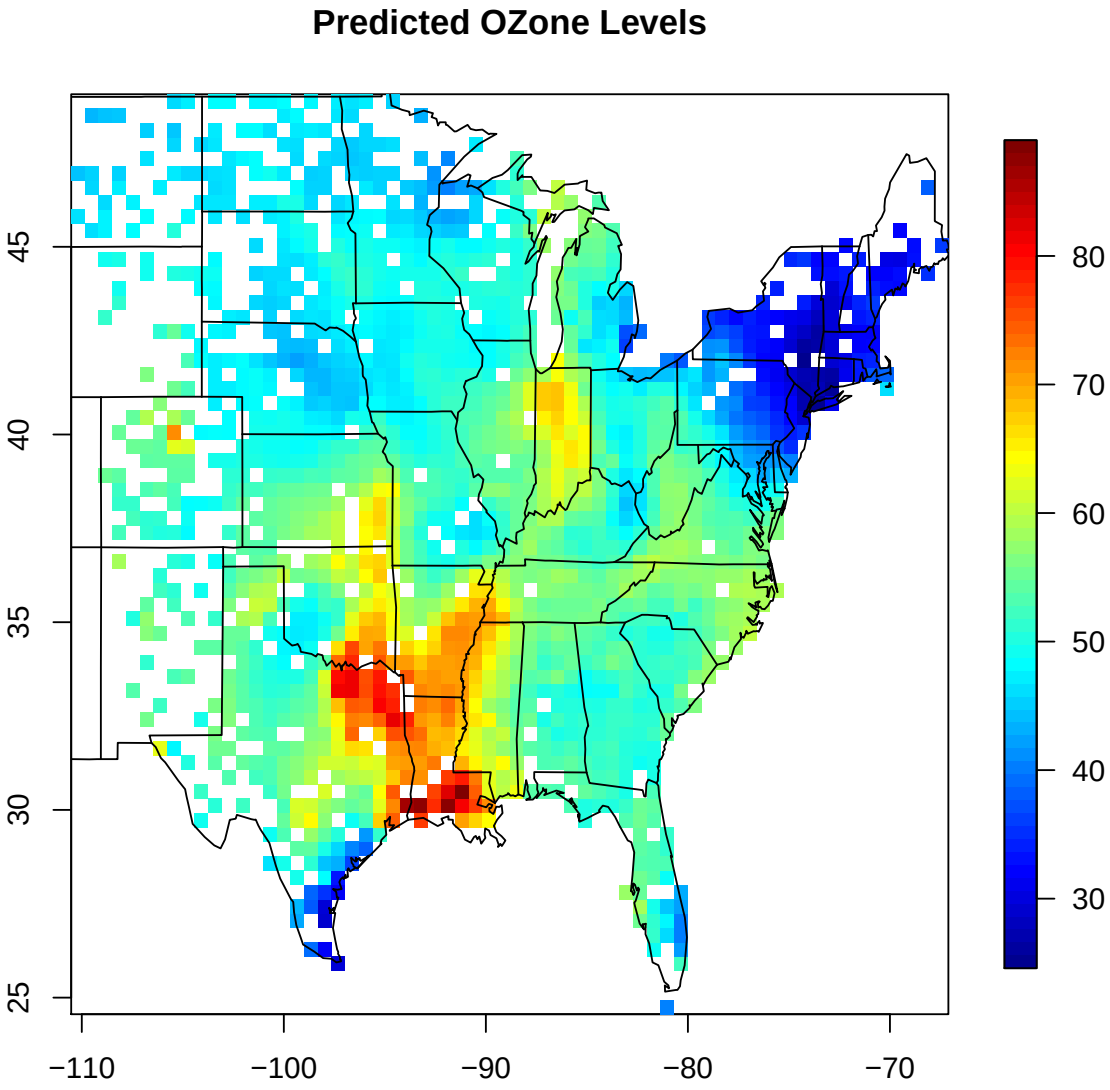

Figure 4: Predicted O_3.

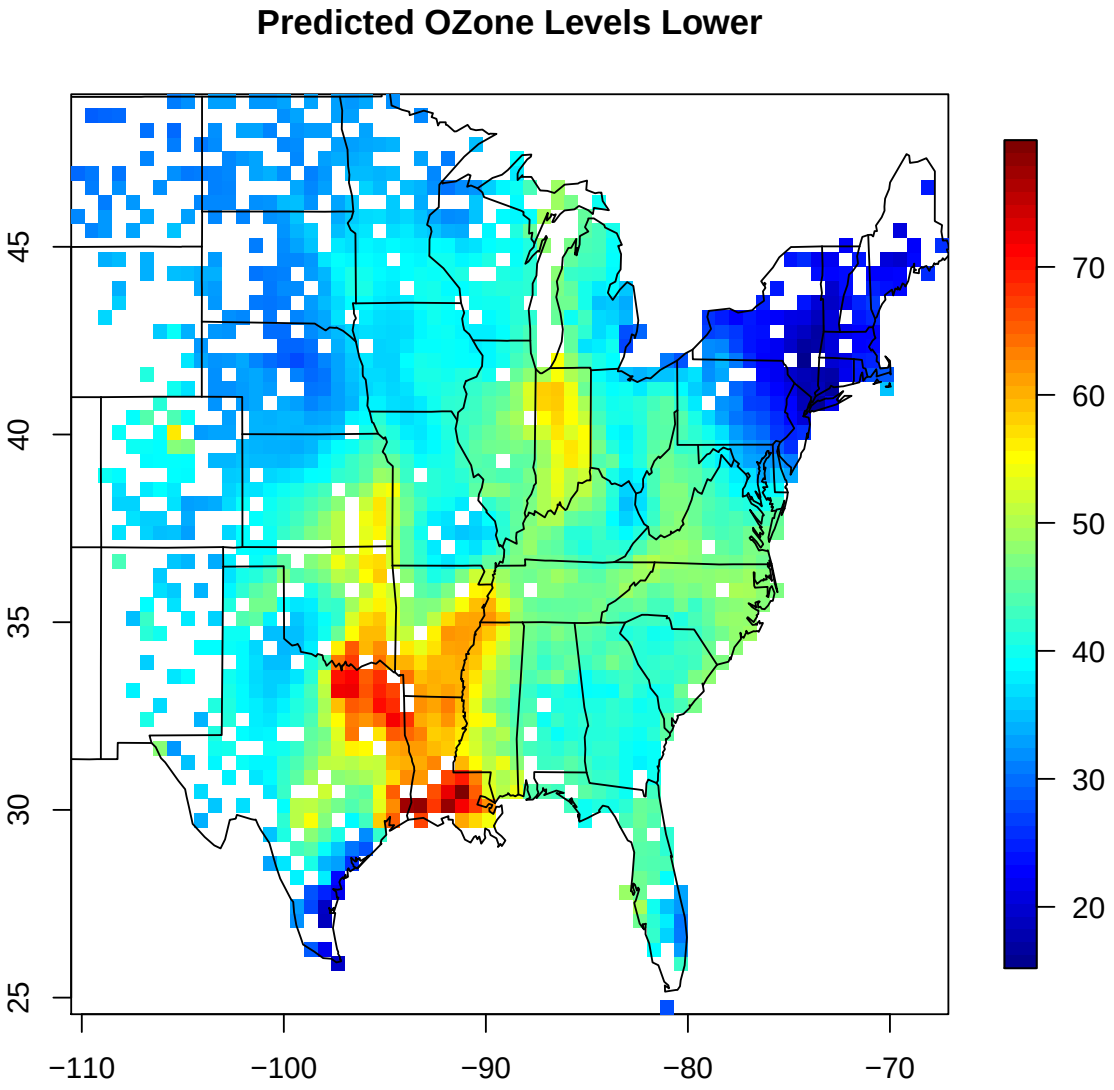

Figure 5: Lower Predicted O_3.

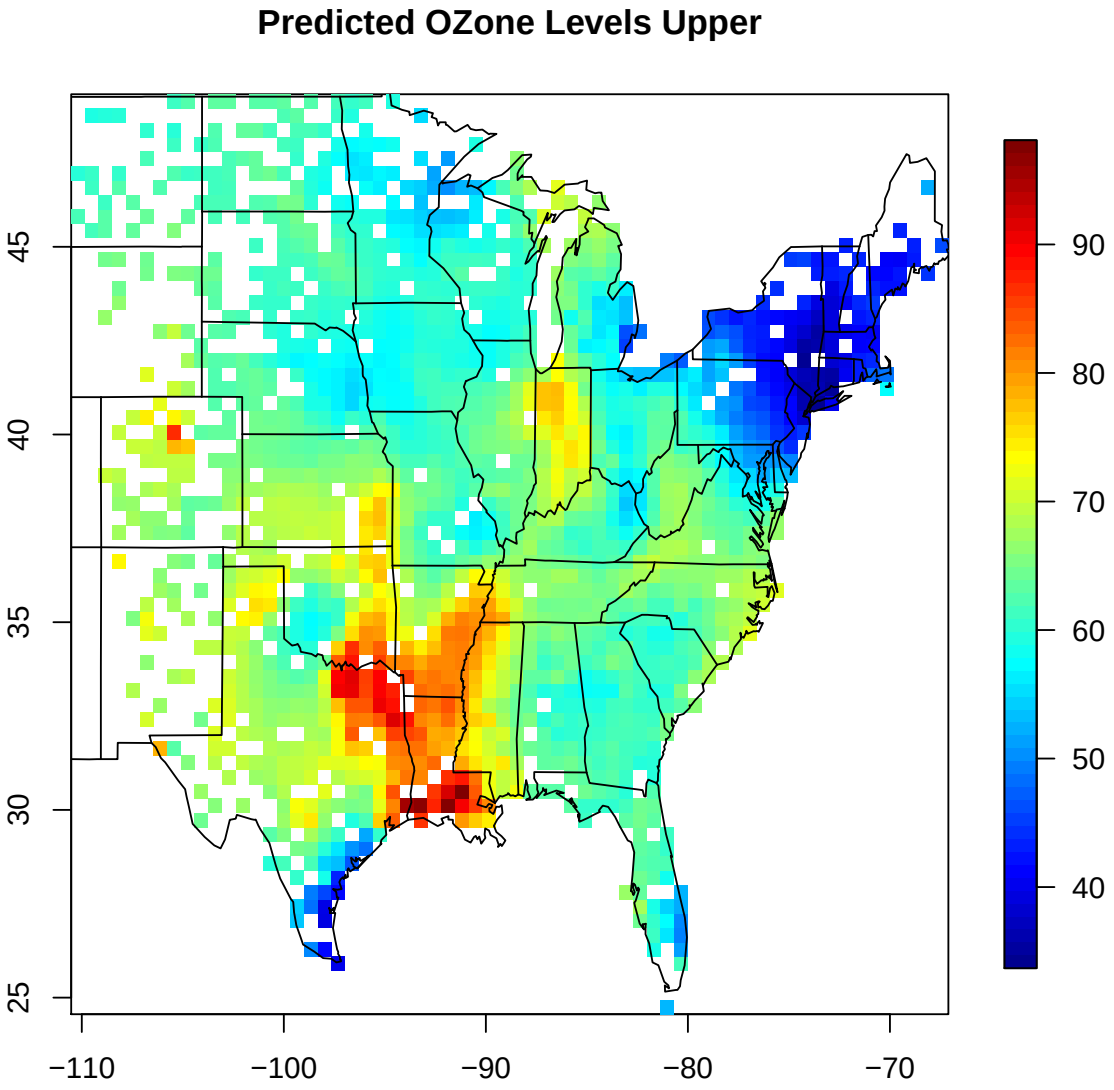

Figure 6: Upper Predicted O_3.

4.5 Main Points

Figure 7 is just a combination of all the figures included so far. We can easily see that corresponding regions have higher O_3 values in the the top left plot than the bottom left plot. We can also see that where there is even a little bit of data in the bottom right graph, the middle right graph borrows more information from the bottom right graph. Where there is only very little data in the bottom right graph, the middle right graph borrows more information from the top right graph.

5 Conclusion

5.1 Potential Alternative Approaches to Investigate

Alternate approaches can be taken to compute the $\boldsymbol{X}(s)$ matrix defined in the Statistical Spatial Model. I chose the 10 nearest locations because I thought it was computationally easy and sensible. That is, O_3 values at a certain location is more affected by O_3 levels near by, but hardly affected by O_3 levels far away. The 20 nearest points could have been chosen. Cross validation could have been used to determine how many of the nearest points to include in the model. Likewise, all the 66,960 points would have been included in the model, but that would be computationally difficult. Forward selection, lasso, ridge, or principal components could also be used in model selection to determine which covariates to include. It could be the case that due to collinearity, every other closest point should be included in the model, but not every closest point, to increase prediction accuracy while reducing collinearity.

5.2 Shortcomings of Gaussian Process

The Gaussian Process is good at fitting models to nonlinear, and spatial data. The Gaussian Process is also convenient in the sense that it is unlike b-splines, where the number of knots needs to be predetermined. One draw-back of the Gaussian Process is that inverting large matrices is computationally expensive. Another draw back of the Gaussian Process is that the rules of determining how to create the covariate matrix $\boldsymbol{X}$ can be quite clumsy. And searching for the model that produces the best (most accurate) predictions will be difficult and time intensive as the algorithms will take a long time to implement.

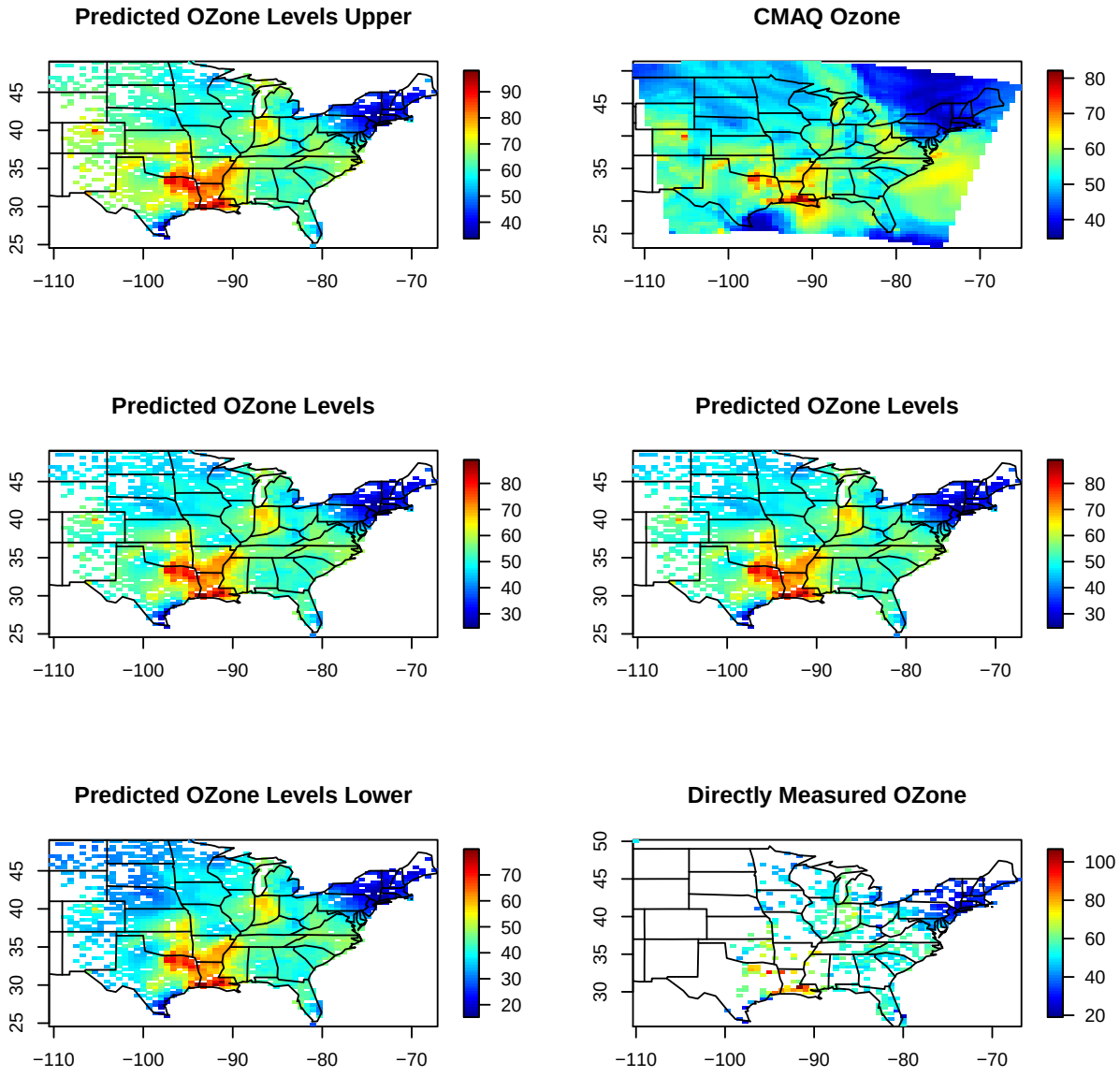

Figure 7: A comparison of the different plots.