Stat 605 Final Project Report

Yukun Fang(yfang67), Mengkun Chen(mchen373), Runze You(ryou3), Mengqi Li(mli653)

November 2020

1 Introduction

"In Sofia, air pollution norms were exceeded 70 times in the heating period from October 2017 to March 2018", citizens' initiative AirBG.info says. During the time, the air pollution has become a serious problem for Sofia, This problem attracted the researchers all around the world.

Air Quality Index has been used to determine the level of air pollution across different regions worldwide. As part of it, the level of particulate matter (PM) is measured as well. This is the term used for a mixture of solid particles and liquid droplets found in the air.

Sofia air quality data set describes daily five air-quality measurements-PM2.5, PM10, and meteorological statistics-pressure, temperature and humidity from July 2017 to June 2019.

For this dataset, we try to figure out whether there are some trends in PM2.5 and PM10 and whether there are some relationships between climate variables and air-quality measurements. We draw time series plots, conduct t-test, make short videos to answer these questions. We find that both meteorological measurements and air-quality measurements (PM2.5 and PM10) influenced by season and the air pollution problems are serious during the winter. Furthermore, there exist relationships between temperature, humidity and PM2.5, PM10.

2 Analysis

2.1 Data description

Meteorological measurements and air pollution measurements are recorded between July 2017 and June 2019. There are two kinds of datasets which contain the air quality data and meteorological data separately. We make **videos** to show the monthly change in PM10 and PM2.5 by longitude and latitude. Here are two geographical distribution maps in our videos [Link to Videos].

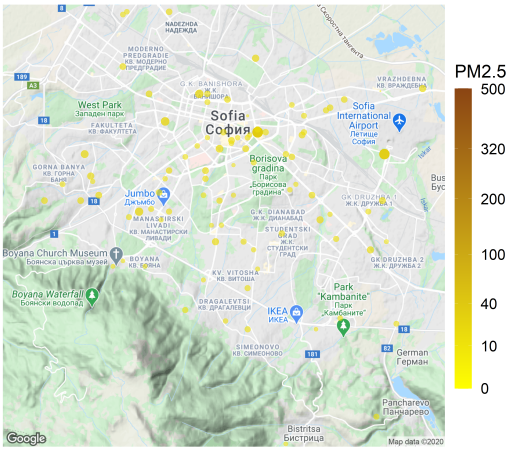

Figure 1: PM2.5 in June 2017

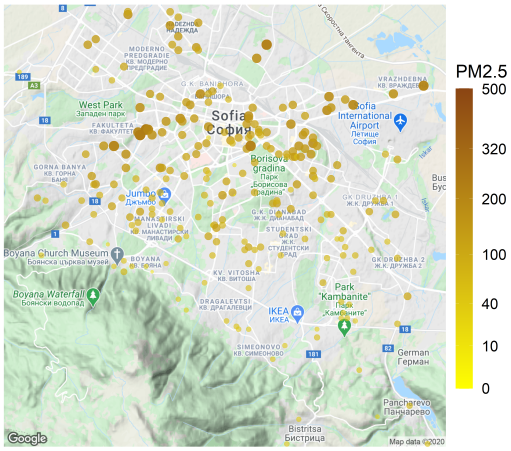

Figure 2: PM2.5 in January 2018

Animation shows that pollution is more serious in winter. From July 2017 to June 2019, areas with more residential communities or with large populations are more polluted than other areas. We noticed that the

vicinity of West Park, which locates in the west of Sofia, is quite polluted in the recorded time span. We will focus on this area to explain our findings.

- Variable Explanation

  sensor_id: sensor identification. Sensors record air pollution measurements and meteorological measurements around every five minutes.

  lat, lon: latitude and longitude coordinates, recording geographical information of each sensor. They are not 1-to-1 matches due to precision of the data.

  P1: PM2.5 concentration

  P2: PM10 concentration

  Pressure, Temperature, Humidity: meteorological measurements reported by sensors.

- Air Pollution Measurements

Table 1: Air Pollution Data Overview

	sensor_id	location	lat	lon	timestamp	P1	P2
1	2888	1453	42.71	23.28	2017-08-01T00:00:01	7.52	6.48
2	3641	1837	42.69	23.36	2017-08-01T00:00:02	10.70	6.10

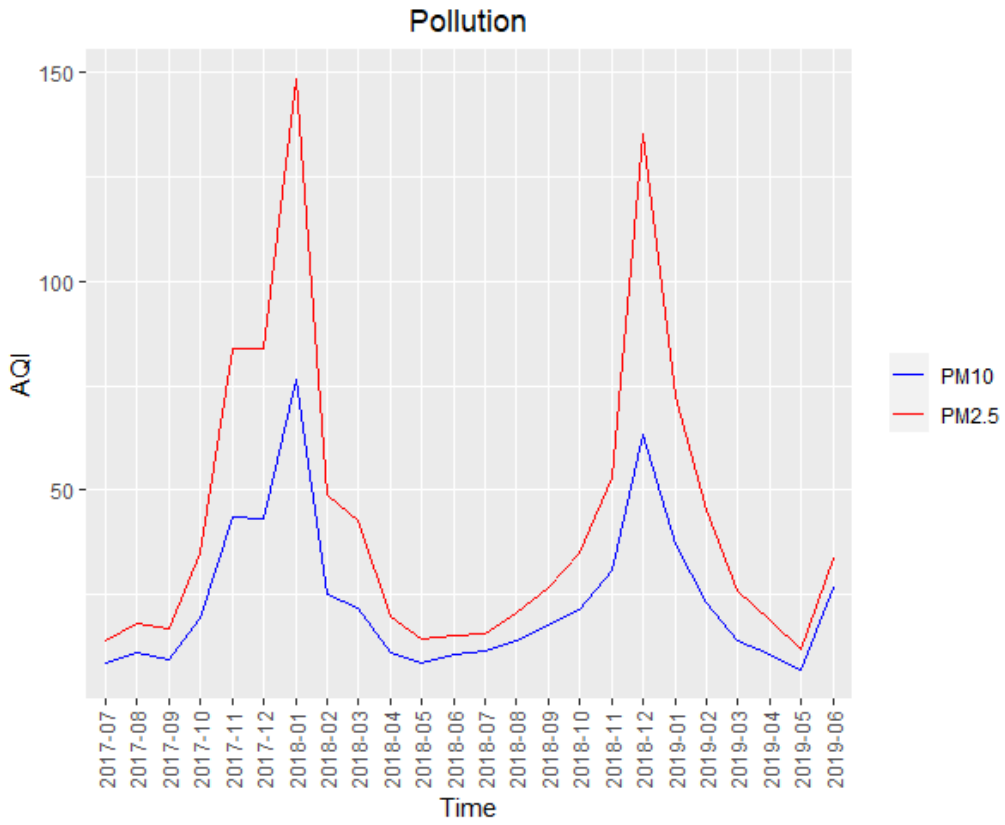

Figure 3: Time series plot for PM2.5 and PM10

Figure 3 indicates that both air pollution measurements have seasonal patterns. They go to the highest level at around January each year and drop to the lowest level in May. Concentration of PM10 is generally lower than that of PM2.5.

- Meteorological Data

Table 2: Meteorological Data Overview

	sensor_id	location	lat	lon	timestamp	pressure	temperature	humidity
1	2266	1140	42.74	23.27	2017-07-01T00:00:07	95270.27	23.46	62.48
2	2292	1154	42.66	23.27	2017-07-01T00:00:08	94355.83	23.06	59.46

We have draw the plots and put it in github and according to the data, pressure in the recorded time span does not show any seasonal pattern. However, humidity shows similar pattern with two air pollution measurements. Humidity reaches its peak at around December every year, This can probably be explained by the climate and topographical characteristics around. Since Sofia locates in northern hemisphere, the temperature is high in summer and low in winter.

2.2 Statistical analysis

- Box plots for air quality measurements

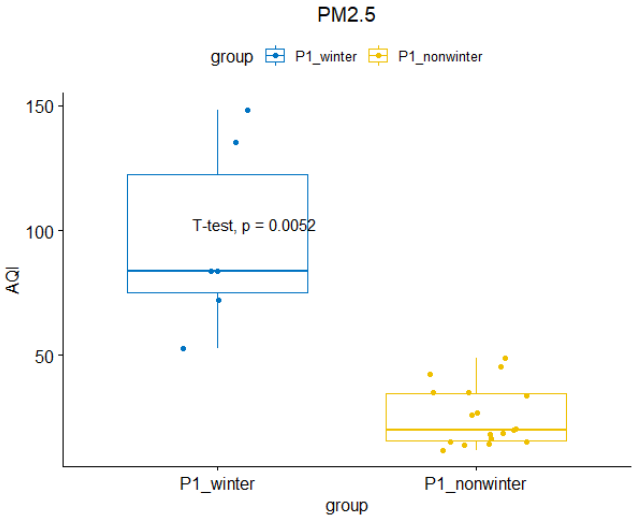

Figure 4: PM2.5

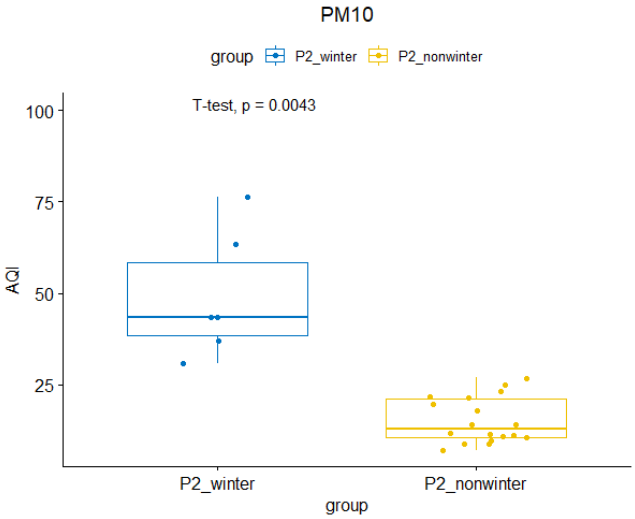

Figure 5: PM10

According to the box plots, we find that both PM2.5 and PM10 concentrations are higher in winter months, which are November, December and January.

- t-test results

Table 3: Test for difference between winter months and non-winter months

	winter mean	non-winter mean	t-statistics	df	p-value
PM2.5	96.0452	25.5388	4.5526	5.3341	0.0052
PM10	49.0558	15.2183	4.6905	5.4434	0.0043

From the test results, we can conclude that the pollution is heavier during the winter with great confidence.

- Correlation between air pollution measurements and meteorological measurements

  We also draw the correlation plot to see the relationship between air-pollution measurements and meteorological measurements.
  Based on the plot, the three stars mean they have great relationship. Clearly PM2.5 and PM10 are highly correlated with temperature and humidity.

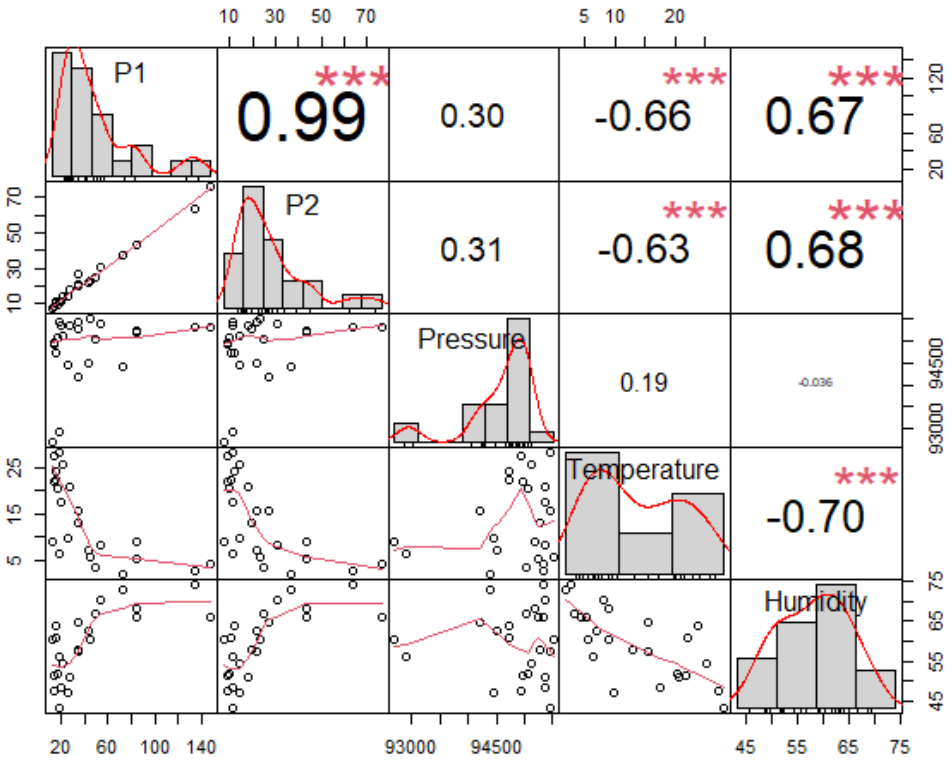

Figure 6: Correlation plot

3 Seasonal GARCH Model to Predict

Air pollution is a serious problem in Sofia during winter. We use times series methods to build the model and predict PM2.5 and P10.
As Figure 7 and Figure 8, scatter plots of each days records, shows, there exists seasonal influence for PM2.5 and PM10 with clear views at the increasing tendency in the second half of a year, and decreasing tendency in the first half of a year.
Besides, with our results getting from the pollution analysis from winter and non-winter, it is important to first reduce the seasonal effect to our time series analysis.

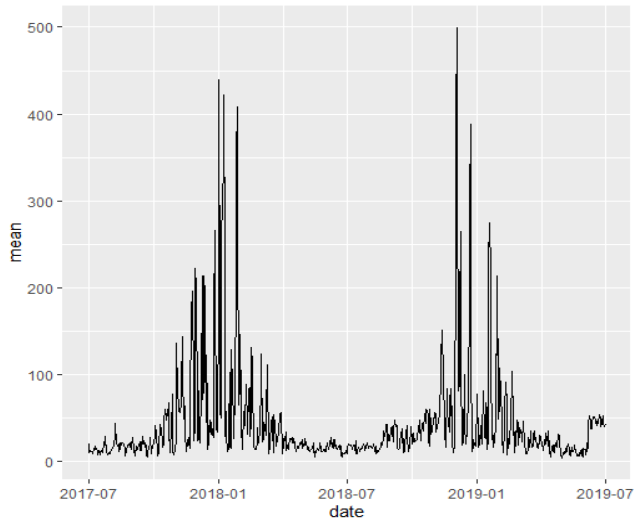

Figure 7: Time series plot for PM2.5

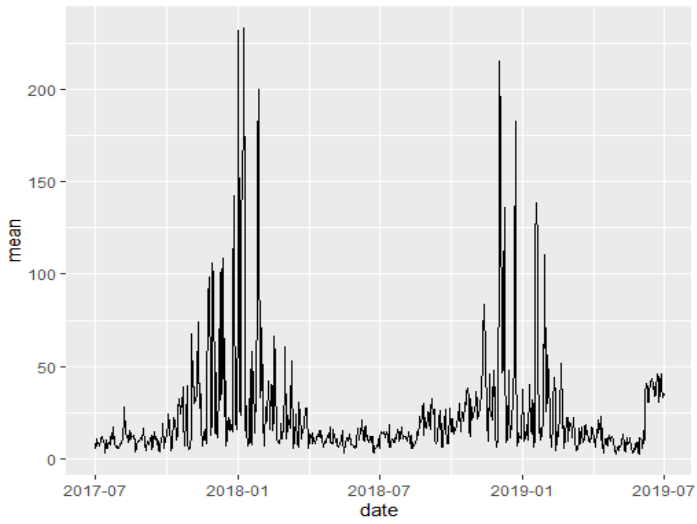

Figure 8: Time series plot for PM10

In order to reduce the seasonal effect, we built ANOVA tables, showing in table 4 and table 5. ANOVA is

meaningful since our data-set, have more than 60 records for each month.
The p-value for the test is much smaller than 0.05 so there exists seasonal effect. We remove seasonal effect and use residuals to build time series model. Residual plots are showing in figure 9 and figure 10.

Table 4: ANOVA of seasonal effects of PM2.5

	Df	Sum Sq	Mean Sq	F value	Pr(¿F)
month	11	748054	68005	26.76	<2e-16 ***
Residuals	713	1812085	2541		

Table 5: ANOVA of seasonal effects of PM10

	Df	Sum Sq	Mean Sq	F value	Pr(¿F)
month	11	160766	14615	23.9	<2e-16 ***
Residuals	713	435939	611		

The blue lines surrounded by marked grey zones are the results from geometric smoothing method. In fact, we mark these zones to make sure the overall seasonal effect is reduced.

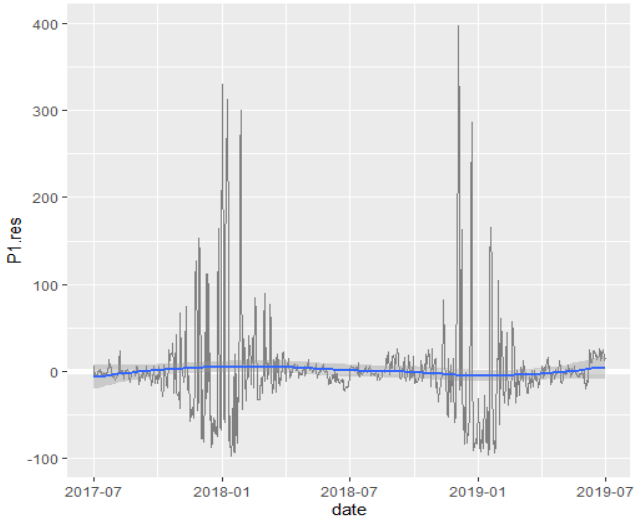

Figure 9: Time series plot for PM2.5
(Remove seasonal effects)

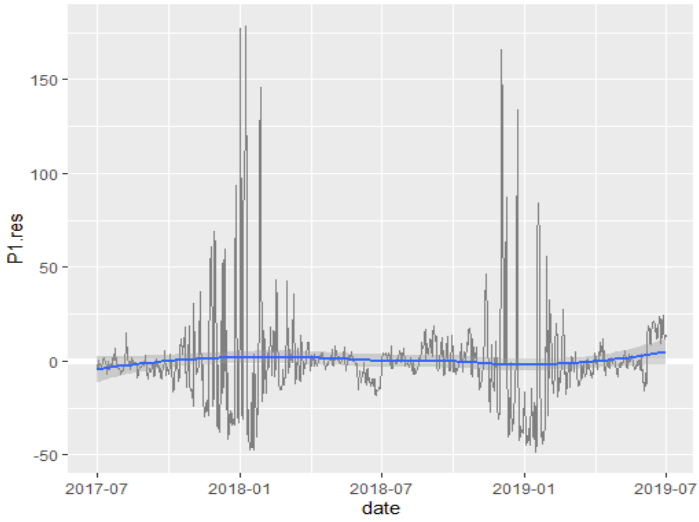

Figure 10: Time series plot for PM10
(Remove seasonal effects)

Considering the obvious volatility, rather than homoscedasticity, we want to use GARCH model to deal with residuals.
A GARCH(1,1), for PM2.5 and PM10, with smaller AIC, 8.686259 and 8.686259, and BIC, 8.717888 and 8.717888, compared with other GARCH model is our choice.
The prediction of future five days (from 2019-07-01 to 2019-07-05) with seasonal influence is showing in the following table 6.

Table 6: Prediction of PM2.5 and PM10 Next 5 days

	1	2	3	4	5
PM2.5	32.9448	34.2625	35.6526	37.1205	38.6718
PM10	27.6259	28.8459	30.1396	31.5123	32.9695

3.1 Difficulties

It is difficult for us to show daily data for two years. Actually we would make efforts to overcome that analyze every time in the data. Drawing pictures and running code in CHTC are quite time consuming.