The Effect of COVID-19 on Air Quality

Richard Watson

September 23, 2020

1 Introduction

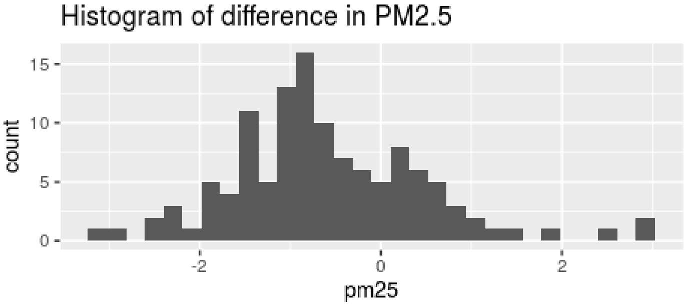
The COVID-19 pandemic has done caused many changes in the US and worldwide: economic decline, massive unemployment, overwhelming healthcare needs, and improved air quality? Recent articles are reporting on findings that suggest that lockdowns due to COVID-19 are related to the improved air quality. The quality of air is largely described in terms of the lack of pollutants that are harmful to humans and the environment. These pollutants are mostly released by human activity, such as industrial work and driving. The claim being suggested by recent articles is that the reduction of human activity due to lockdowns may have led to a reduction in harmful pollutants and improved air quality. The objective of this report is to investigate whether or not pollutants have actually decreased and if so, find out what factors can be used to predict this decrease accurately. Specifically, this report will zoom in on the Southeastern US and use data regarding PM2.5, a particularly dangerous variety of particulate matter that is aptly named for its incredibly small size of 2.5 micrometers, from April to June in 2019 and 2020 to check for significant differences and investigate factors that can be used to model this difference.

2 The Data

The data was collected from a multitude of sources in an attempt to generate a model that could successfully describe the trends in the response and then be used for prediction.

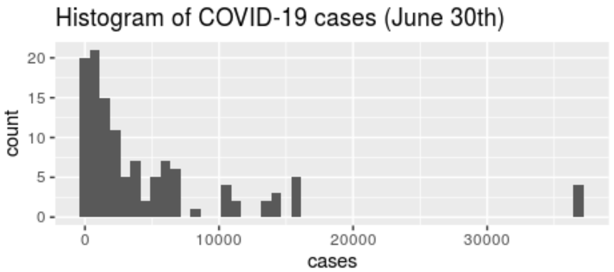
2.1 The Response

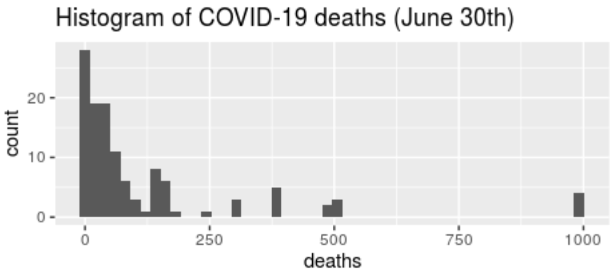
The response variable used in this analysis will be the difference in the average concentration of PM2.5 for April-June 2019 and April-June 2020, that is to say $Y(s) = \bar{Y}_{2020}(s) - \bar{Y}_{2019}(s)$, in Virginia, North and South Carolina, Georgia, and Florida. Only collection sites with more than 10 samples for the year were used in the analysis, which after filtering is 120 sites.

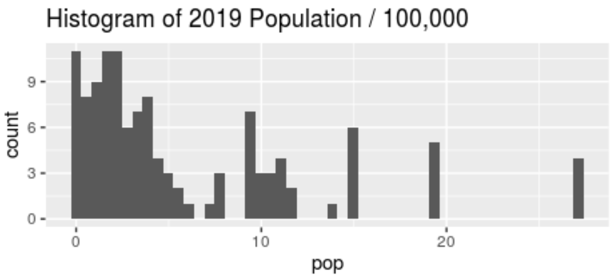


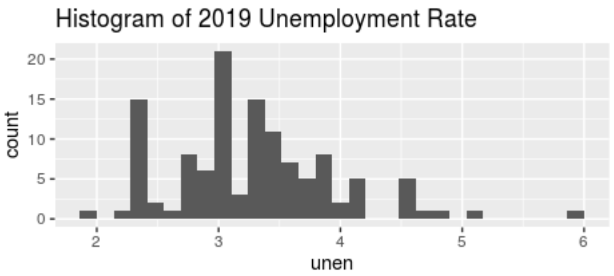
2.2 The Covariates

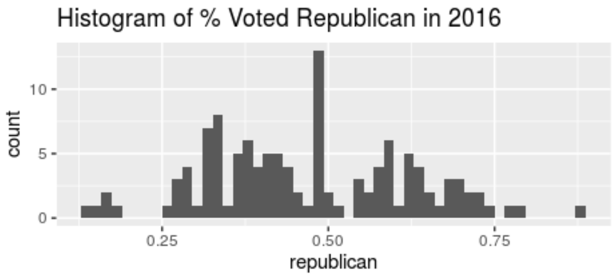
Other than the latitude and longitude information, 6 other variables were used to try to describe the response. The first two try to capture the affect of COVID-19 during the early months of the pandemic. They are the total amount of deaths and cases reported on June 30th, 2020. Next, the population on a county level divided by 100,000 estimated in 2019. We also have the unemployment rate for 2019 as a implicit measure of human activity; areas with more unemployment will likely have a smaller reduction in human activity than areas with less unemployment. Finally we have the percentage of registered voters that voted for the 2016 Republican presidential nominee and the Democratic presidential nominee. This can be seen as a summary estimate for political affiliation and other demographic information.

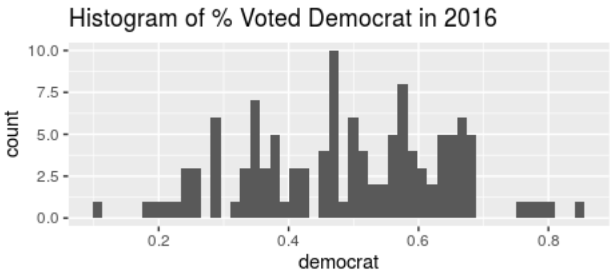












3 The Models

Four models were created in attempt to describe and predict the response variable. Before describing each of the models, it is important to define terms that will used in the model definitions:

S: partial 2nd order spatial model ($S_i = \beta_0 + \beta_1 lat_i + \beta_2 lon_i + \beta_3 lat_i * lon_i$)

C: term for COVID-19 cases

D: term for COVID-19 deaths

P: term for 2019 population divided by 100,000

U: term for 2019 unemployment rate

B: term for percent voted Democrat in 2016 (Blue)

R: term for percent voted Republican in 2016 (Red)

The first two models attempt to describe the response using demographic information split across party lines. The idea here being that trends in demographics may describe the difference in human behavior before and during a lockdown scenario:

$$Y_{1i} = S_i + \beta_4 P_i + \beta_5 U_i + \beta_6 R_i + Z_i + \epsilon_i$$

$$Y_{2i} = S_i + \beta_4 P_i + \beta_5 U_i + \beta_6 B_i + Z_i + \epsilon_i$$

where the spatial covariance for both models is exponential with $\tau^2 = 0.3$, $\sigma^2 = 0.3$, and $\phi = 1.5$. The third model attempts to describe the response using COVID-19 data and population data:

$$Y_{3i} = S_i + \beta_4 C_i + \beta_5 D_i + \beta_6 P_i + Z_i + \epsilon_i$$

where the spatial covariance for both models is exponential with $\tau^2 = 0.3$, $\sigma^2 = 0.3$, and $\phi = 1.5$. The final model attempts to describe the response using all of the covariates, combining demographic information with COVID-19 information:

$$Y_{4i} = S_i + \beta_4 C_i + \beta_5 D_i + \beta_6 P_i + \beta_7 U_i + \beta_8 R_i + \beta_9 B_i + Z_i + \epsilon_i$$

where the spatial covariance for both models is exponential with $\tau^2 = 0.3$, $\sigma^2 = 0.275$, and $\phi = 1.5$.

4 The Best Model

To choose the best model, we perform a 5-fold cross validation test:

	AIC	BIC	MSE	COVERAGE
Model 1	289.56	317.43	0.68	92.50
Model 2	289.75	317.63	0.68	92.50
Model 3	292.22	320.10	0.64	94.17
Model 4	296.06	332.30	0.70	0.00

Table 1: Results of 5-fold Cross Validation

In terms of AIC and BIC, Model 1 performed the best. Model 3 has the best coverage and MSE, but losses to Model 1 in AIC and BIC. Since Cross Validated MSE is a more direct estimate of model accuracy than AIC and BIC and the difference in AIC and BIC between Model 1 and 3 is relatively small, Model 3 will be selected as the best model.

4.1 Assumptions of the Best Model

In the making of Model 3, we assumed that the response was normal and the spatial covariance could be modeled using an exponential fit. We also assume that the residuals are isotropic, but given the small amount of data points, this is difficult to check. The fit for the variogram is not that good and the histogram is not Uniform which suggests that the model is breaking our assumptions, but based on the plots, this violation is not egregious.

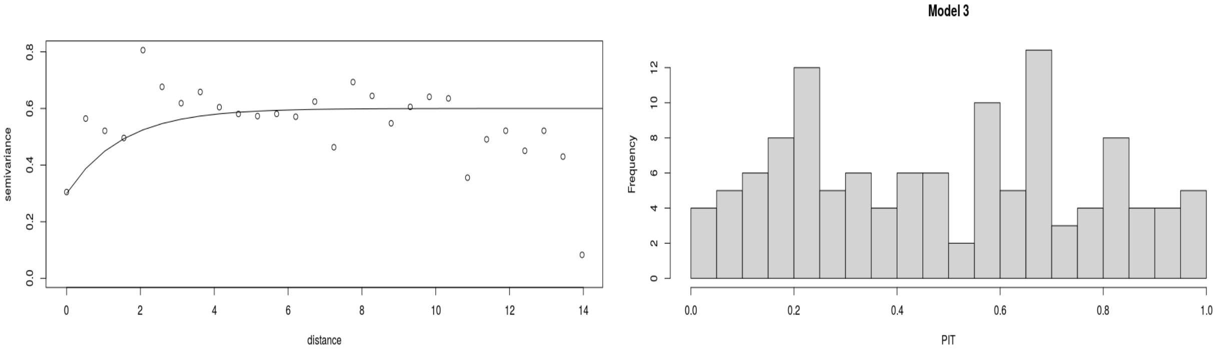


Figure 4: Variogram with fit(Right) and Histogram of probability integral transform statistics applied to Model 3

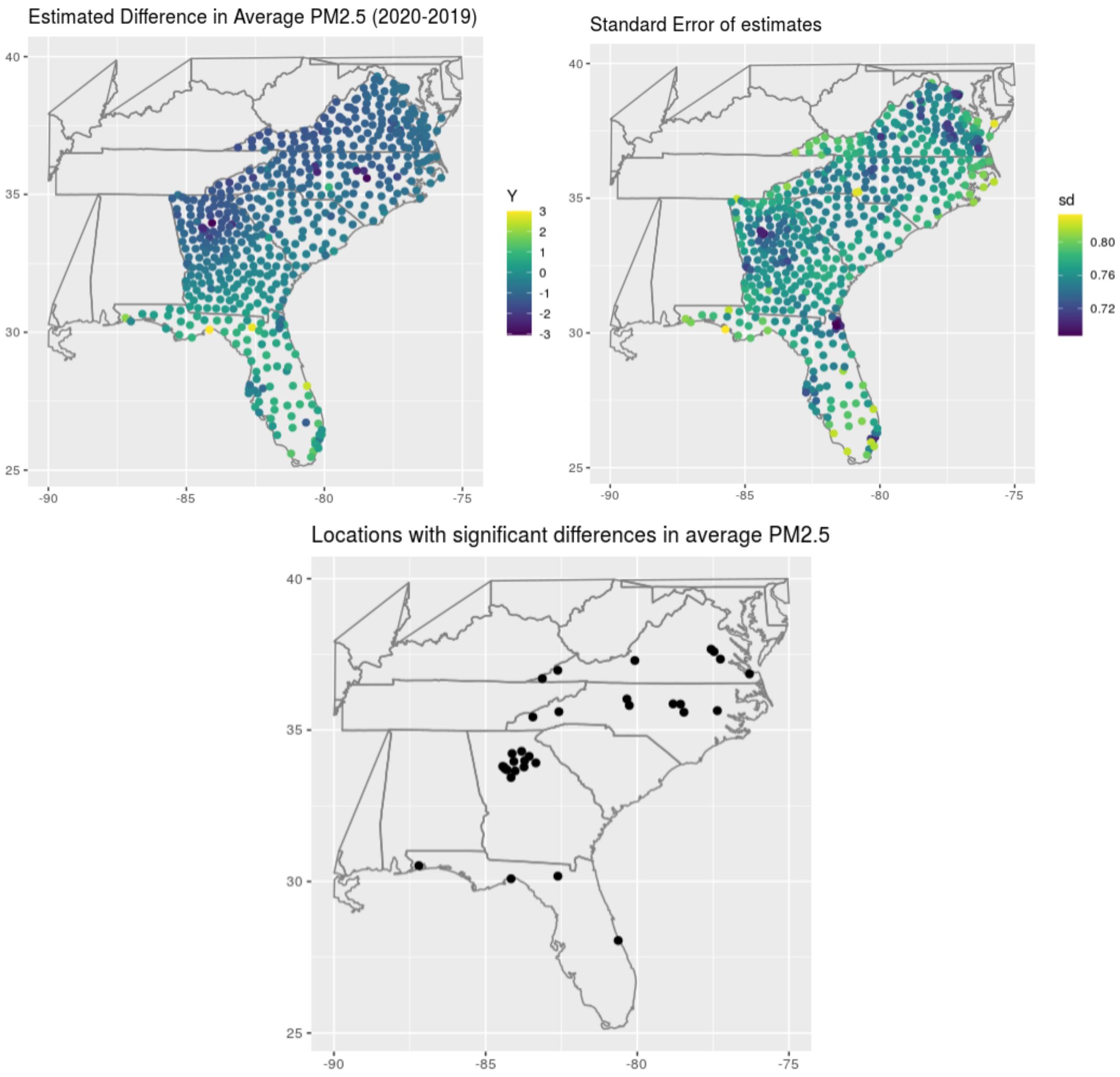
4.2 Parameter Estimates

Surprisingly, Model 3 is a reduced spatial population model in disguise. The parameter estimates below suggest that besides not being significant, the total COVID-19 cases and deaths have no effect on the response. All of the other parameters are either significant or fairly close to being considered significant. Moving on to interpretations, the intercept is irrelevant here as it represents the mean response at a latitude and longitude of 0 where there exists no one (literally as it is the middle of the Atlantic Ocean). The model suggests that as one moves northwards, westwards, northeastwards, the response will increase. Finally the estimate for population suggests that as population increases, the response decreases. This follows intuition in that more populous regions will tend to see a greater reduction in human activity and thus pollution when lockdowns are in effect than less populous regions.

	Estimate	Std error	Z	P value
Intercept	-71.5947	42.6112	-1.6802	0.0929
lat	2.2352	1.2472	1.7921	0.0731
lon	-0.9772	0.5260	-1.8580	0.0632
latlon	0.0306	0.0154	1.9798	0.0477
cases	0.0000	0.0001	0.8013	0.4230
deaths	0.0005	0.0013	0.4052	0.6854
pop	-0.0989	0.0392	-2.5239	0.0116

Table 2: Parameter estimates for Model 3

5 Results & Conclusions


Estimated Difference in Average PM2.5 (2020-2019)

Standard Error of estimates

Locations with significant differences in average PM2.5

Figure 5: Maps regarding estimates of Model 3

The model estimates that for the most part, the PM2.5 difference will be between 1 and -1. Taking into account the standard error of the estimates, this would mean that a majority of the locations do not have significant differences. In fact, only 32 locations were found to be significant. These locations are mostly very dense urban areas, such as Atlanta, Raleigh, Orlando, and Richmond, so the model agrees with the claim suggested by recent articles, but only for extremely urban areas.