ST 533 Spatial Statistics Exam 1

Nate Wiecha

Analysis of PM2.5 Data During COVID-19

Introduction

COVID-19 has had many impacts, including, possibly, on air pollution as a result of the pandemic. As our outcome variable was air pollution, I analyzed whether COVID-19 was related to the outcome. I analyzed the change in mean PM2.5 levels between April – June 2019, and April – June 2020, in order to determine whether the prevalence of COVID-19 in a county is associated with any change in particulate matter levels between these two time periods.

Data

My primary data source is the publicly available EPA daily air quality data. These data have PM2.5 measurements at sites throughout the United States. This analysis focuses on the Southeastern U.S.: Florida, Georgia, North Carolina, South Carolina, and Virginia. I used data for two time periods: April – June 2019, and April – June 2020. At each station for which there were at least ten observations in both time periods (which was all stations), I averaged the measurements from each of the two three-month periods to obtain a 2019 average and a 2020 average at each site. Then I took the difference between the two averages to obtain the change between average PM2.5 levels from 2019 to 2020 at each site.

I hypothesized that the change in air quality might be influenced by the prevalence of COVID-19 cases around the monitoring sites. If the area of a site was strongly affected by COVID-19, it is plausible that lockdowns or lessening of travel, industry, and other activity could cause less pollution, thus reducing the PM2.5 relative to a similar time in the previous year.

I used the COVID-19 cases per U.S. county, cumulative as of June 30, 2020, compiled by the New York Times, and the estimated county populations as of July 1, 2019 from the US Census. The covariate was the county's cases divided by the county's population times 10,000, then logged, since there was a very wide range of values before taking the log. A limitation is that 2020 county-level population estimates were not available, which likely means that my estimate of cases per 10,000 population is a slight overestimate. This is not likely to greatly influence the regression results as it is unlikely for counties to experience major changes in population, relative to each other, from July 2019 – July 2020.

Figure 1 provides a graphical display of counties' prevalence of COVID-19, as well as the monitoring sites' locations and change in PM2.5 from 2019 to 2020.

Table 1 presents descriptive statistics aggregated over the monitoring sites. Note that for the COVID-related variables, since these variables are originally at the county-level, but analyzed for each site, this means that some counties are included multiple times, if multiple sites are in the same county, while counties with no monitoring sites are not included in the summary.

Variable	Min	1st Quartile	Median	3rd Quartile	Max
Mean PM2.5 Change	-3.05	-1.23	-0.76	0.10	3.00
COVID cases per 10K Pop (County of site)	2.39	45.31	65.91	88.45	231.49
Log COVID Cases per 10k Pop (County of site)	0.87	3.81	4.19	4.48	5.44

Table 1. Descriptive statistics for monitoring sites

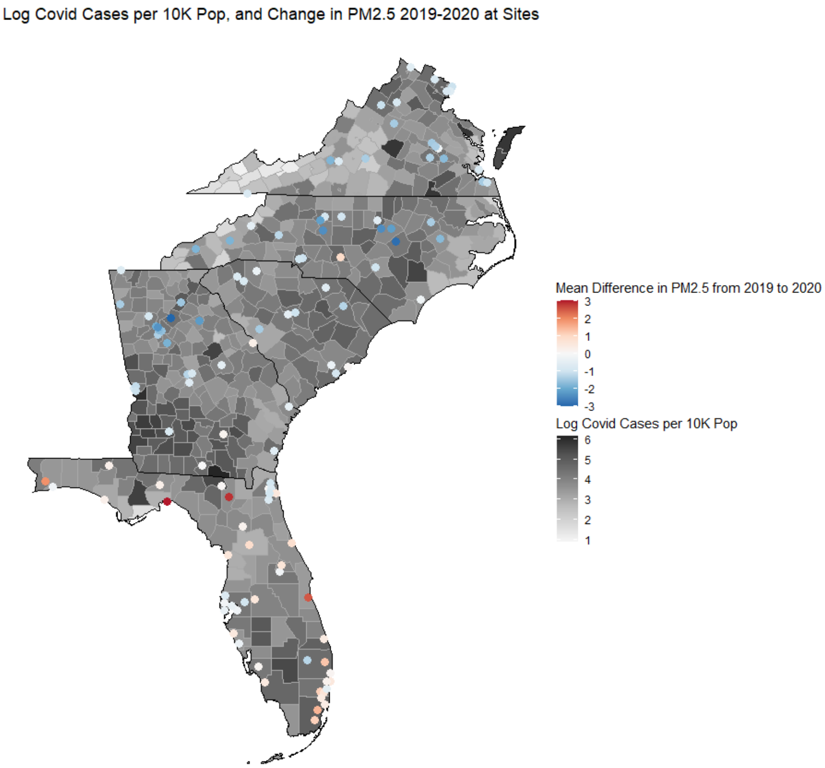

Log Covid Cases per 10K Pop, and Change in PM2.5 2019-2020 at Sites
Mean Difference in PM2.5 from 2019 to 2020
3
2
1
0
-1
-2
-3
Log Covid Cases per 10K Pop
6
5
4
3
2
1

Figure 1 Log COVID-19 cases/10,000 population for each county, and change in PM2.5 from 2019-2020 at each monitoring site

Methods

I examined whether COVID-19 prevalence in a county predicted change in PM2.5 from 2019 to 2020 by using a geostatistical linear regression model. The model takes the general form

$$Y_i = \mu_i + Z_i + \varepsilon_i$$

Where μ_i is a linear combination of the predictors, Z_i is spatial covariance, and ε_i is unexplained other variance.

The model for the mean is:

$$\begin{aligned}\mu_i = \beta_0 + \beta_1(longitude) + \beta_2(longitude^2) + \beta_3(latitude) + \beta_4(latitude^2) \\ + \beta_5(longitude)(latitude) + \beta_6(\log(Cases\ per\ 10{,}000\ pop))\end{aligned}$$

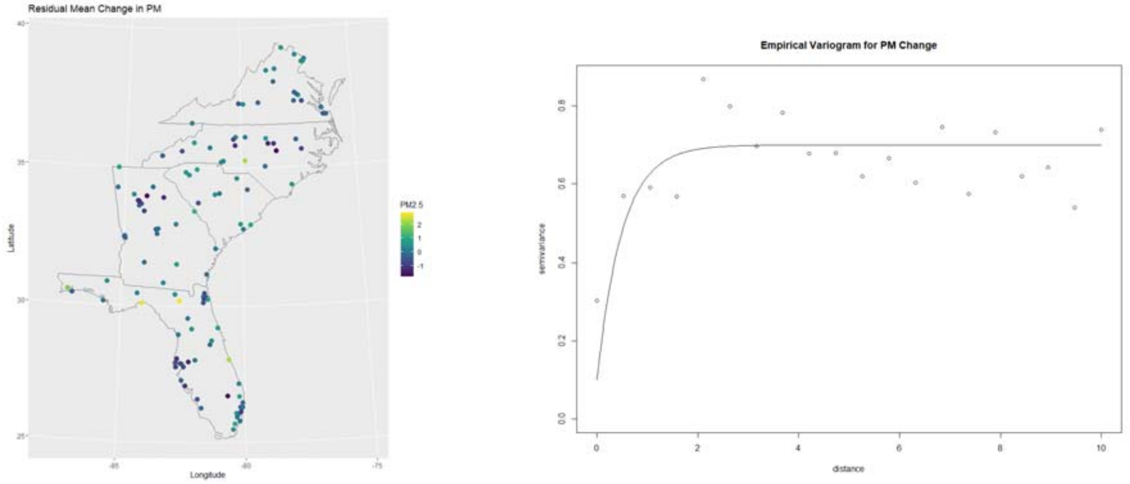
The two most obvious choices for the spatial covariance models are exponential and Matern. I tried both. I assumed isotropy in both cases. There is no obvious particular angle of spatial dependence visible in the plots.

Exponential covariance model: $Cov(Z_i, Z_i) = \sigma^2 \exp\{\frac{-d_{ij}}{\phi}\}$

Matern covariance model: $Cov(Z_i, Z_i) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{d_{ij}}{\phi}\right)^{\nu} K_{\nu}\left(\sqrt{2\nu}\frac{d_{ij}}{\phi}\right)$

Model Comparisons

As exploratory analysis, I fit the linear model for the mean, μ_i, and examined the residuals, and used a variogram to visually examine the spatial dependence. From the map of the residuals, there are some clusters of similar points. These clusters do not appear to be oriented in a particular angle, so there does not appear to be visual evidence of anisotropy. There also does not appear to be a different spatial relationship in different parts of the map, so stationarity appears to be a reasonable assumption.



Figures 2, 3: Residual PM change after fitting a linear model (Left). Empirical variogram with exponential covariance function overlaid (parameters picked manually) (Right).

An exponential function fits well. The variogram indicates a potential nugget, but as the covariance is decreasing rapidly as distance approaches zero, it is not entirely clear from the variogram whether a nugget is appropriate. I fit four models, trying both exponential and Matern covariance functions, with and without a nugget. I used AIC, BIC, and cross-validation metrics of mean-squared error (MSE) and coverage of the prediction intervals, using 12 cross-validation folds, to assess model fit. The number of folds was the lowest number that did not cause problems with the numerical optimization to obtain the MLE. These problems may have occurred due to different sites having the same county-level covariate values, as well as similar PM2.5 measurements if the sites were close together—i.e., several sites may have had very similar values for all variables, which at certain fold sizes may cause problems. The results of the models are summarized in Table 2.

	AIC	BIC	CV MSE	CV Coverage
Exp w/ nugget	295.72	323.56	0.709	90.00%
Exp w/o nugget	293.75	318.83	0.704	90.00%
Matern w/ nugget	297.52	328.18	0.749	85.83%
Matern w/o nugget	295.52	323.39	0.749	85.83%

Table 2. Model summary statistics

The exponential covariance without nugget fit best by all metrics, so that is the model I chose to use in my final analysis. In this model 90% of cross-validation 95% prediction intervals included the true value and the mean-squared cross-validation error was .704.

Model Checking

The main assumption I made for this model is that the spatial covariance can be modeled using an isotropic exponential function. The exponential function seemed to fit the empirical variogram well, so that shows the assumption of an exponential function is not unreasonable. To check isotropy, I also conducted a directional variogram as shown below.

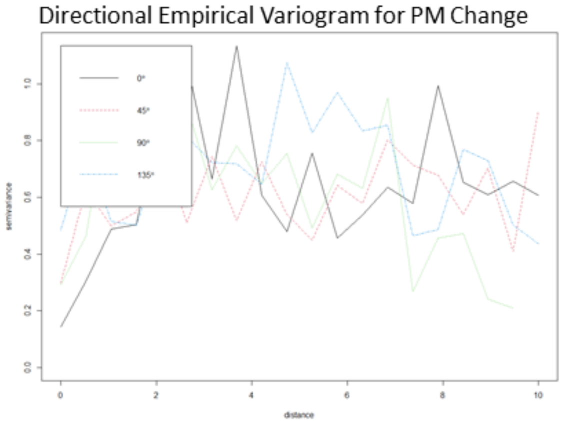


Figure 3 Directional Empirical Variogram

The variogram does not appear particularly different for any angles, confirming that isotropy is a reasonable assumption.

Final Model Summary

The final model, using an exponential covariance function and no nugget, was estimated by maximum likelihood using the geoR package. The parameter estimates and statistics are below.

Parameter	Estimate	Standard Error	Z	P-value
Intercept	-135.30	169.05	-0.80	0.4235
Longitude	-2.22	3.57	-0.62	0.5331
Latitude	2.96	1.91	1.55	0.1216
Longitude^2	-0.06	0.02	-0.31	0.7584
Latitude^2	-0.00	0.01	-0.03	0.9722
Longitude·Latitude	0.04	0.02	1.86	0.0631*
Log(COVID Cases/10K Pop)	-0.23	0.14	-1.70	0.0900*
Partial Sill	0.692			
Range parameter	0.232			

Table 3. Final model parameter estimate summary (* = statistically significant at the 10% level)

None of the coefficients were statistically significant at the 5% level. In particular, the prevalence of COVID-19 by county did not have a statistically significant association with the change in PM2.5 from 2019 to 2020 at the 5% level, although it did at the 10% level. The latitude and longitude variables could be suffering from collinearity, inflating their p-values even though they likely are contributing to the predictive accuracy.

Interpretations of coefficients:

- Longitude, Longitude squared, and longitude times latitude:
    - Holding latitude constant, a one-unit change in longitude from lon_1 to $lon_2=(lon_1+1)$ changes the mean PM2.5 change by $-2.22 - 0.06(lon_2^2 – lon_1^2) + 0.04(lat)$.
- Latitude, latitude squared, and longitude times latitude:
    - Holding longitude constant, a one-unit change in latitude from lat_1 to $lat_2=(lat_1+1)$ changes the mean PM2.5 change by $2.96 + 0.04(long)$.
- Note that the interaction term between latitude and longitude dominates the effect of any of the other latitude or longitude coefficients as it multiplies either latitude which is between approximately 25 to 40, or longitude which is between approximately -90 to -75.
- An increase by ten percent of COVID-19 case per 10,000 county residents decreases the average PM2.5 change by $(.23)\log(1.1) \approx 0.022$.
- Partial Sill: The total variation in the response is 0.692 units of PM2.5
- Outside a distance of 0.696 units, the observations are no longer correlated.

Spatial Prediction

I conducted prediction using Kriging, using the final model. The results are illustrated in Figures 5, 6, and 7.

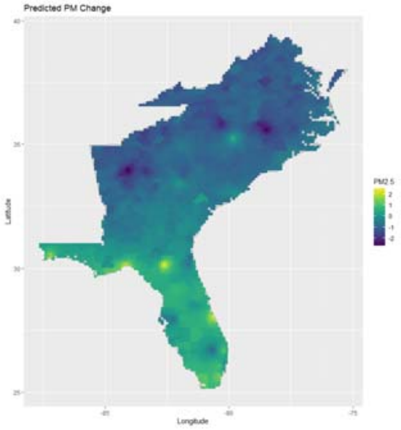


Figure 4 Predicted PM change via Kriging

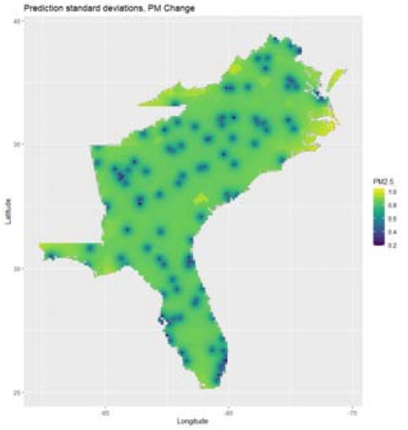


Figure 5 Prediction standard deviations

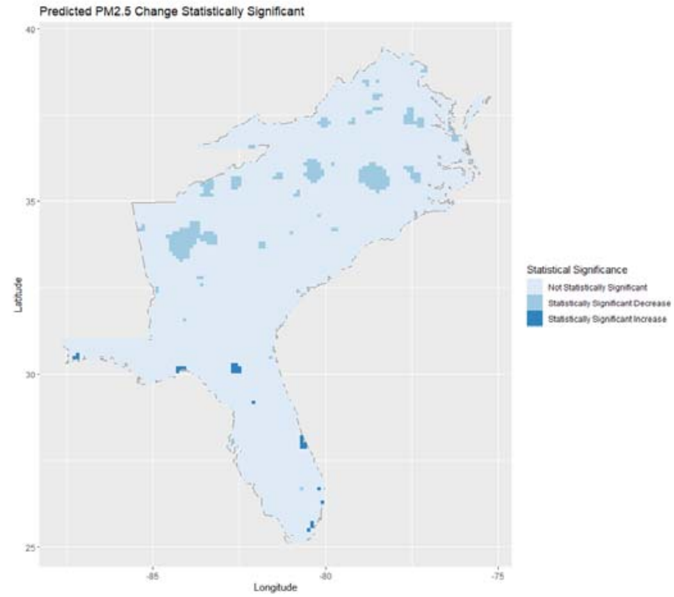

Predicted PM2.5 Change Statistically Significant
40
35
30
25
Latitude
-85
-80
-75
Longitude
Statistical Significance
Not Statistically Significant
Statistically Significant Decrease
Statistically Significant Increase

Figure 6 Locations where the prediction is statistically significant.

Figure 7 indicates where the predicted change was statistically significant, and was produced by determining where the prediction interval for PM2.5 change excluded 0 at the 95% confidence level.

Conclusions

I find some evidence, although not particularly strong, that prevalence of COVID-19 in counties is associated with change in PM2.5 from 2019 to 2020 at air quality monitoring sites in those counties. Several areas of the southeastern United States likely had decreases in PM2.5 concentration from 2019 to 2020, during April – June, based on Kriging predictions. Some smaller areas likely had increases in PM2.5 levels. More research might be able to identify other covariates that also influenced the change in PM2.5 and perhaps better isolate the association, if any, with prevalence of COVID-19.