Does the Lock-down Policy change the air quality in 2020 across the Southeastern US?

Hunter Jiang

1 Summary

In this paper, we use a spatial linear regression model to analyze the difference in the average of PM2.5 concentration between 2020 and 2019 at some observation sites located at Southeastern US. Then, we implement a Kriging model to predict the total trend. Interestingly, we find the change in average of PM2.5 concentration from weekdays, which is the results of the lock down policy, together with the location information can explain our response very well, but those from weekend does not.

2 Introduction

People have been working from home during the COVID-19 pandemic, and there is rumor that the air quality is getting better. In this paper, we keep our eyes on the air quality data from EPA about southeastern states (VA, NC, SC, GA, and FL) during April-June 2019 and 2020. We will try to answer (a) whether is a statistically significant change in the average of PM2.5 concentration, (b) if there is a difference, does such change spatial correlated, (c) what is the potential reason of this change (if it exist), and (d) Predict the difference in mean between 2019 and 2020 on these areas on a fine grid.

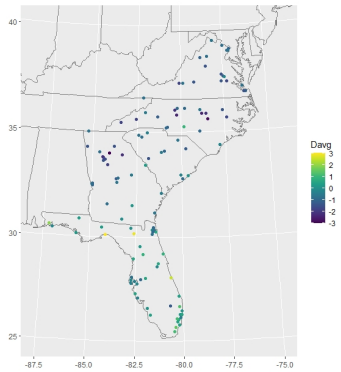

Figure 1: Difference in PM2.5 Concentration (2020 - 2019)

3 Data Description

The csv file downloaded form the EPS website[1] contains 20 columns of data, from which we picked out date, daily mean PM2.5 concentration, and site's information from the April-June. Then we merge data by site to get the response variable of interest, the difference in mean of PM2.5 concentration, and other potential covariates in our model.

Firstly, we include the **Longitude** and the **Latitude** of the observation site into our model. The reason we do this is a very strong trend in uneven pattern between different latitude (we can see from figure (1)).

The second covariate that we interested in is the **State** that observation site from. From figure (2) we can clearly find out that Florida has the smallest mean in the distribution plot, compared with other states. It could also be the state-wise reason which cannot be explained by the longitude and latitude only.

The third covariate of interest is the difference in the average of PM2.5 concentrations on **Weekdays** and **Weekends**. The idea of this may be ill-defined because we are using parts of the data to explain itself, but it can be found in figure (3) that the distribution of PM2.5 concentration has changed since 2019. Besides, we find a very strong and positive correlation between the difference in weekdays and difference in total, but not between weekends and total.

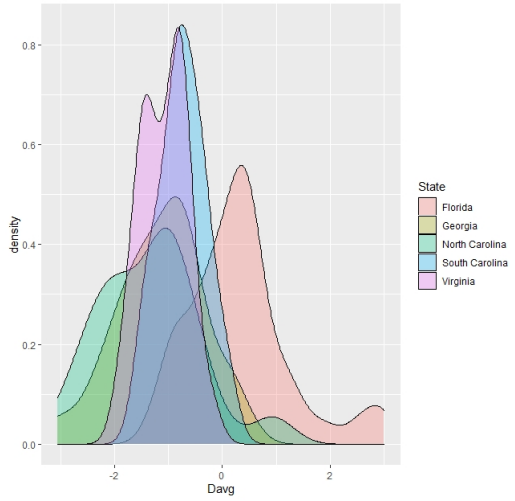

Figure 2: Difference by State

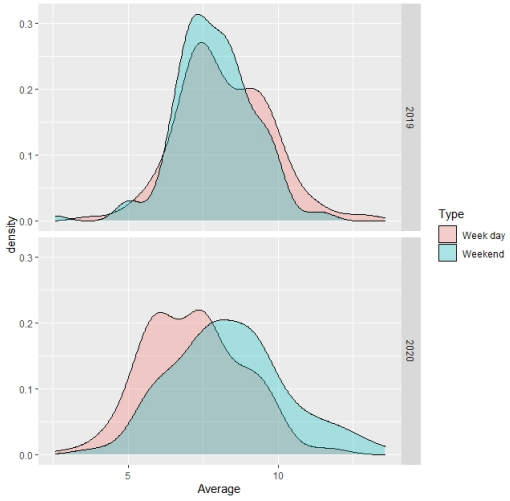

Figure 3: Difference by Weekday/end

[1]https://www.epa.gov/outdoor-air-quality-data/download-daily-data

4 Methods

There are two methods in our approach. We will first fit several Spatial Linear Model (SLM) with exponential (spatial) correlation, and then we implement the best model with Kriging to predict the difference in mean on a fine grid.

4.1 SLM with Exponential Correlation

Assume we already observed n data points with the response variable y, covariates $\boldsymbol{x}$, and location $\boldsymbol{s}$. For the data set $\boldsymbol{T} = \{\boldsymbol{Y}; \boldsymbol{X}, \boldsymbol{s}\} = \{(y_i; \boldsymbol{x}_i, \boldsymbol{s}_i)\}_{i=1}^n$, the basic model in this paper, spatial linear model, could be expressed as

$$Y_i = \beta_0 + \boldsymbol{X}_i\boldsymbol{\beta} + Z_i + \epsilon, \tag{1}$$

where Y_i is the random variable for response y_i, $\boldsymbol{X}_i$ is a random vector for covariates $\boldsymbol{x}_i$, Z_i is correlated to the location information $\boldsymbol{s}_i$, and $\epsilon \sim N(0, \sigma^2)$ is the error term.

The difference between the ordinary regression method and model (1) is the spatial correlated term, $\boldsymbol{Z}_i$. There are many spatial correlation functions available, and we choose the most common one, the exponential correlation, which can be expressed as:

$$\rho(d) = e^{-d/\phi} + c, \tag{2}$$

where $\phi > 0$ is the spatial range parameter, and $c \geq 0$ is the nugget parameter.

4.2 Simple Kriging in Prediction

For anther data set where we know some covariates correspond with their locations, says $\boldsymbol{R} = \{(\boldsymbol{x}_j, \boldsymbol{s}_j)\}_{j=1}^m$, and a pre-trained model on data set $\boldsymbol{T}$, we can predict the response $\boldsymbol{Y}_0$ on the new locations after defining $Cov(Y_0, Y_i) = \Sigma_0(\hat{\boldsymbol{\theta}})$ by

$$\hat{y_j} = \hat{\beta}_0 + \boldsymbol{x}_j\hat{\boldsymbol{\beta}} + \Sigma_0(\hat{\boldsymbol{\theta}})\Sigma_0(\hat{\boldsymbol{\theta}})^{-1}\{\boldsymbol{Y} - \hat{\beta}_0 - \boldsymbol{X}\hat{\beta}_1\}. \tag{3}$$

5 Model Comparisons

Denote the difference in mean as Y, the location at the point as $\boldsymbol{s} = (long, lat)$, difference in mean among weekdays and weekends as (dw, de), and the state factor as S. We have following 6 models as candidates.

$$Y = \beta_0 + \beta_1 long + \beta_2 lat + Z + \epsilon, \tag{4}$$

$$Y = \beta_0 + \beta_1 long + \beta_2 lat + \beta_3 dw + Z + \epsilon, \tag{5}$$

$$Y = \beta_0 + \beta_1 long + \beta_2 lat + \beta_3 de + Z + \epsilon, \tag{6}$$

$$Y = \beta_0 + \beta_1 long + \beta_2 lat + \beta_3 S + Z + \epsilon, \tag{7}$$

$$Y = \beta_0 + \beta_1 long + \beta_2 lat + \beta_3 dw + \beta_4 S + Z + \epsilon, \tag{8}$$

$$Y = \beta_0 + \beta_1 long + \beta_2 lat + \beta_3 de + \beta_4 S + Z + \epsilon. \tag{9}$$

For each model, we run two sub-model with and without the nugget term in the exponential correlation. Thus, we will finally have 12 candidates in total.

As for comparing method, we use standard 5-fold Cross Validation. The performance matrices are MSE= $||\hat{\boldsymbol{Y}} - \boldsymbol{Y}||^2$, MAE= $||\hat{\boldsymbol{Y}} - \boldsymbol{Y}||_1$, and COR= $corr(\hat{\boldsymbol{Y}} - \boldsymbol{Y})$. The best performance within all candidates are highlighted in bold font.

Model	Nugget	MSE	MAE	COR	Model	Nugget	MSE	MAE	COR
(4)	yes	0.68	0.61	0.64	(7)	yes	0.65	0.60	0.66
	no	0.67	0.60	0.65		no	0.65	0.59	0.66
(5)	yes	**0.06**	**0.20**	**0.97**	(8)	yes	0.07	0.21	**0.97**
	no	0.09	0.24	0.96		no	0.07	0.21	**0.97**
(6)	yes	0.23	0.37	0.89	(9)	yes	0.25	0.37	0.88
	no	0.28	0.40	0.87		no	0.28	0.40	0.87

Table 1: Model Performance Results

From table (1) we can find out that model (5) with a nugget term outperform others. The correlation between our self-train prediction and the true response is as high as 0.97, which means using such covariates can explain most information in the response variable. Practically, we may argue that the difference in average PM2.5 may indeed varies among different locations, and partially due to the lock down / work from home policy.

6 Model Checking

6.1 Variogram Analysis

Model (5) assumes that the variogram of the data is an exponential with evenly distributed parameters, so we need to check the trend of variograms to make sure about that. Figure (4) is the total variogram plot, with a fitted exponential model, which fits well to real variogram. Figure (5) is the sub-variogram of four regions in the South-East U.S., we use $long = -82$ and $lat = 33$ as the border of our sub-region. From this figure we can find that the semivariance is relatively even among four regions.

6.2 Data Imputation

Owing to the fact that we are going to predict the response variable in a finer grid, we need to do some data imputation also via simple Kriging method with $long$, lat, $long^2$, lat^2, and $long * lat$. Figure (6) shows the standard error plot of each locations. We can find that the absolute value of se is less than 1, which means that our model do make sense in this case.

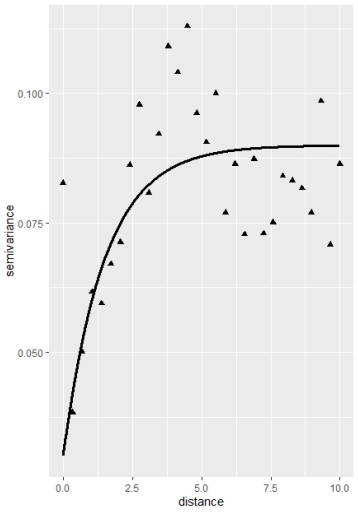

Figure 4: Total Variogram

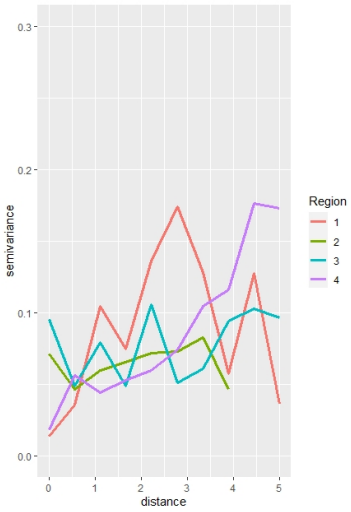

Figure 5: Sub-Variogram

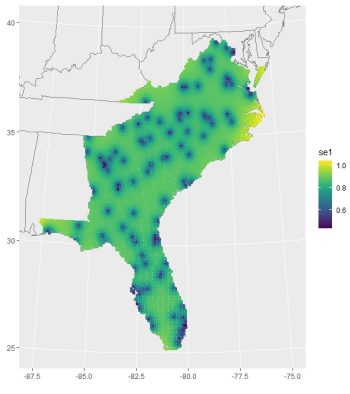

Figure 6: SE(Imputation)

7 The Final Model

$$Y = -6.86 - 0.08long + 0.0185lat + 0.9023dw + Z + \epsilon, \quad (10)$$

$$\rho(d) = e^{-d/3.146} + 0.0469. \quad (11)$$

When the (longitude, latitude, weekday diff average) change 1 unit, the response will change (-0.08, 0.019, 0.9023) unit with other fixed. When all covariates are 0, the average value of difference is -6.86. The spatial effect range is about $3\phi = 3 * 3.146$, the partial still is 0.048, and the estimated nugget effect is 0.0469.

8 Spatial Prediction

Figure (7) shows the difference in the average PM2.5 concentration between 2020 and 2019, with correspond standard error in figure (8). Then we calculate the confidence interval for each point by $\mu \pm 1.96se$, and check whether or not it contains 0 to determine the significance, which is showed in figure (9).

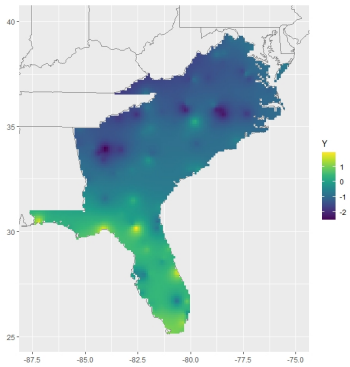

Figure 7: Prediction

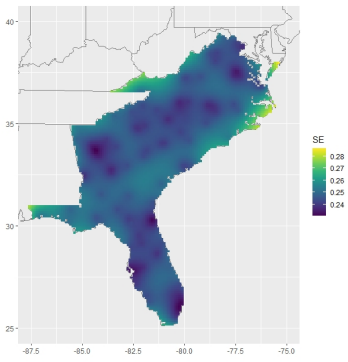

Figure 8: SE(Prediction)

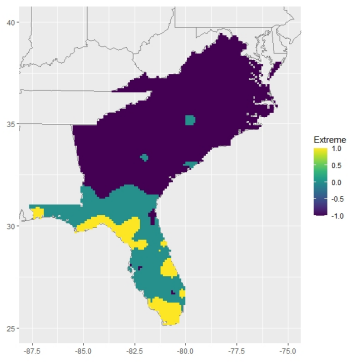

Figure 9: Significance

9 Conclusions

No page left, see Summary. Note that codes used in this paper could be downloaded from `https://github.com/HunterJiang97/PM2.5SouthEastenUS`.