Ozone in the Northeast United States

Mickey Warner

1 Abstract

We fit a Gaussian process to ozone measurements at 107 locations in the northeastern United States. We found there to be linear trends in longitude and latitude as well as an interaction between the two. We attempted to fit a GP with a Matérn correlation and observation error, but this appeared to be unnecessary as our estimated implied there to be little to no covariance between observations.

2 Introduction

2.1 Data

The region we consider here comprises New England and a few states below along the east coast. Along this roughly 700 miles of east coast we have 107 measurement locations (Figure 1). Measurements are made up of yearly ozone averages (in parts per million) for 2015. Included in our data is the altitude (meters above sea level) at each location.

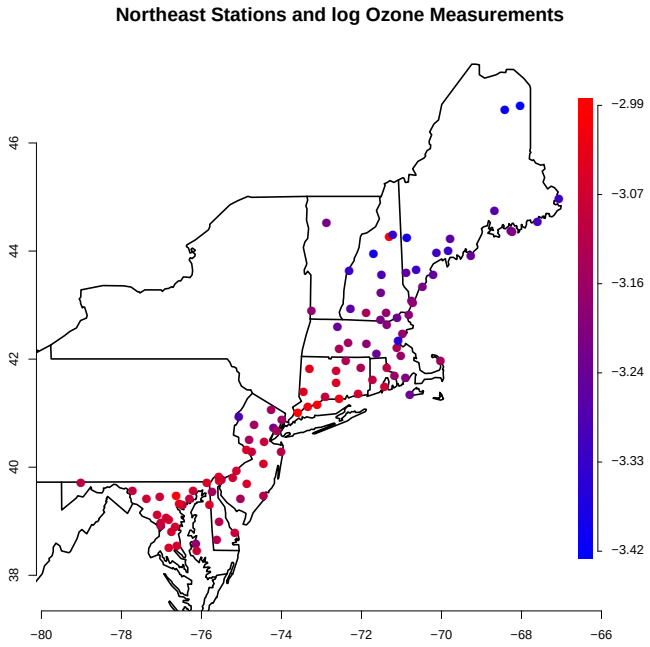

Figure 1: Measurement locations of (log) ozone readings. Altitude not shown.

2.2 Methods

To analyze this data we will first explore possible trends. This can be achieved by fitting a classic linear regression with a variable selection method. We select variables from linear, quadratic, and interaction terms based on BIC. The chosen variables are then used as part of the mean in our GP. We will validate our model by looking at DIC.

In determining an appropriate correlation function, we can look at possibly anisotropies in the data by binning distances to estimate semivariances. Here, we will analyze the residuals from our chosen linear model. This is discussed more in section 3.

We fit our model on a training data set of 96 locations, leaving about 10% out. If our model is a good fit, we should see reasonable predictions for the 11 locations from the test data set.

3 Exploratory data analysis

3.1 Trends

We explore possible trends in the data. Before we get to that, we make two adjustments to the data. First, we take the log of the ozone measurements to get closer to normality. Second, since we have a couple of observations with really high altitude, we add one to the altitude and take the log. This deals with the stations at 0 altitude and reduces the effect of possible leverage points.

Figure 2 hints at the possibility of some linear trends along some of the spatial dimensions. We fit a regression with linear and quadratic terms in each direction as well as interactions. Covariates are selected with a step-wise procedure with BIC as its selection criterion.

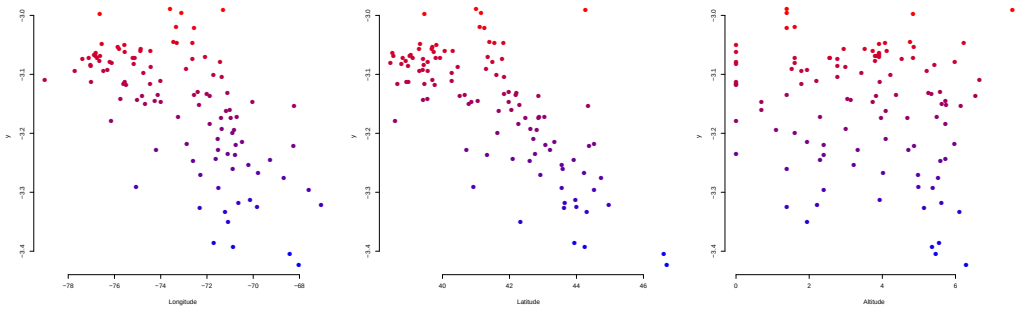

Figure 2: Each of three location types versus log ozone. The gray dots are the observed data, the black dots are the regression fit, and the liness represent the residuals.

The step-wise procedure selects an intercept, longitude, latitude, and an interaction between longitude and latitude. Altitude does not appear to have an influence on ozone.

3.2 Semivariograms

Using the residuals we compute a binned semivariogram from the data. We also look at the semivariogram in four directions to see if there is an anisotropy. The plots are given in Figure 3. We see from the right plot in Figure 3 that each direction produces more-or-less the same semivariogram, there is little difference between the curves. This suggests that we do not have enough evidence to believe there is anisotropy.

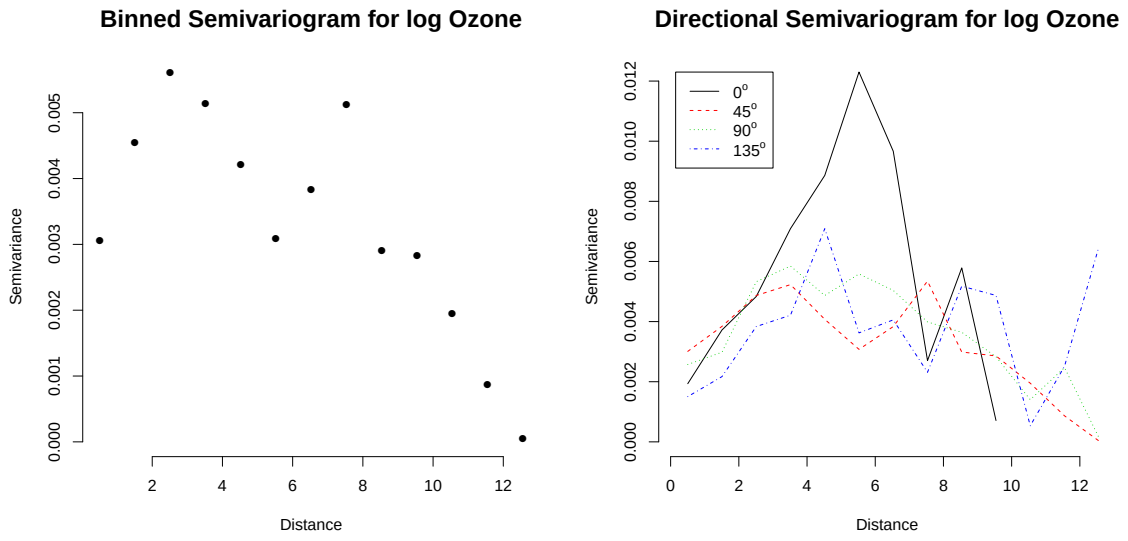

Figure 3: Omnidirectional semivariogram (left) and directional semivariograms (right).

Some possible issues with this include how we are calculating our distances. We use euclidean distance on longitude and latitude, but this is likely to be a poor or inconsistent measurement of distance given the large geographic area within which we are working.

Another potential problem is that our omnidirectional semivariogram (left, Figure 3) seems to be decreasing. This could be because of a really small effective range, which is itself pretty disconcerting. A very small ϕ would essentially mean that there is little to no spatial correlation between observations.

We do a least squares fit to the omnidirectional semivariogram for parameters in the Matérn correlation. The function we are minimizing is

$$f(u; \sigma^2, \phi) = \sum_{i=1}^{n} \left[\sigma^2(1 - \rho(u_i, \phi, \nu)) - y_i\right]^2$$

where u_i is the distance obtain from the omnidirectional semivariogram, y_i is the empirical semivariance, ϕ is the range, ν is the smoothness, ρ is the Matérn correlation function, and σ^2 is related to the sill. We fix ν to be 0.5, 1, 1.5, and 2.5 and estimate the remaining parameters by maximizing f. The estimates for the parameters are shown in top right corners of the plots in Figure 4. From these, we do see a very small ϕ.

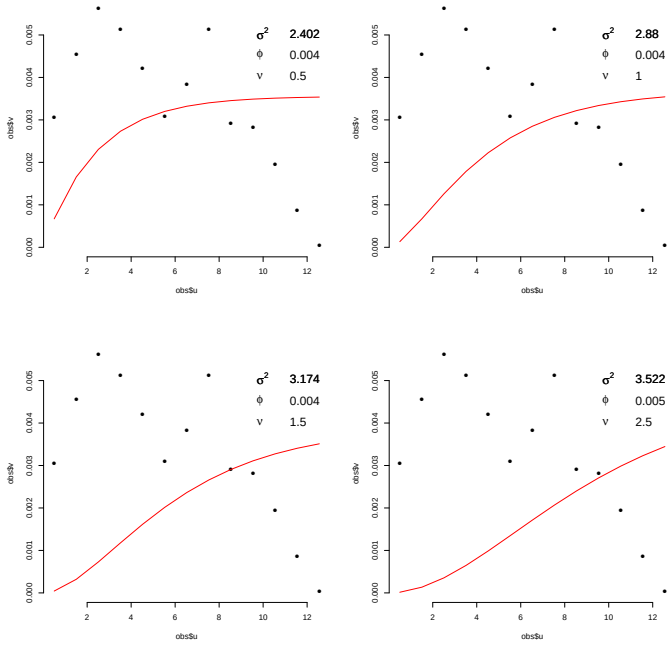

Figure 4: LSE for the Matérn correlation.

A nugget effect was not included because this lead to a $\phi = 0$, which I didn't want. This was because I was insistent on fitting a GP, but with $\phi = 0$ there is no need for a GP since all the covariances will be zero. That, or I spent a lot of time doing and understanding the wrong thing. Anyway.

4 Model

The model we fit is given by

$$\begin{aligned} X &= \mu + \epsilon, \qquad \epsilon \sim N(0, \tau^2 I) \\ \mu &= D\beta + v, \qquad v \sim N(0, \sigma^2 R(\psi)) \end{aligned}$$

where $D = D(s)$ makes up the design matrix at each location $s \in S$. In our case, D has a column of ones (intercept), the longitudes, latitudes, and interactions terms. The matrix $R(\psi)$ is computed using the Matérn correlation function. We are left with

$$X \sim N(D\beta, \tau^2 I + \sigma^2 R(\psi)).$$

We let $\gamma^2 = \tau^2/\sigma^2$, leading to

$$\begin{aligned} X &\sim N(D\beta, \tau^2(I + 1/\gamma^2 R(\psi))). \\ X &\sim N(D\beta, \tau^2 K(\psi)). \end{aligned}$$

where γ^2 is absorbed into the vector $\psi = (\phi, \nu, \gamma^2)$. For simplicity, we write $K(\psi) = K$.

The previous result leads to the following likelihood

$$L(\beta, \tau^2, \psi) \propto |K|^{-1/2}(\tau^2)^{-n/2} \exp\left\{-\frac{1}{2\tau^2}(X - D\beta)^\top K^{-1}(X - D\beta)\right\}.$$

We obtain posterior samples using the blocking strategy described on pages 6 and 7 of the slides on Bayesian inference. From these, we have

$$\begin{aligned}
\beta|\tau^2, \psi, X &\sim N(\hat{\beta}, \tau^2 D^\top K^{-1} D) \\
\tau^2|\psi, X &\sim IG(a + (n-k)/2, b + S^2/2) \\
p(\psi|X) &\propto |K|^{-1/2}|D^\top K^{-1} D|^{-1/2}(S^2 + 2b)^{-\frac{n-k}{2}-a} p(\psi)
\end{aligned}$$

where $\hat{\beta}$ is the solution to $D^\top K^{-1} D\beta = D^\top K^{-1} X$ and $S^2 = (X - D\hat{\beta})^\top K^{-1}(X - D\hat{\beta})$. I expect τ^2 to be small, so I let $a = 3$ and $b = 2$, which provides an inverse gamma with finite mean and variance, centered around 1.

For ψ, we have $p(\psi) = p(\phi, \nu, \gamma^2) \propto p(\phi)p(\gamma^2)$, where ν is constrained to be one of the values in $\{0.5, 1, 1.5, 2.5, 3.5\}$. We let ϕ and γ^2 have gamma distributions each with mean and variance 1, as these are also expected to be small.

5 Results

Summaries of the posterior parameters are given in Tables 1 and 2. Under the specified model, we have a DIC of -183.326. It is at this point we suspect either our model or our sampler. We were expecting a much smaller ϕ, but our posterior results show otherwise. In fact, the posteriors for ϕ and γ^2 were very comparable to their priors, under a range of priors. This was an area of great concern.

	Mean	2.5%	97.5%
β_0 (intercept)	8.212	-45.068	63.040
β_1 (longitude)	0.132	-0.595	0.875
β_2 (latitude)	-0.264	-1.549	0.991
β_3 (lon $\times$ lat)	-0.003	-0.020	0.014
τ^2	0.045	0.034	0.060
ϕ	2.937	0.956	6.277
γ^2	1.736	0.195	4.916

Table 1: Posterior parameter means and 95% interval bounds (the full model).

Each β_i has 95% credible bounds that contain zero, an indication that we are overparametrizing the model. When we fit the model to the intercept only, we obtain a DIC of -203.058 which is substantially better than the first model.

ν	0.5	1	1.5	2.5	3.5
Prob	0.002	0.036	0.116	0.321	0.522

Table 2: Posterior probabilities (bottom) of ν being a specific value (top).

The posteriors for the smoothness ν are shown in Table 2. The increasing behavior seemed to suggest that we have a very smooth process, but this would seem to contradict the hypothesis that $\phi \approx 0$. The model was fit to include higher ν, but everything above 3.5 had about the same posterior mass.

Posterior predictions based on the full model for the hold out sample are given in Figure 5. Again, this brings us to another concern with the model. Our predictive variances are much too large to be realistic. The only consolation to all of this is that our predictive means are similar to the observed values, but this doesn't seem to be a result of a stellar Gaussian process.

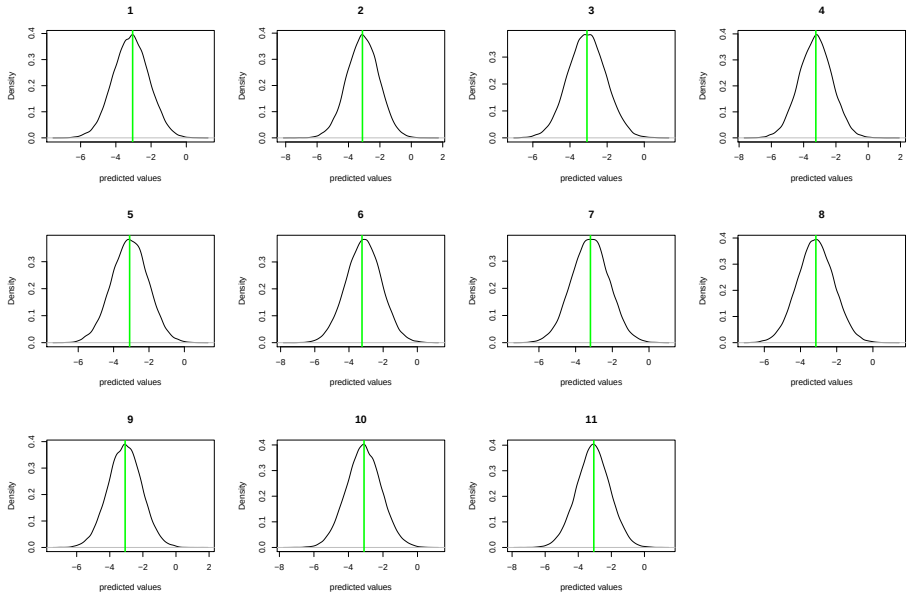

Figure 5: Predictive distributions for the hold out sample. The green lines mark observed data.

6 Conclusions

Clearly, some more work is needed to properly analyze this data set. I am very much in doubt that we should observe posterior estimates as the ones we obtained. This could be due to a poor model specification (very likely) as well as some bug in my code (also likely).