Study on Shanghai PM 2.5
Final Report

April 28, 2019

Consultants: Wei Zhang
Zhu Wang
Client: Feifang Hu

Abstract

OBJECTIVE To examine the PM2.5 problem is getting better or worse in the past years in Shanghai. And predict the PM2.5 level in the future years. Then give suggestions to reduce air pollution.

METHODS We used univariate linear regression method and Mann-Kendall test to test the trend of monthly average and maximum PM 2.5 values. Moreover, we fitted SARIMA model, exponential smoothing models, and GAM models to predict the future Shanghai PM 2.5.

RESULTS Based on the results of analysis, both monthly average and maximum PM 2.5 have a decreasing trends. For predicting the monthly average values, SARIMA model shows the best results. And GAM 1 model is the best model for predicting monthly maximum PM 2.5 values.

CONCLUSION We can conclude that there is a decreasing trend for PM2.5 levels in recent years, with an obvious seasonal variation. The prediction of future indicates that the air quality is getting better. But people and government still need to take close attention and active actions to improve the air quality.

1 Introduction

In this study, PM2.5 data of Shanghai was analyzed to determine whether PM2.5 level has increased or decreased. Shanghai is a global financial center and transport hub which is located on China?s east coast. The air pollution in Shanghai is not as severe as in many other Chinese cities, but still substantial by world standards. PM2.5 has a great influence on human?s health and is one of the major sources of air pollution. Therefore, the study of PM2.5 is of great significance. The study mainly focuses on three questions: i) What are the relevant factors for PM2.5? How are they affecting PM2.5? ii) Is PM2.5 getting better or worse in the past several years? iii) Is there any suggestions to improve air quality?

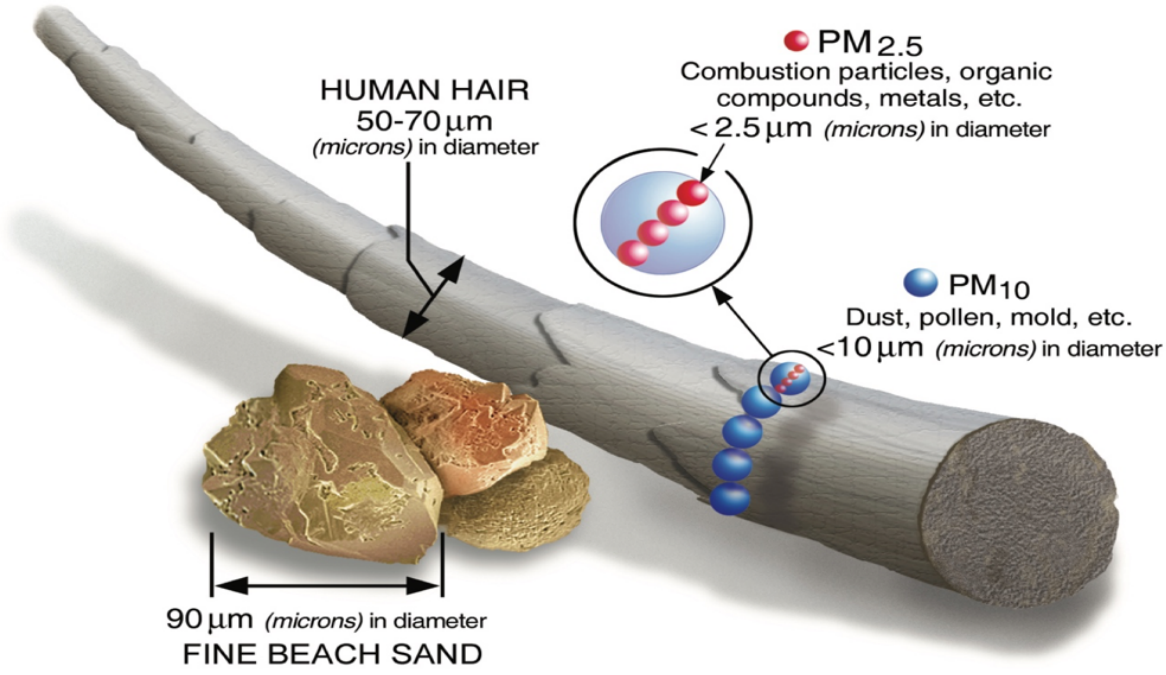
In this project, we are going to use statistical methods to answer the above questions.

2 Background and Dataset

2.1 Background

PM stands for particulate matter, also called particle pollution, is a mixture of solid particles and liquid droplets found in the air. PM2.5 is a kind of fine inhalble particles with diameters that are generally 2.5 micrometers and smaller which means it can get deep into human?s lung and some may even get into human?s bloodstream. These particles come from either natural such as volcanoes, dust storms or human activities like burning of fossil fuels in vehicles and various industrial processes. PM2.5 has severe harmful effect on both human?s health, for example irregular heartbeat or decreased lung function, and the environment. In 2013, the ESCAPE study involving 312,944 people in nine European countries revealed that there was no safe level of particulates and that for every increase of 10 micrograms per cubic meters in PM2.5, the lung cancer rate rose 36%. For those issues, people need to pay highly attention to air pollution especially PM2.5.

The local government of Shanghai launched two important policies to tackle the serious air pollution problem in Shanghai. One is Clean Air Action Plan in 2013, which aims to reduce the concentration of PM2.5 by 20 percent in five years. Shanghai?s clean air action plan is the first regulation in China to prevent and control volatile compounds that make up PM2.5. And the other is Shanghai?s Environmental Protection Plan in 2015, the goal is to reduce the average PM2.5 concentration to 48 micrograms per cubic meters by the end of 2017.



2.2 Data Description

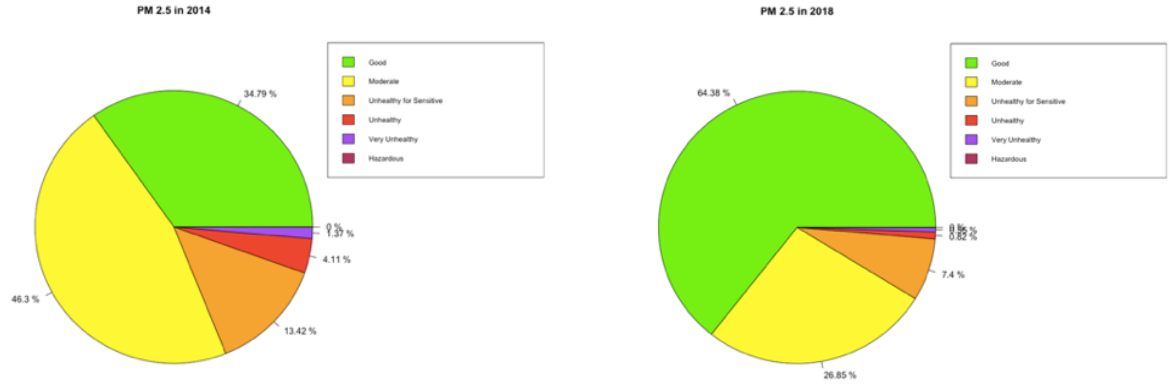
The data of this study were collected from several different sites (Appendix 2.1). The first dataset is from the US Embassy, which contains hourly data from 2012 through June 2017. The second dataset is from Shanghai Environmental Monitoring Center, the dataset consists of measurements from 2016 through 2019. This dataset provides not only PM2.5 values, but also values for other pollution sources such as PM10, CO, SO2. The third dataset is from Shanghai Municipal Bureau of Ecology and Environment, which provides 2018 hourly data of different blocks in Shanghai. The fourth dataset consists of 2016-2019 Shanghai weather data, which is from tutiempo.net. We also get supplementary data of PM2.5 values from aqistudy. To sum up, we get a total of 13 variables, and more than 50,000 observations.

At the outset, we check for missing values for all datasets. The problem of missing value exists in three of the datasets (Appendix 2.2). So we use EM algorithm to do the data imputation. We uses R to procedure the dataset. The packages we used include 'stat','mgcv','forecast','tseries' etc. The explicit R code are attached in the Appendix.

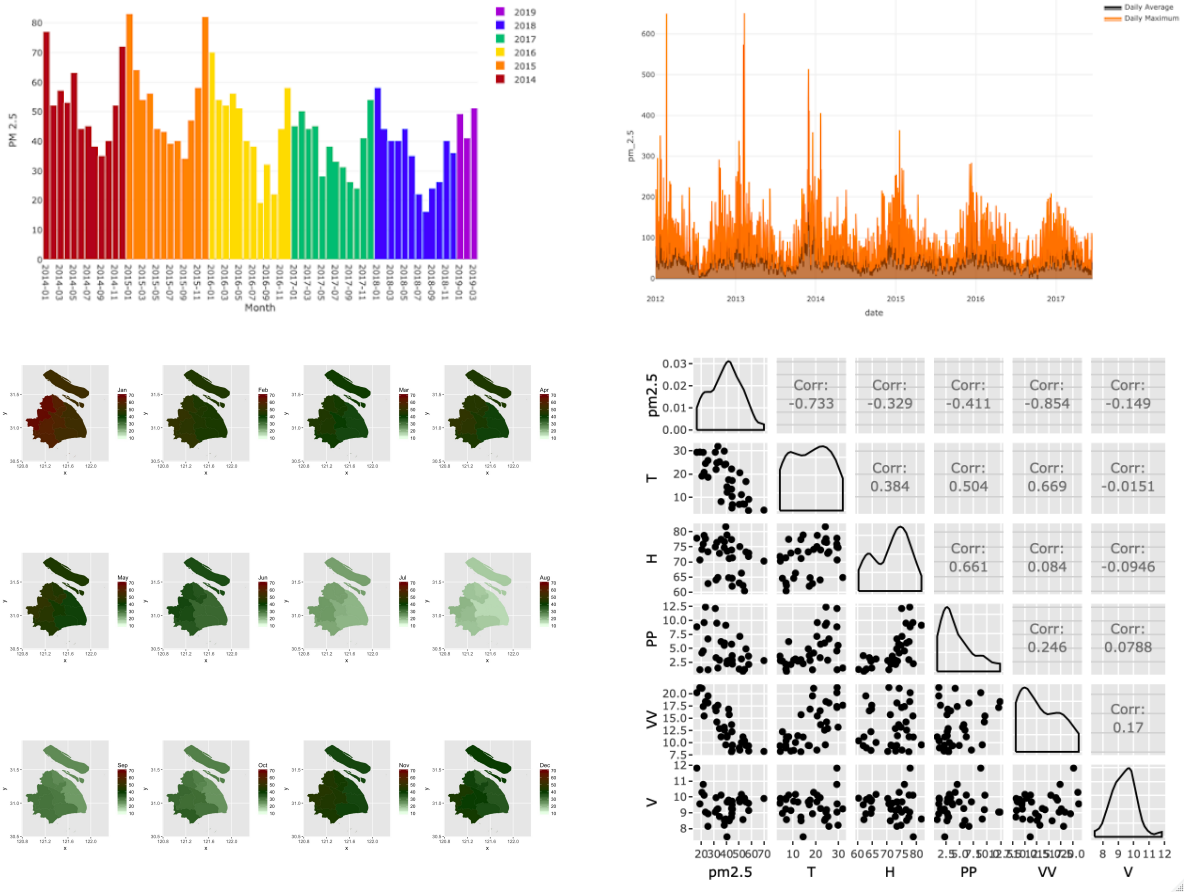
3 Data Analysis and Results

3.1 Exploratory Data Analysis

First of all, we take a look at the trend of AQI in recent years. By world?s standard, we divided the air quality into 6 groups from good to hazardous. As shown in the pie chart below, we can clearly see a better air quality from 2014 (left) to 2018 (right).

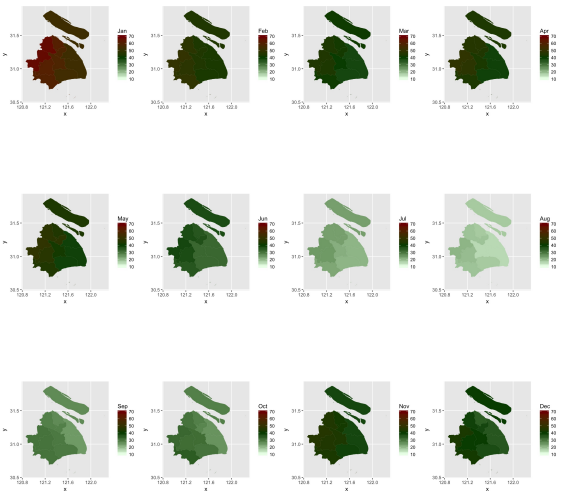


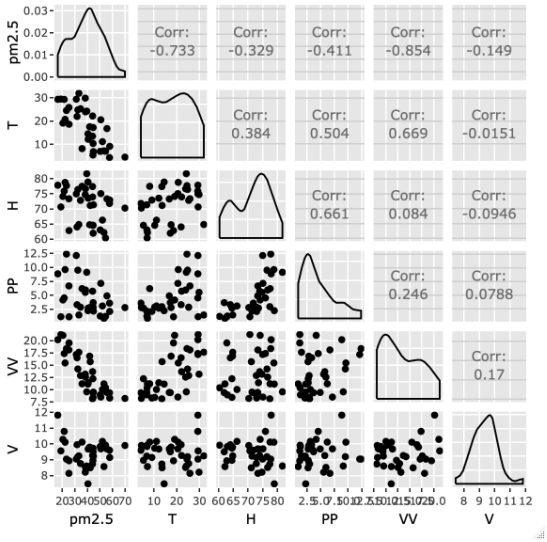
Then we use the monthly average data from AQI study to take an overall look at the changing trend for PM2.5 from January 2014 to March 2019. It shows a decreasing trend yearly, and we can see a seasonal pattern within years (left).



Also, we use the hourly data from the US Embassy, we calculated the daily average PM2.5 and daily maximum PM2.5. The changing trend (right) coincides with the analysis above. Then we test the correlation of PM2.5 between other pollution source and the correlation of PM2.5 between weather (Appendix 3.1). We can conclude that PM2.5 has a strong positive relationship with PM10, SO2, NO2, and NO. For weather factors, temperature and visibility have strong negative relationship with PM2.5. For more exploration, we use spatial analysis method to check out whether

the air quality in Shanghai vary for different districts. As shown above, the overall air quality in January is worse than August. And the regions near the sea have better air quality than inner-land.





3.2 Statistical Analysis

In this section, we use several methods to analysis whether there is a trend in Shanghai pm 2.5 and fit several models to predict the future pm2.5 values.

Combine Datasets

In order to detect a general trend of pm 2.5 in recent years, we used the 2012-2017 dataset from the US Embassy (original data) to derive average pm 2.5 values and maximum pm 2.5 values in every month. After a t-test analysis (Appendix) that showed no difference between Embassy dataset and aqistudy dataset (new data), we combined these two dataset and repeated the steps in analyzing the US Embassy data. We added the US Embassy data to aqistudy data from the range of 2012 to 2014. So the new data contains pm 2.5 concentration values up to date from 2012.

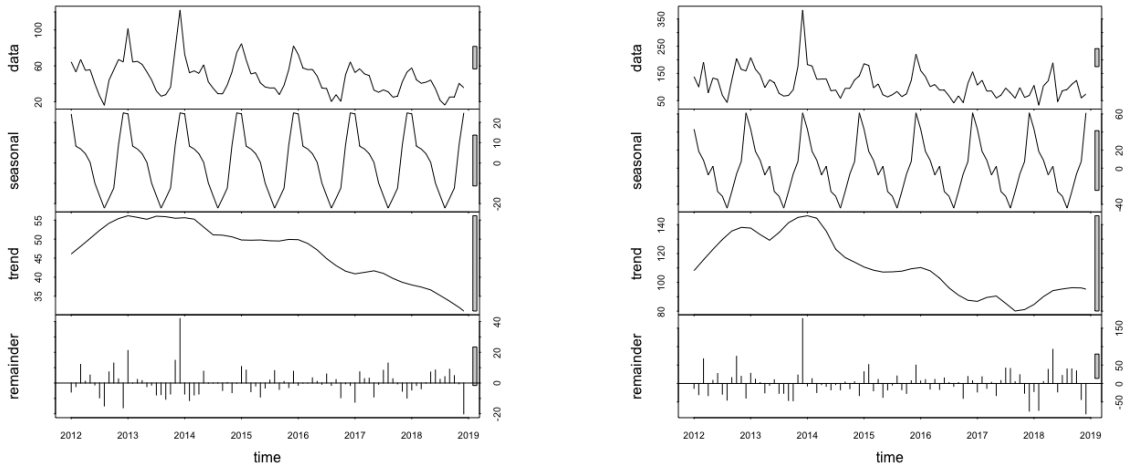
3.2.1 Trends Analysis

Time Series Components

Assume the time series data can be decomposed as an additive model:

$$y_t = S_t + T_t + R_t,$$

where y_t is the data, S_t is the seasonal component, T_t is the trend-cycle component, and R_t is the remainder component.

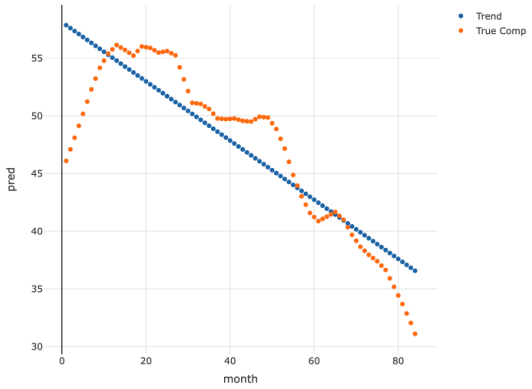


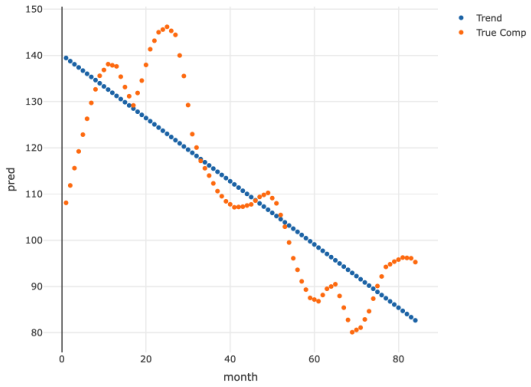
We decomposed monthly average PM 2.5 and maximum PM 2.5 datasets by the Seasonal and Trend decomposition using Loess (STL) method, which is a robust method for decomposing time series. Above are the plots of decomposition of the two datasets, the left is the decomposition of PM 2.5 monthly average, and the right is the monthly maximum. The first line is the data plots; the second line is the seasonal components; the third line is the trend components and the bottom line is the residuals when the seasonal and trend components have been subtracted. To analyze the trend, we directly applied trend analysis methods to the trend-cycle components only.

Univariate Linear Regression Method

After extracted the trend cycle components, we set the PM 2.5 monthly average value as the response variable, and month as the explanatory variable to fit a univariate linear regression model. As the plot with true component and fitted trend shows on the left below, we can see that there may have a downward trend in the monthly average PM 2.5 concentration. The estimate of this model is -0.279 and the p-value < 0.05 also proved that the this estimate is significant. Hence, we can conclude that the monthly average PM 2.5 values have a downward trend in 2012 to 2017.

We did the same steps to the monthly maximum pm 2.5 values. The estimate is -0.746, which is also significant based on the p-value. Compared with the estimates between monthly average, the monthly maximum has a steeper downward trends, which means that the degree of PM 2.5 is milder in 2017 compared with 2012 (summary results in Appendix).





Mann-Kendall Test

We did a Mann Kendall Trend Test to confirm our results in the univariate linear regression method. The M-K test is a non-parametric test which used to analyze time series data for monotonic trends. The hypothesis test is designed as following:

H_0: There does not exist a trend

H_a: There exists a trend

From the summary below, the p values tell that both monthly average and maximum are significant. And the tau values are negative, this means that there are increasing trend for both monthly average and maximum PM 2.5 values.

```
Score =  -2535 , Var(Score) = 74404.34
denominator =  3741
tau = -0.678, 2-sided pvalue =< 2.22e-16
```

```
Score =  -2398 , Var(Score) = 76985.34
denominator =  3828
tau = -0.626, 2-sided pvalue =< 2.22e-16
```

From the results of univariate linear regression method and Mann-Kendall test, we can conclude that the Shanghai PM 2.5 concentration is decreasing in the recent years.

3.2.2 Model Fitting

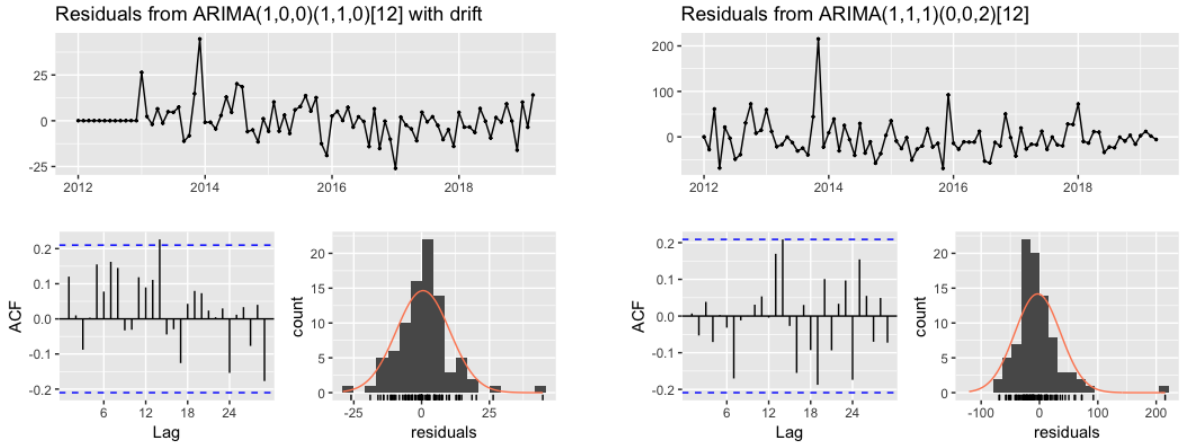
After analyzing the trend of PM 2.5 in recent year, we wanted to predict whether pm 2.5 values based would decrease in the future. ARIMA models and exponential smoothing are the two most common way to approach time series forecasting. Moreover, based on the trend and seasonal components, we also fitted several generalized linear models.

SARIMA Model

ARIMA model is known as autoregressive integrated moving average model, and seasonal ARIMA is an extension of ARIMA with a seasonal component. The equation of general SARIMA $(p,d,q)(P,D,Q)_m$ is given by following:

$$\Phi\left(B^m\right)\phi(B)\nabla_m^D\nabla^d X_t = \Theta\left(B^m\right)\theta(B)Z_t$$

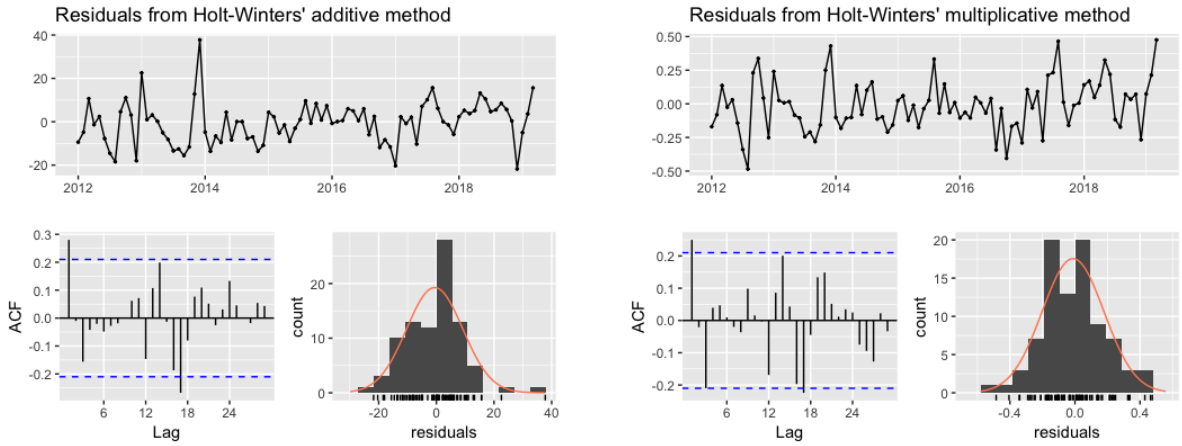
Using the R function 'auto.arima' in the 'forecast' package, we fit a SARIMA$(1,0,0)(1,1,0)_{12}$ and a SARIMA $(1,1,1)(0,0,2)_{12}$ respectfully for monthly average pm 2.5 and monthly maximum pm 2.5, and here are the residuals for these two models.



As the plots above show that the ACF of residuals for both monthly average and monthly maximum are between the dot lines, the residuals appear to be white noise. The Ljung-Box test (See Appendix) also shows that the residuals have no remaining autocorrelations. Thus, the SARIMA models can be used to forecast both monthly average and maximum PM 2.5 values.

Exponential Smoothing Method

Exponential smoothing method weighed the observation exponentially with more weights on the recent observations and less weights on the past. This method is commonly used in seasonal time series data. In our project, we chose to use Holt-Winters' seasonal method, which decomposed the predictor with three components: a level l_t, a trend b_t, and a seasonal s_t. There are two variations for Holt-Winters' method, one is the Holt-Winters' Additive method, and another is the Holt-Winters' Multiplicative method, which decomposed the predictor into different formats with the same components. We tried both additive method and multiplicative method for monthly average and monthly maximum PM 2.5 values. Here are the residuals plots for these two methods. The top is the plots for monthly average and the bottom is maximum.



For both residuals plots of average and maximum (see maximum residuals plots in Appendix), the residuals of Holt Winters′ Additive method show more likely to be white noise, since the distribution of the residuals are closed to normal. However, the p-values of the Ljung-Box test results for all models show insignificant. Hence, the exponential smoothing methods might not be the best models for prediction.

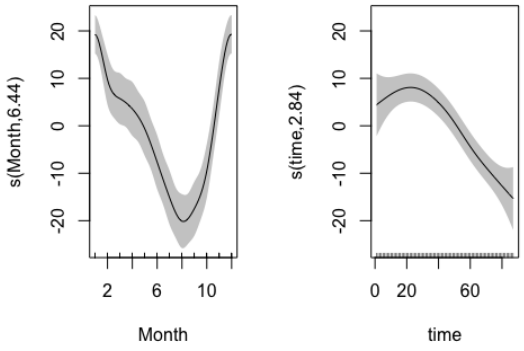
Generalized Additive Models

Generalized additive model is a kind of generalized linear model, in which the linear predictor depends on smooth function of variables rather than variables themselves. GAM has been widely used in seasonal time series analysis. The model is shown as below:

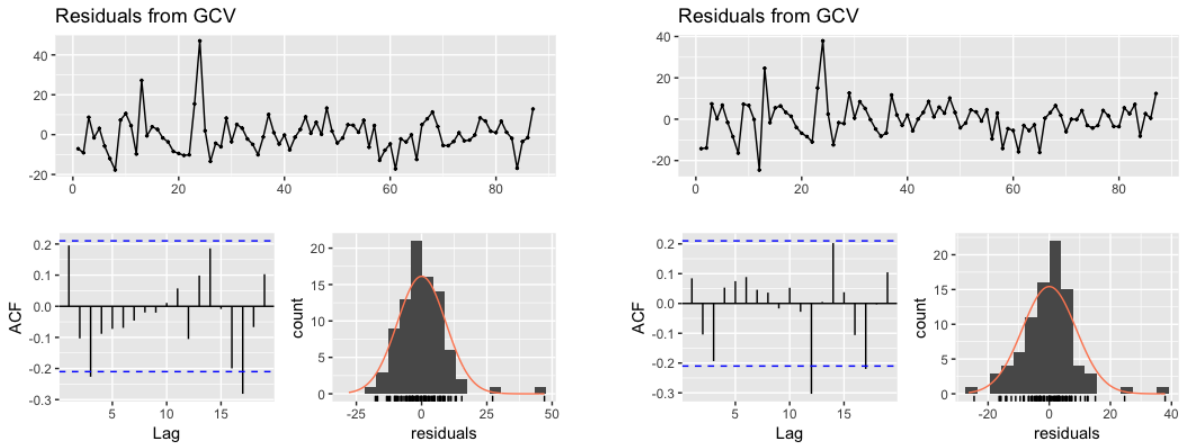
$$y = \beta_0 + f_{\text{seasonal}}\ (x_1) + f_{\text{trend}}\ (x_2) + \varepsilon, \quad \varepsilon \sim N\left(0, \sigma^2 \mathbf{\Lambda}\right)$$

We tried two different smoothing methods to fit the monthly average and maximum PM 2.5 values: i) smoothing the time and month components separately and ii) smoothing the time and month components plus a product interaction term. Below is the splines plots for the monthly average values for model i), the model directly extract the trend and seasonal components for the average values. The table on the right is the estimate degree of freedom for two GAM models of average PM 2.5 values. As we can see that the GAM 2 decreases the degree of freedom for all smoothing variables, with increased r-sq. value. The estimates are all significant based on the F-test.

When checking the residuals plots for these two GAM models of monthly average PM 2.5 values, we can see that the residuals are both satisfied the normal assumptions. Moreover, based on the Breusch-Godfrey test for serial correlation, the residuals show no correlation with each other.



	GAM 1	GAM 2
s(Month)	6.444 ***	4.867 ***
s(Time)	2.840 ***	1.829 ***
ti(Month,Time)		2.232 **
R-Square	0.703	0.736



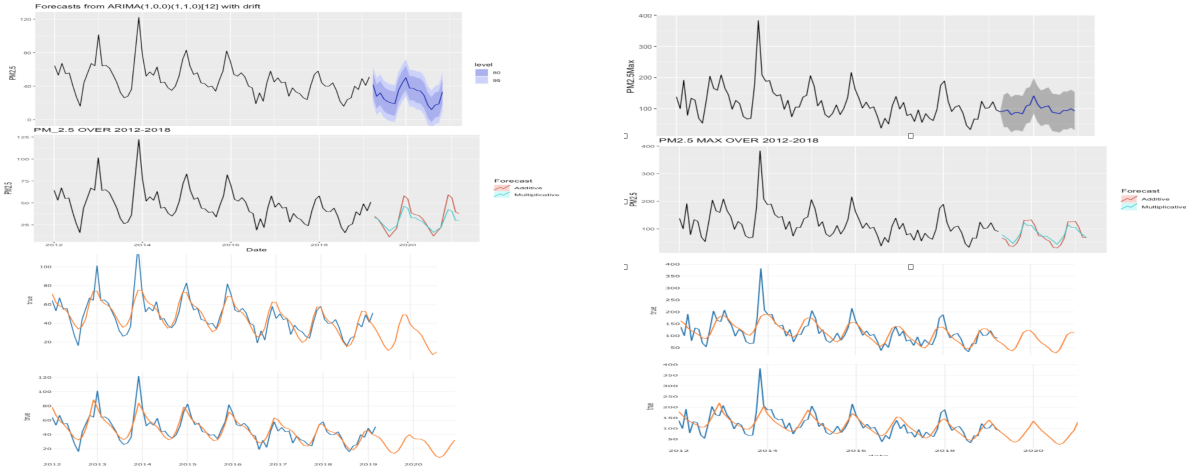
However, for average maximum PM 2.5 values, the residuals plots of GAM 2 show that the residuals might not satisfied the normal assumptions (see Appendix). Therefore, for monthly average values, the GAM models can be used to predict future average PM 2.5. But the GAM 2 might not be precisely models for monthly maximum PM 2.5.

3.2.3 Forecast Results

We have already fitted 5 different models to predict monthly average and maximum PM 2.5, here is the table for the AIC values of different models.

	SARIMA	HW Additive	HW Multiplicative	GAM 1	GAM 2
Avg.	583.81	818.2102	798.7903	656.3680	645.8053
Max.	903.85	1059.748	1040.291	888.0311	893.6073

For monthly average PM 2.5 data, SARIMA model has the lowest AIC value; For monthly maximum PM 2.5 data, GAM 1 model has the lowest AIC value. Hence, we can conclude that the best models for monthly average value prediction is SARIMA$(1,0,0)(1,1,0)_{12}$. As for the monthly maximum prediction, the best model is GAM 1. The prediction results for each models are showed below. Left plots are the forecast of monthly average, and right plots are the forecast of monthly maximum. All plots show downwards trend in predicting future PM 2.5 values.



4 Conclusion and Suggestions

The results of the statistical analysis above indicate that the Shanghai PM 2.5 values have a downwards trend in both monthly average and maximum. The predictions of the models also show a slightly decrease in the future Shanghai PM 2.5 degree. Based on the results, we provided some suggestions that could help improving Shanghai air quality.

Suggestions

According to the director of Shanghai Environmental Protection Bureau, there are external and internal causes for the long-term and severe pollution situation in Shanghai. A cold front brought polluted air from northern China cities, which may use coal-burning boilers for central heating system, while a high pressure cyclone in south of Shanghai prevents dirty air from being blown further to the sea. Besides, tree density decreased in late Autumn may also be a reason that air pollution is more serious during winter for the trees can no longer trap dust. And straw burning in provinces nearby during October to December polluted the air as well. To tackle the air pollution problem, the local government launched the clean air action plan and environmental protection plan, including emissions from vehicles, industry, surrounding provinces and agriculture sectors.

In terms of suggestion, the government should provide strict supervision and take penalization into action. For industrial, the carbon emission trading scheme is a good method to control the total emission. However, new technologies and upgraded facilities can help to reduce the emission from the source. For Agricultural, straw burning should be strictly banned. The government can fund the construction of factories for straw power generation. For transportation, the government has extended subsidies for renewable energy cars. Meanwhile, Shanghai tightened its ban on Yellow Label vehicles from outer ring roads and adopted the V emission standards for all new vehicles. Although a lot of work have be done, the PM2.5 problem cannot be solved overnight. We still need to take long-term action to improve the air quality.

5 Reference

- [1] Ole Raaschou-Nielsen; et al. (July 10, 2013). "Air pollution and lung cancer incidence in 17 European cohorts: prospective analyses from the European Study of Cohorts for Air Pollution Effects (ESCAPE)". The Lancet Oncology. 14 (9): 813?22.
- [2] Ghassan B. Hamra,1 NeelaGuha,1 Aaron Cohen; et al. (September 2014). "Outdoor Particulate Matter Exposure and Lung Cancer: A Systematic Review and Meta-Analysis". Environmental Health Perspectives. 122 (9): 906?11. doi:10.1289/ehp.1408092.
- [3] Hyndman, Rob J. & Athanasopoulos, George. & OTexts.com, issuing body. (2014). Forecasting : principles and practice. Heathmont, Victoria: OTexts.com
- [4] Forecasting: Principles and Practice. 8.9 Seasonal ARIMA Models, otexts.com/fpp2/seasonal-arima.html.
- [5] Laurinec, Peter. ?Doing Magic and Analyzing Seasonal Time Series with GAM (Generalized Additive Model) in R ? Peter Laurinec ? Time Series Data Mining in R. Bratislava, Slovakia.? Doing Magic and Analyzing Seasonal Time Series with GAM (Generalized Additive Model) in R ? Peter Laurinec ? Time Series Data Mining in R. Bratislava, Slovakia., petolau.github.io/Analyzing-double-seasonal-time-series-with-GAM-in-R/

Appendix

2.1 Data Description

2012-2017 Hourly Data from US Embassy

Site	Parameter	Date..LST.	Year	Month	Day	Hour	Value	Unit
Shanghai	PM2.5	1/1/2012 0:00	2012	1	1	0	112	_/m
Shanghai	PM2.5	1/1/2012 1:00	2012	1	1	1	113	_/m
Shanghai	PM2.5	1/1/2012 2:00	2012	1	1	2	115	_/m
Shanghai	PM2.5	1/1/2012 3:00	2012	1	1	3	144	_/m
Shanghai	PM2.5	1/1/2012 4:00	2012	1	1	4	152	_/m
Shanghai	PM2.5	1/1/2012 5:00	2012	1	1	5	138	_/m

2016-2019 Daily Data from Shanghai Environmental Monitoring Center

id	date	pm2.5	pm10	o3	so2	no2	co	aqi	quality
1118	2016-01-01	79	58	41	17	75	23	79	良
1117	2016-01-02	133	87	32	26	104	35	133	轻度污染
1116	2016-01-03	207	114	17	28	124	45	207	重度污染
1115	2016-01-04	165	83	45	27	102	35	165	中度污染
1114	2016-01-05	36	NA	47	13	35	15	47	优
1113	2016-01-06	64	55	37	18	44	18	64	良

2018 Hourly Data of Different Blocks from Shanghai Municipal Bureau of Ecology and Environment

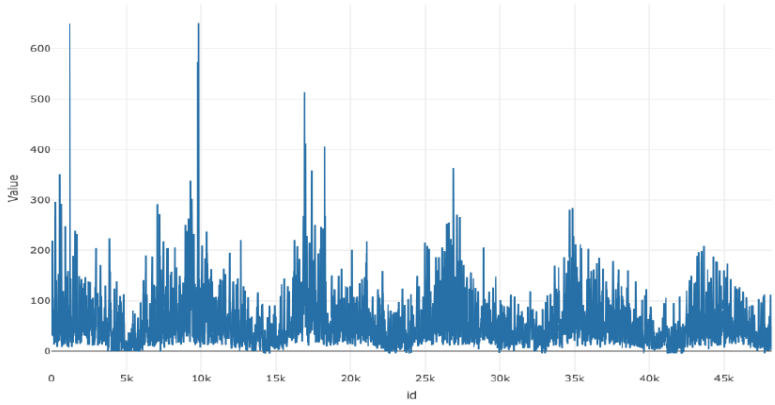
time	site	aqi	level	pm25	pm10	co	no2	ozone1hour	ozone8hour
12/1/2018 1:00	Shiwuchang	67	2	35	83	0.5	58	52	52
12/1/2018 1:00	Hongkou	60	2	26	70	0.5	43	64	64
12/1/2018 1:00	Xuhui	59	2	29	68	0.4	55	48	48
12/1/2018 1:00	Yangpu	68	2	34	86	0.5	62	38	38
12/1/2018 1:00	Qingpu	77	2	33	103	0.4	50	34	34
12/1/2018 1:00	Jingan	63	2	22	75	0.3	45	76	76

2016-2019 Shanghai Weather Data from Tutiempo.net

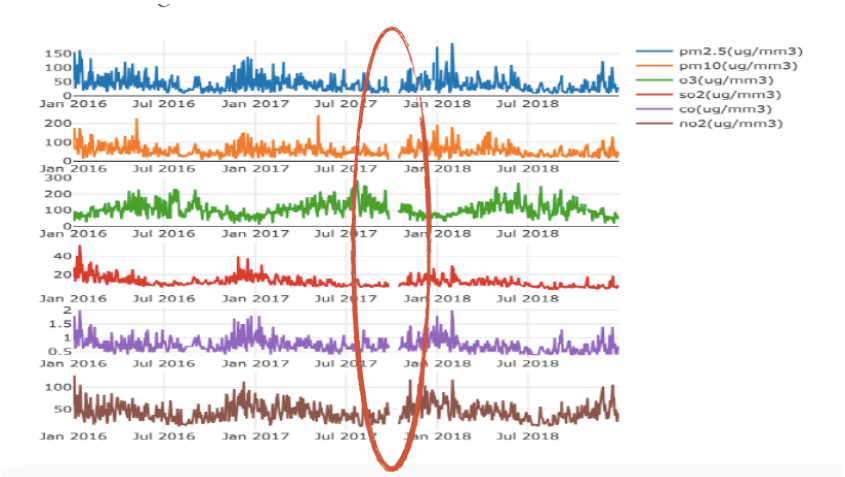
Y	M	D	T	TM	Tm	SLP	STP	H	PP	VV	V	VM	VG	FG
2016	1	1	9.1	12.5	-0.9	1029	1028.1	67	0	7.6	6.7	10.7	-	0
2016	1	2	11.1	16.9	7.3	1022.6	1021.8	62	0	5	5.4	10.7	-	0
2016	1	3	12.1	15.7	7.3	1021.3	1020.5	78	1.78	3.1	3.1	3.5	-	0
2016	1	4	11.3	13.5	9	1021.4	1020.6	89	9.14	2.7	11.7	18	-	0
2016	1	5	8.6	9.2	7.7	1025.2	1024.4	81	1.27	14.8	13	18	-	0
2016	1	6	7.4	9.2	4	1028.4	1027.5	74	0	8.5	9.1	18	-	0

2.2 Missing Values

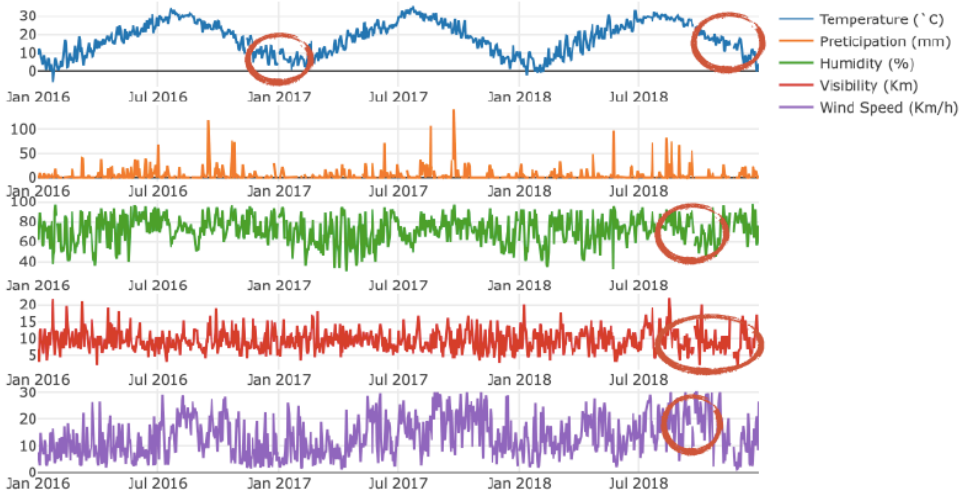
2012-2017 Hourly Data from US Embassy



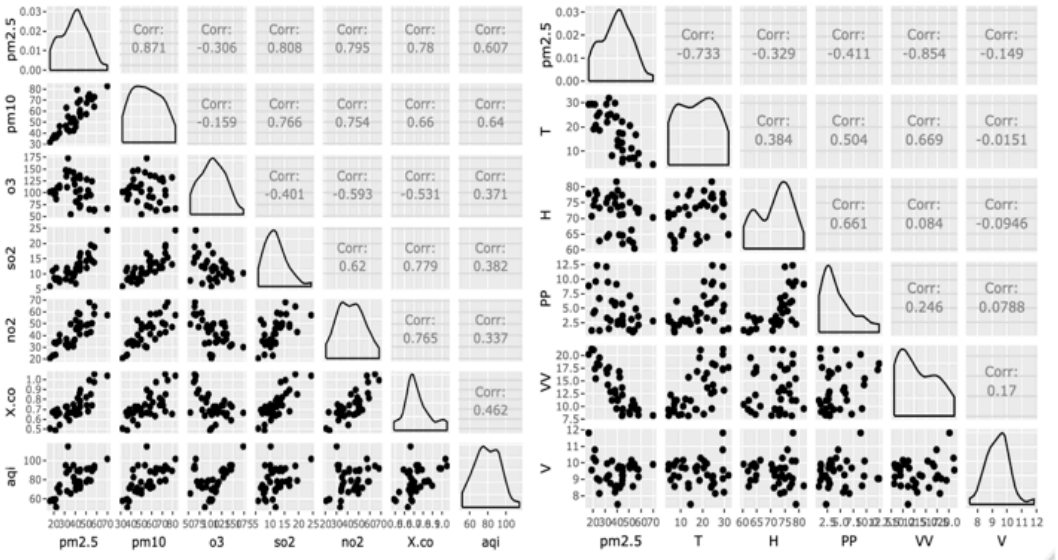
2016-2019 Daily Data from Shanghai Environmental Monitoring Center



2016-2019 Shanghai Weather

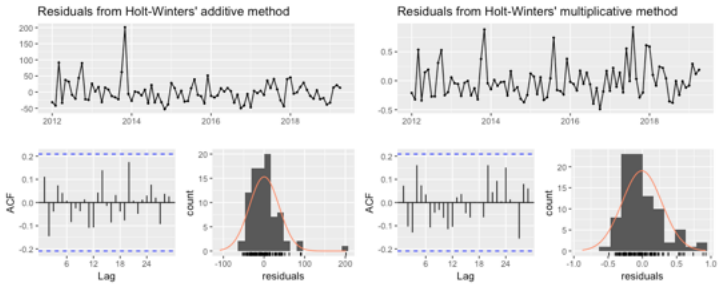


3.1 Correlation

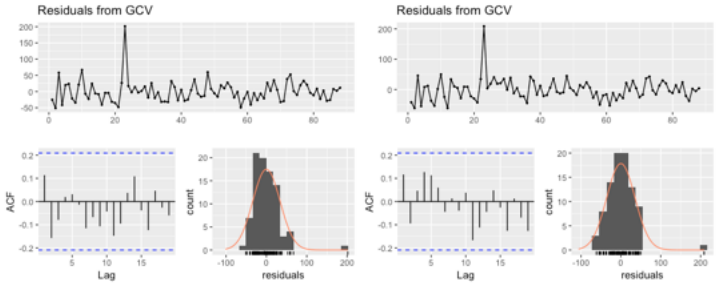


3.2 Statistical Analysis

Maximum residuals for Holt Winters'



Maximum residuals for GAM



Codes

```
## New data
pollution=read.csv("air pollution.csv")
pollution_mon=read.csv("monthly.csv")
### Compare
pollution$date=as.Date(pollution$date)
compare1=merge(pollution,dat2,by="date",all=TRUE)
plot_ly(data=compare1,x=~date) %>%
    add_lines(y=~pm_2.5,name="Old") %>%
    add_lines(y=~pm2.5,name="New",opacity=0.8) %>%
    layout(yaxis=list(title="Values"))
d1=dat2[c(1:731),]
k=c(d1$pm_2.5,pollution$pm2.5)
compare2=data.frame(date=compare1$date,pm2.5=k)
pollution_mon$month=seq(from=as.Date("2014-01-01"),length.out=63,by="month")
names(combin)[2]="month"
compare3=merge(pollution_mon,combin,by='month',all=TRUE)
plot_ly(data=compare3,x=~month) %>%
    add_lines(y=~pm2.5mon,name="Old") %>%
    add_lines(y=~pm2.5,name="New") %>%
    layout(yaxis=list(title="Values"))
d2=combin[c(1:24),]
d=c(d2$pm2.5mon,pollution_mon$pm2.5)
compare4=data.frame(month=compare3$month,pm2.5=d)
### Trend Analysis
#### mann-kendall trend test
library(Kendall)
tss=ts(compare4$pm2.5,frequency=12,start=c(2012,1))
decompose1=stl(tss,s.window="periodic")
plot(decompose1)
trend11=decompose1$time.series[,2]
str(trend11)
trend12=data.frame(month=c(1:87),trendmon=trend11)
trendmodel3=lm(trendmon~month,data=trend12)
summary(trendmodel3)
predict=predict(trendmodel3,type="response")
predict2=data.frame(month=c(1:87),trendmon=predict)
plot_ly() %>%
    add_lines(data=predict2,x=~month,y=~trendmon,add=TRUE,name="Trend") %>%
    add_lines(data=trend12,y=~trendmon,x=~month,name="True Comp")
mk=MannKendall(trend11)
```

```
summary(mk)
#### Model fit
compare4$Month=c(rep(c(1:12),7),1,2,3)
compare4$time=c(1:87)
library(mgcv)
m1=auto.arima(tss)
m0=hw(tss,seasonal="additive")
m2= hw(tss,seasonal="multiplicative")
m3=gam(pm2.5~s(Month,bs="cc",k=12)+s(time),data=compare4)
m4=gam(pm2.5~s(Month,time),data=compare4)
period=1
m5=gam(pm2.5~te(time,Month,k=c(period,12),bs=c("cr","ps")),data=compare4)
m6=gam(pm2.5~s(Month,bs="ps",k=12)+s(time,bs="cr",k=period)+ti(time,Month,k=c(period,12),bs=c("cr","ps")),
       data=compare4,family=gaussian)
m7=gam(pm2.5~t2(Month,time,k=c(period,12),bs=c("ps","cr"),full=TRUE),data=compare4)
summary(m1)
ggtsdiag(m1)
checkresiduals(m1)
summary(m0)
checkresiduals(m0)
summary(m2)
checkresiduals(m2)
summary(m3)
par(mfrow=c(1,2))
plot(m3,shade=TRUE)
checkresiduals(m3)
summary(m4)
summary(m5)
summary(m6)
checkresiduals(m6)
summary(m7)
le=data.frame(time=c(87:107),Month=c(5:12,1:12,1))
#predi1=predict(m1,type="response")
#predi2=predict(m2,type="response")
predi3=predict(m3,type="response")
pr3=predict(m3,le,type="response")
predi4=predict(m4,type="response")
pr4=predict(m4,le,type="response")
predi5=predict(m5,type="response")
pr5=predict(m5,le,type="response")
predi6=predict(m6,type="response")
pr6=predict(m6,le,type="response")
predi7=predict(m7,type="response")
pr7=predict(m7,le,type="response")
```

```
dataframe=data.frame(date=seq(from=as.Date("2012-01-
01"),length.out=108,by="month"),true=c(compare4$pm2.5,rep(NA,21)),p3=c(predi3,pr3),p4=c(predi4,pr4),p5=c(predi5,pr5),p6=c
(predi6,pr6),p7=c(predi7,pr7))
p2=plot_ly(data=dataframe,x=~date) %>%
      add_lines(y=~true) %>%
      add_lines(y=~p3) %>%
      layout(showlegend=FALSE)
p3=plot_ly(data=dataframe,x=~date) %>%
      add_lines(y=~true) %>%
      add_lines(y=~p4)    %>%
      layout(showlegend=FALSE)
p4=plot_ly(data=dataframe,x=~date) %>%
      add_lines(y=~true,color="bkack") %>%
      add_lines(y=~p5)    %>%
      layout(showlegend=FALSE)
p5=plot_ly(data=dataframe,x=~date) %>%
      add_lines(y=~true) %>%
      add_lines(y=~p6)    %>%
      layout(showlegend=FALSE)
p6=plot_ly(data=dataframe,x=~date) %>%
      add_lines(y=~true,color="bkack") %>%
      add_lines(y=~p7)    %>%
      layout(showlegend=FALSE)
subplot(p2,p3,p4,p5,p6,nrows=5)
ggplot2::autoplot(tss) +
   autolayer(m0, series="Additive",PI=FALSE) +
   autolayer(m2, series="Multiplicative",
     PI=FALSE) +
   xlab("Date") +
   ylab("PM2.5") +
   ggtitle("PM_2.5 OVER 2012-2018") +
   guides(colour=guide_legend(title="Forecast"))
m1 %>% forecast(h=20) %>%
   autoplot() +
   ylab("PM2.5")
AIC(m3,m4,m5,m6,m7)
### max
max=compare2
max$Year=format(compare2$date, "%Y")
max$Month=format(compare2$date, "%m")
max1=max %>%
      group_by(Year,Month) %>%
      summarise(pm_2.5max = max(pm2.5))
library(Kendall)
```

```
tssm=ts(max1$pm_2.5max,frequency=12,start=c(2012,1))
decompose2=stl(tssm,s.window="periodic")
plot(decompose2)
trend21=decompose2$time.series[,2]
str(trend21)
trend22=data.frame(month=c(1:88),trendmon=trend21)
trendmode23=lm(trendmon~month,data=trend22)
summary(trendmode23)
predict3=predict(trendmode23,type="response")
predict4=data.frame(month=c(1:88),trendmon=predict3)
plot_ly() %>%
      add_lines(data=predict4,x=~month,y=~trendmon,add=TRUE,name="Trend") %>%
      add_lines(data=trend22,y=~trendmon,x=~month,name="True Comp")
mk=MannKendall(trend21)
summary(mk)
#### Model fit
max1$time=c(1:88)
library(mgcv)
max1$Month=as.numeric(max1$Month)
a1=auto.arima(tssm)
a0=hw(tssm,seasonal="additive")
a2= hw(tssm,seasonal="multiplicative")
a3=gam(pm_2.5max~s(Month,bs="cc",k=12)+s(time),data=max1)
a4=gam(pm_2.5max~s(Month,time),data=max1)
period=1
a5=gam(pm_2.5max~te(time,Month,k=c(period,12),bs=c("cr","ps")),data=max1)
a6=gam(pm_2.5max~s(Month,bs="ps",k=12)+s(time,bs="cr",k=period)+ti(time,Month,k=c(period,12),bs=c("cr","ps")),
          data=max1,family=gaussian)
a7=gam(pm_2.5max~t2(Month,time,k=c(period,12),bs=c("ps","cr"),full=TRUE),data=max1)
summary(a0)
checkresiduals(a0)
summary(a1)
ggtsdiag(a1)
checkresiduals(a1)
summary(a2)
checkresiduals(a2)
summary(a3)
par(mfrow=c(1,2))
plot(a3,shade = TRUE)
checkresiduals(a3)
summary(a4)
summary(a5)
summary(a6)
summary(a7)
```

```
le=data.frame(time=c(88:108),Month=c(5:12,1:12,1))
#predi1=predict(m1,type="response")
#predi2=predict(m2,type="response")
pred3=predict(a3,type="response")
pre3=predict(a3,le,type="response")
pred4=predict(a4,type="response")
pre4=predict(a4,le,type="response")
pred5=predict(a5,type="response")
pre5=predict(a5,le,type="response")
pred6=predict(a6,type="response")
pre6=predict(a6,le,type="response")
pred7=predict(a7,type="response")
pre7=predict(a7,le,type="response")
dataframe2=data.frame(date=seq(from=as.Date("2012-01-
01"),length.out=109,by="month"),true=c(max1$pm_2.5max,rep(NA,21)),p3=c(pred3,pre3),p4=c(pred4,pre4),p5=c(pred5,pre5),p6
=c(pred6,pre6),p7=c(pred7,pre7))
p22=plot_ly(data=dataframe2,x=~date) %>%
      add_lines(y=~true) %>%
      add_lines(y=~p3) %>%
      layout(showlegend=FALSE)
p33=plot_ly(data=dataframe2,x=~date) %>%
      add_lines(y=~true,color="black") %>%
      add_lines(y=~p4)    %>%
      layout(showlegend=FALSE)
p44=plot_ly(data=dataframe2,x=~date) %>%
      add_lines(y=~true) %>%
      add_lines(y=~p5)    %>%
      layout(showlegend=FALSE)
p55=plot_ly(data=dataframe2,x=~date) %>%
      add_lines(y=~true) %>%
      add_lines(y=~p6)    %>%
      layout(showlegend=FALSE)
p66=plot_ly(data=dataframe2,x=~date) %>%
      add_lines(y=~true,color="bkack") %>%
      add_lines(y=~p7)    %>%
      layout(showlegend=FALSE)
subplot(p22,p33,p44,p55,p66,nrows=5)
ggplot2::autoplot(tssm) +
      autolayer(a0, series="Additive",PI=FALSE) +
      autolayer(a2, series="Multiplicative",
                  PI=FALSE) +
      xlab("Date") +
      ylab("PM2.5") +
      ggtitle("PM2.5 MAX OVER 2012-2018") +
```

```
      guides(colour=guide_legend(title="Forecast"))
a1 %>% forecast(h=20) %>%
      autoplot() +
      ylab("PM2.5Max")
AIC(a3,a4,a5,a6,a7)
kk=data.frame(pm2.5avg=compare4$pm2.5,pm2.5max=max1$pm_2.5max[1:87],date=compare4$month)
plot_ly(data=kk,x=~date) %>%
      add_lines(y=~pm2.5avg,name="Monthly Average",fill='tonexty') %>%
      add_lines(y=~pm2.5max,name="Montly Maximum",fill='tonexty')
gamm=gam(pm2.5~s(pm10)+s(o3)+s(so2)+s(no2)+s(co),data=pollution)
summary(gamm)
pred8=predict(gamm)
pred8=data.frame(date=pollution$date,fit=pred8,true=pollution$pm2.5)
plot_ly(data=pred8,x=~date) %>%
      add_lines(y=~fit,add=TRUE,name="Fit") %>%
      add_lines(y=~true,name="True Value",opacity=0.5)
ppp=compare1[c(732:1995),]
pppp=compare3[c(25:84),]
t.test(ppp$pm_2.5,ppp$pm2.5,p.adjust.method = "BH")
t.test(pppp$pm2.5,pppp$pm2.5mon,p.adjust.method = "BH")
```