Time Series Analysis of Beijing PM2.5 Pollution

Haoxuan Derek Liu (hdl5hz)
Department of Statistics, University of Virginia

Introduction

Understanding air pollution trends is an important task to combat unhealthy air quality levels and mitigate illnesses caused by pollution. Beijing, China is infamous for high levels of PM2.5, which are fine inhalable particles with diameters 2.5 micrometers or smaller. These particles are the standard for air pollution level measurements, and have long been used as indicators for citizens in and around the city on whether or not preventative measures should be taken on any given day. Understanding the patterns of PM2.5 levels and creating statistical forecasts can help with experimental design that aims to reduce pollution levels year-round.

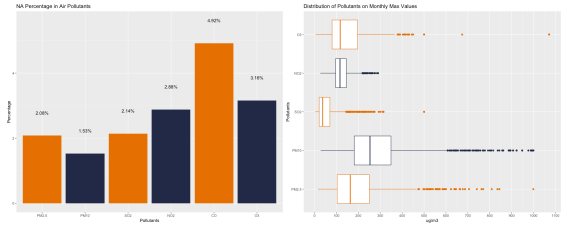
Objective

The goal of this project is to understand the underlying trends of PM2.5 levels in Beijing, China, and how some time-series modelling methods may or may not be appropriate for the specific data set. This project will also extend to forecasting PM2.5 levels for time periods immediately following the data set's end date.

Data Description

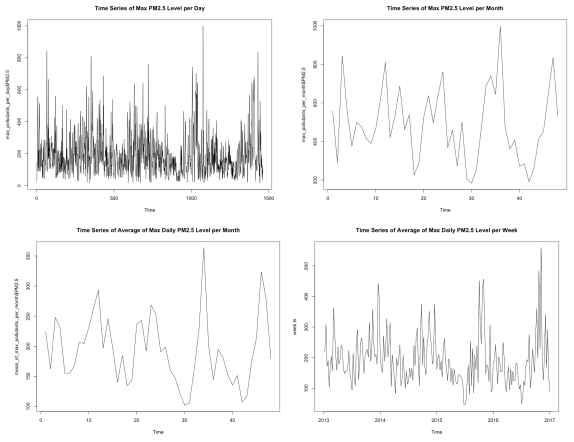
Variable Name	Variable Description
No	Row number
year	Year of data in this row
month	Month of data in this row
day	Day of data in this row
hour	Hour of data in this row
PM2.5	PM2.5 concentration (ug/m^3)
PM10	PM10 concentration (ug/m^3)
SO2	Sulphur Dioxide concentration (ug/m^3)
NO2	Nitrogen Dioxide concentration (ug/m^3)
CO	Carbon Monoxide concentration (ug/m^3)
O3	Ozone concentration (ug/m^3)
TEMP	Temperature (degrees Celcius)
PRES	Pressure (hPa)
DEWP	Dew point temperature (degree Celcius)
RAIN	Precipitation (mm)
wd	Wind direction
WSPM	Wind speed (m/s)
station	Name of the air-quality monitoring site

Exploratory Data Analysis



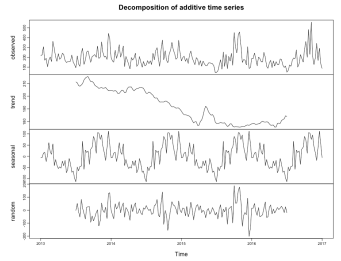
Exploring through the raw data set, there are a total of 18 variables and 420768 hourly observations. The data ranges between "2013-03-01 00:00: 00 UTC" and "2017-02-28 23:00:00 UTC". There are a total of 74027 missing values. However, the percentage of missing values in our variable of interest: PM2.5, is only 2.08%, which is negligible and can be omitted for analysis. The summary for PM2.5 is as follows:

Min	1Q	Med	Mean	3Q	Max	NA
2.00	20.0	55.0	79.8	111	999	8739

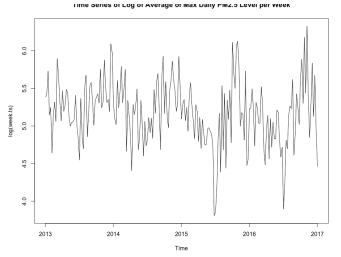


Further exploring PM2.5 through data manipulation, we can view the observations as a time series in many different time intervals. A note before describing each time series graph: The maximum value of each station at each observation is used because that is the level of concern at any given interval for citizens living in the city.
The top left shows the maximum PM2.5 Level per day, aggregating the 12 stations and their hourly data. The top right shows the maximum PM2.5 value by month, aggregating the per day time series array. The bottom left shows the average maximum value per month, which shows no significant trend difference than the maximum value per month, other than the scale.
After working with day, month and average per month arrays, the ACF diagnostics indicate that the trends are strongest on the weekly level, hence the final time-series data is the weekly average of maximum PM2.5 per day, shown in the bottom right.

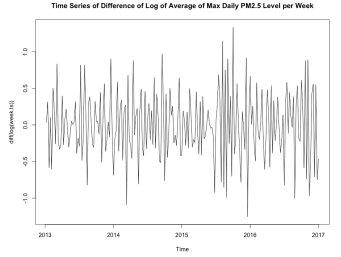
Exploratory Data Analysis (Cont.)



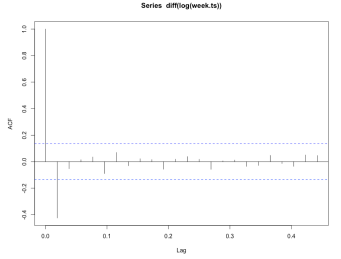
By decomposing the additive time-series of the weekly data, we can see that there is a downward trend that is slowly decreasing over this particular period of time. The seasonal graph suggests potential seasonality when selecting a model.



The original time series of the weekly average maximum for the was not stationary, as can be seen through the decomposition, as it doesn't have a constant mean and variance. By taking the log of the array, we get the new time-series graph on the left, which looks much more stationary than the original.



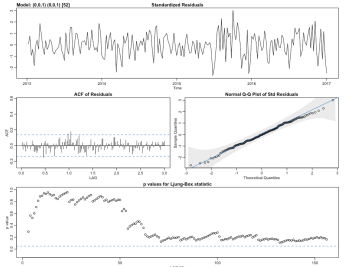
By differencing the log of the time series, the graph becomes even more stationary and looks very similar to white noise. In this case, we can work with the data to build a SARIMA model.



The ACF of the differenced log of the weekly data suggests a spike at lag 0.02, which suggests a MA(1) process since the frequency of the data is set at 52.

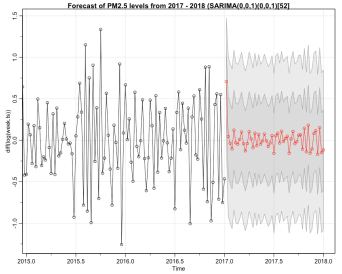
The final SARIMA model that is used is a SARIMA(0,0,1)(0,0,1)52, which has AIC=199.86, and BIC=209.87.

SARIMA Model and Forecast



The final SARIMA model looks to pass every diagnostic:

- Standardized residuals has no clear pattern, similar to white noise
- Sample ACF does not indicate any significant autocorrelation
- QQ-plot suggests the log transformation has made the data normal.
- The Ljung-Box statistics indicate a great fit.



Forecasting 2017-2018 data using the SARIMA model, it appears that the intervals are well calculated that matches previous patterns, and that the point estimates are conservative.

Conclusion

The SARIMA model seems to have a pretty accurate range of prediction and performs well under all assumptions. It seems that the data suggests pollution levels are going to stay quite constant from historic data. There isn't any publicly accessible data between 2017 and 2018 on Beijing pollution level, and recent data from 2019-2021 are too far from the original data for any accurate predictions. Air pollution levels are still an on-going concern in Beijing, and through the seasonal time-series analysis, we can infer that if no substantial change is implemented, the pollution levels will persist.

References

- UCI Machine Learning Repository, Beijing Multi-Site Air-Quality Data Set https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data
- Rob J Hyndman and George Athanasopoulos, Forecasting Principles and Practice (2nd Ed)
- Professor Jeffrey Woo, STAT 5170 Spring 2020 Lectures