Beijing $PM_{2.5}$ Air-Quality Analysis

Isabella Chittumuri
Professor Dana Sylvan
STAT 715: Time Series Analysis
City University of New York, Hunter College

Abstract

Beijing, China is known to experience one of the worst air pollution worldwide. The U.S Environmental Protection Agency (EPA) collaborates with China's Ministry of Environmental Protection (MEP) to offer guidance on ambient air quality standards for six principal pollutants that can be harmful to public and environmental health [1]. In this project, we studied the statistical time series analysis of $PM_{2.5}$ concentration levels in Beijing to model, forecast, and determine air quality levels.

---

1. Introduction

1.1 What is $PM_{2.5}$ and its health risks?

$PM_{2.5}$ is particulate matter in the air that has a diameter of less than 2.5 micrometers. These particles are formed as a result of burning fuel, atmospheric chemical reactions, and forest fires. Since potential health damage caused by air pollutants depends on both concentration and duration of exposure, $PM_{2.5}$ is measured by a 24-hour average index. EPA established a primary and secondary standard which states that "an area meets the standard if the 98th percentile of 24-hour $PM_{2.5}$ concentrations in one year, averaged over three years, is less than or equal to 35 μg/m3" [2].

Primary air quality standards provide public health protection. Because of its small size, $PM_{2.5}$ can penetrate the respiratory tract, deep into the lungs and sometimes enter the circulatory system. Long-term exposure to this particulate matter can lead to cardiovascular and respiratory diseases. Secondary air quality standards provide environmental health protection. $PM_{2.5}$ can affect the stability of ecosystems and contribute to climate change. It can change weather patterns, acidify bodies of water, deplete soil nutrients and damage forests [3].

1.2 Data Collection

This project used the "Beijing Multi-Site Air-Quality Dataset" created by Song Xi Chen, donated in September 2019 to University of California (UCI) Machine Learning Repository. This data was collected by the Beijing Municipal Environmental Monitoring Center from 12 nationally controlled air quality monitoring sites. It was then matched with the nearest weather station from the China Meteorological Administration, which established accuracy and validity. This data is classified as a time series because it is a sequence of observations recorded at regular time intervals. The sampling scheme used was systematic because it was collected at every hour of every day from

March 1st, 2013 to February 28th, 2017, within all 12 monitoring sites. However, there were numerous missing values from all monitoring sites which made the data incomplete with unequal probability.

The data was available in csv format, and it was consistent in the measurement of 18 variables and 420,768 observations. From these variables, we were most interested in the hourly $PM_{2.5}$ concentration levels (in μg/m3) from the Wanliu site. The data collected for this specific pollutant had a total of 35064 observations, with 382 missing values. The hourly date was parsed into four separate columns, one for each year, month, day, and hour.

1.3 Data Processing

First, we combined the four hourly date columns into one cohesive a datetime column. Then, we filled in the missing data with values that lie on a linear curve between existing data points. Finally, we consolidated the 35064 datetime observations into 1461 daily observations, by taking the mean of every 24 hours. This created our target population of 24-hour average $PM_{2.5}$ concentration levels (in $\mu g/m^3$) from March 2013 to February 2017.

In this report we will explore, model, and forecast daily $PM_{2.5}$ levels to predict the spectrum of how unhealthy Beijing's air quality will be from years 2017 to 2020. All the analysis was done using Python 3.6.9.

2. Exploratory Data Analysis

2.1 Data Observations

To understand the distribution of our data values, we computed its descriptive statistics. The mean was 83.47 $\mu g/m^3$, representing the average of 24-hour $PM_{2.5}$ concentration levels within the target population. The median was 64.63 $\mu g/m^3$, representing the middle value of the data once it was set in numeric order. There was no distinct mode, although there was high frequency of numbers between 20-40 $\mu g/m^3$. The difference between the mean, median, and mode suggested a right-skewed distribution. In this case, the median was a better measure of central tendency, than the mean.

The 24-hour $PM_{2.5}$ concentration levels ranged from 4.29 to 481.29 $\mu g/m^3$, which gave a difference of 477 $\mu g/m^3$. Variance and standard deviation measure how dispersed the data values are around the mean. The target population had a variance of 4990.04 $(\mu g/m^3)^2$, with a standard deviation of 70.64 $\mu g/m^3$.

Figure 1. Density Plot of 24-hour $PM_{2.5}$ Levels

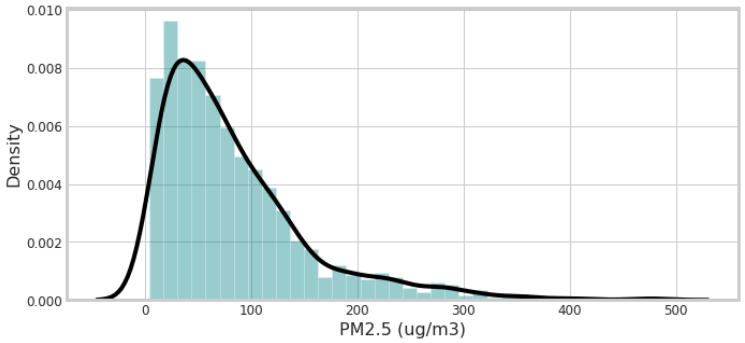


Figure 2. Box and Whisker Plot of 24-hour $PM_{2.5}$ Levels Indexed by Month

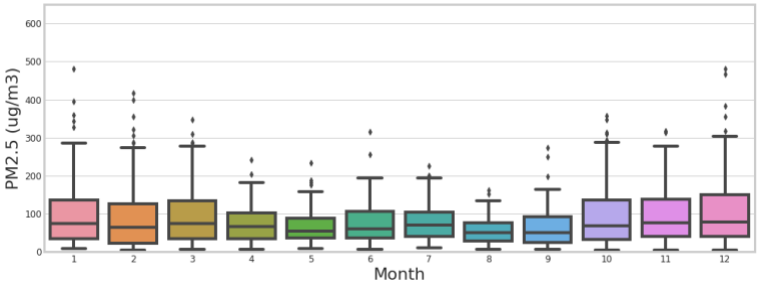


Figure 1 depicts a right-skewed unimodal density distribution. This distribution indicated the possibility of outliers within the target population. To determine which numbers were outliers, we found the upper fence, which was 226.67 $\mu g/m^3$. This meant that any values greater than 226.67$\mu g/m^3$ were outliers. However, **Figure 2**

shows that the outliers differ by month. From months 1-3 and 10-12, outliers start around 300 $\mu g/m^3$. For months 4-9, outliers start around 200 $\mu g/m^3$. This meant that outliers vary monthly throughout the year.

2.2 Time Series Decomposition

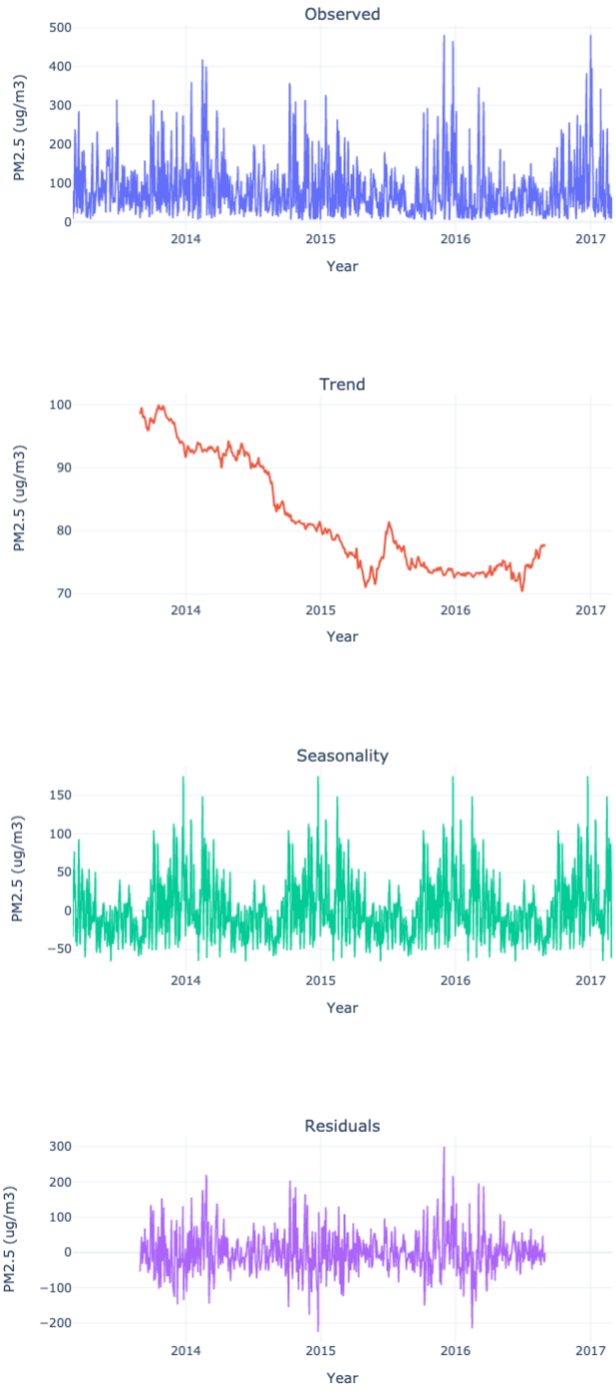
We decomposed the time series to understand the target population's inherent nature. **Figure 3** depicts our time series decomposition consisting of four plots: observed, trend, seasonal, and residual. The observed plot shows the 24-hour average $PM_{2.5}$ concentration levels over time. We saw that concentration levels fluctuate seasonally in Beijing. The concentration levels tend to be high in the beginning of the year and then decrease drastically six months later. This pattern continued over the four years the data was collected on.

The trend component of a time series represents the overall direction and long-term movement of the data values. In our trend plot, we saw that through 2013 to 2015 there is a constant downward movement of $PM_{2.5}$ concentration levels. However, towards the end of 2015 and the beginning of 2016, there was a minor upward spike in concentration levels. For the reminder of the period, the concentration levels went back down and seemed to stabilize around 75 $\mu g/m^3$. Overall, the trend showed a consistent linear decline, with a slight curve as it potentially reaches equilibrium.

The seasonal component of a time series represents the oscillation within yearly variations that is steady over time, direction, and magnitude. Our seasonal plot shows yearly seasonal shifts in $PM_{2.5}$ concentration levels. This confirmed what we presumed in our observed plot: 24-hour $PM_{2.5}$ levels periodically fluctuate in value approximately every six months in Beijing.

The residual component of a time series represents the random unexplainable parts of the data that cannot be assigned to trend or seasonality. The remaining data in our residual plot looked to be inconsistent in values and therefore may not have constant mean or variance [4].

Figure 3. Time Series Decomposition of 24-hour $PM_{2.5}$ Levels



2.3 Yearly Patterns

To clearly identified the seasonal components, we grouped the target population by season and took the yearly mean. **Figure 4** shows the yearly average of $PM_{2.5}$ levels per season. Seasons 1, 2, 3 and 4 respectively represent winter, spring, summer and fall. From this plot, we concluded that concentration levels are significantly higher during winter and significantly lower during summer. Fall and spring concentration levels stay relatively in between the other two seasons.

In the same way, we looked for patterns by grouping the target population by quarters and taking the yearly mean. **Figure 5** shows the yearly average of $PM_{2.5}$ levels per quarter. Quarter 1 is from January to March, quarter 2 is from April to June, quarter 3 is from July to September, and quarter 4 is from October to December. We saw that quarter 1 and 4 steadily remained the highest quarters of $PM_{2.5}$ concentration levels just as quarter 2 and 3 remained the lowest. However, the order between the two groups switched depending on the year.

3. Stationarity

Stationarity can affect the ability to understand and model data. A time series is considered stationary when its statistical properties such as mean, variance and autocorrelation are constant over time.

3.1 Statistical Properties

First, we checked our time series for stationarity by calculating the rolling statistics, also known as moving averages. We calculated and plotted the rolling statistics for mean and standard deviation, seen in **Figure 6**. Since our target population was 24-hour averages, we took the rolling statistics of every 30 days to approximate monthly averages. As we result, our time series was weakly stationary since the rolling mean and standard deviation was relatively constant, with a few shifts in direction.

Figure 4. Yearly Average of 24-hour $PM_{2.5}$ Levels per Season

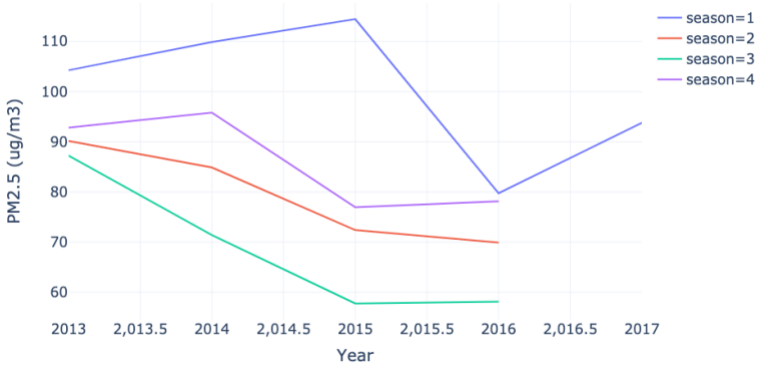
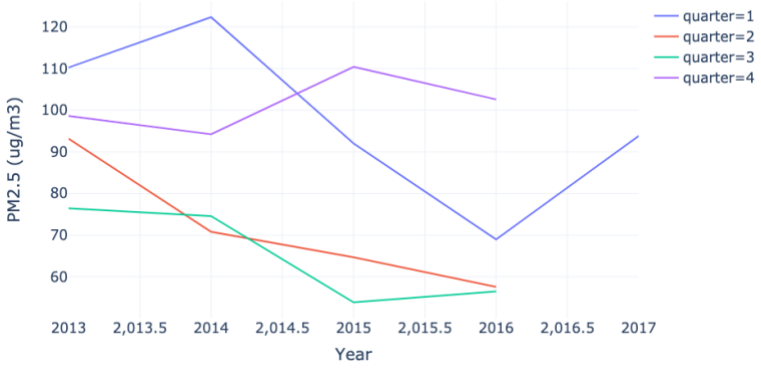


Figure 5. Yearly Average of 24-hour $PM_{2.5}$ Levels per Quarter



Next, we plotted the autocorrelation function (ACF) and partial autocorrelation function (PACF) with 95% confidence intervals, seen in Figure 7. Any values that are outside the confidence intervals suggest a correlation. The ACF shows how a time series is correlated with its past values at different time steps, or lags. The PACF shows how past and future values are related in a time series. It represents the correlation between two points at a time interval, after removing the effects of intervening correlations [cite]. Both our ACF and

PACF plots showed that PM2.5 levels were significantly correlated at lags 1, 2, and 3.

Figure 6. Monthly Rolling Statistics of 24-hour $PM_{2.5}$ Levels

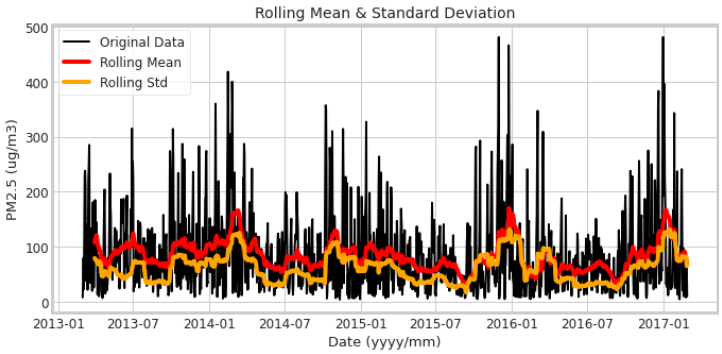
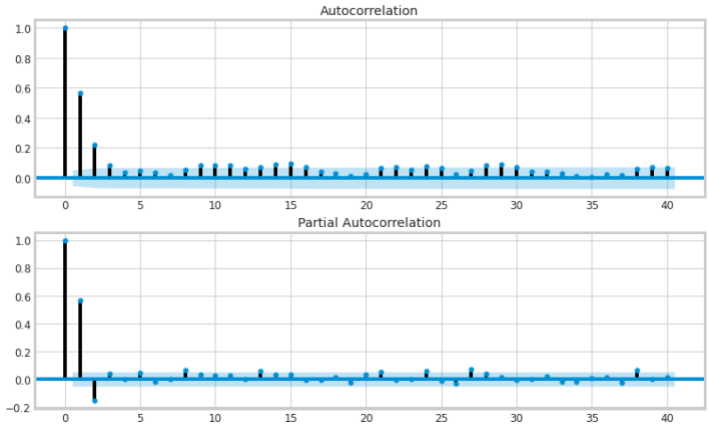


Figure 7. ACF and PACF of 24-hour $PM_{2.5}$ Levels



3.2 Unit Root Tests

A unit root causes random walk, or stochastic trend, random noise has a long-term effect on the time series. Unit root tests such as Augmented Dickey-Fuller (ADF) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) can help determine stationarity. The ADF tests the null hypothesis that a unit root is present with the alternative hypothesis that the series is stationary. The null and alternate hypothesis for the KPSS test is opposite that of the ADF test. For ADF, if the test statistic is less than the critical values and if the p-value is less than alpha level 0.05, we can reject the null hypothesis. These conditionals are inversely true for KPSS [4]. **Table 1** shows the results of our unit root tests.

Table 1. Unit Root Test Results

Unit Root Test	ADF	KPSS
Test Statistic	-1.11e+01	0.44
p-value	3.88e-20	0.06
#Lags Used	7	24
Critical Value (1%)	-3.44	0.74
Critical Value (2.5%)	-----	0.57
Critical Value (5%)	-2.86	0.46
Critical Value (10%)	-2.57	0.35

The ADF test statistic was less than its critical values and its p-value was less than alpha level 0.05. Based on this, we rejected the null hypothesis. The KPSS test statistic was lower than all critical values except the critical value for 10% and its p-value is only 0.01 greater than the alpha level 0.05. Based on this, there was a minor unit root effect in our time series. Since our ADF test resulted in stationary and our KPSS test resulted in non-stationary, this suggested that our time series needs to be differenced.

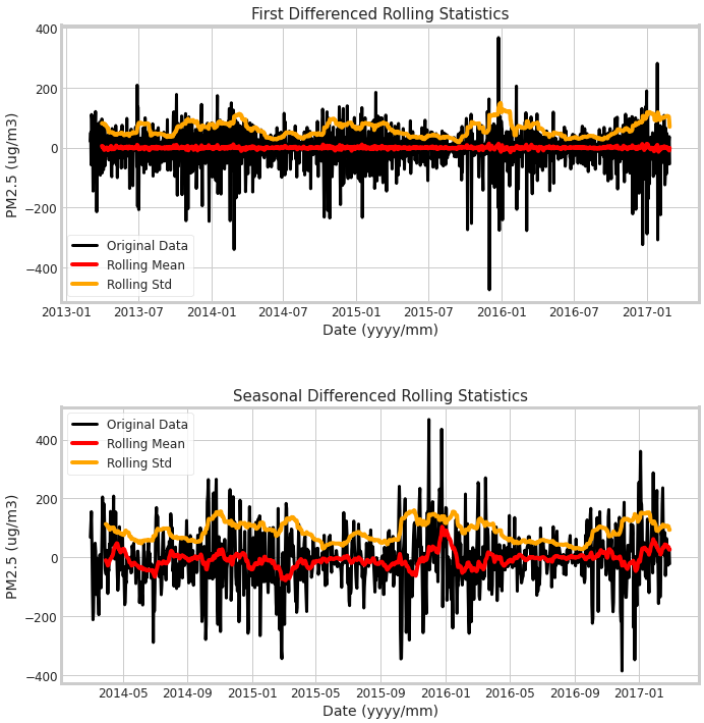
3.2 Differencing

Differencing is a change between consecutive observations in a time series. It is type of transformation can help stabilize the mean of a time series and reduce the effect of trend and seasonality. A first difference is the difference between an observation and the previous observation, while a seasonal difference is similar except that the observations are from the same season [4].

We executed both first and seasonal differencing and ran their unit root tests. For both types of differencing, we got similar ADF results from that of the original time series. For the KPSS test we got more significant results with the first difference. To see how the differencing affected our time series, we plotted their rolling statistics seen in **Figure 8**. We saw that the rolling statistics for the

first difference was more constant about zero than that of the seasonal difference. Subsequently, we achieve a stationary time series after differencing once.

Figure 8. Differenced Rolling Statistics of 24-hour $PM_{2.5}$ Levels



4. Modeling

4.1 Model Selection

We trained and tested three different types of time series models to fit our target population. Firstly, we modeled an Autoregressive Integrated Moving Average (ARIMA) because our series needed to be differenced to be stationary. An ARIMA is a combination of an auto-regressive model (AR) and moving average model (MA). It has three parameters and is written in the form ARIMA(*p*,*d*,*q*). *p* represents the number of AR terms, *d* represents degree of differencing, and *q* represents the number of MA terms. To find the most optimal parameters, we ran a grid search to find the best model based on the Akaike Information Criterion (AIC), an estimator of the out-of-sample prediction error.

Secondly, since our time series had a strong seasonal component, we modeled a Seasonal ARIMA (SARIMA). This model has four additional parameters and is written in the form of SARIMA(*p*,*d*,*q*)(*P*,*D*,*Q*)[*m*]. *P* represents the number of seasonal AR terms, *D* represents the integrated order of the seasonal process, *Q* represents the number of seasonal MA terms, and *m* represents the seasonal length of the series. Similarly to the first model, we ran a grid search to optimize all seven parameters based on the AIC.

Lastly, we used Facebook's open-source model, Prophet. "Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with multiple yearly effects" [5]. We customized this model to fit linear growth, yearly seasonality, quarterly seasonality, weekly seasonality, and China's holidays. The holidays included were: Chinese New Year, Dragon Boat Festival, Labor Day, Mid-Autumn Festival, National Day, New Year's Day, and Tomb-Sweeping Day.

Table 2. Time Series Model AIC Results

Model	AIC
ARIMA(3,1,4)	15970.661
SARIMA(2,1,1)(0,1,1)[12]	15911.823
Prophet(5)	12250.355

Table 2 shows the AIC results of our three models. The ARIMA grid search resulted in a first differenced AR(3) process and MA(4) process. The SARIMA grid search resulted in a first differenced AR(2) and MA(1) process with a seasonal first differenced AR(0) process, MA(1) process and a seasonal length of 12. From the original ARIMA model, the AIC decreased by 58 when seasonality was introduced. However, from the SARIMA, the AIC significantly decreased by 3661 when the other four yearly effects were modeled in Prophet. Therefore, instead of overfitting, Prophet more

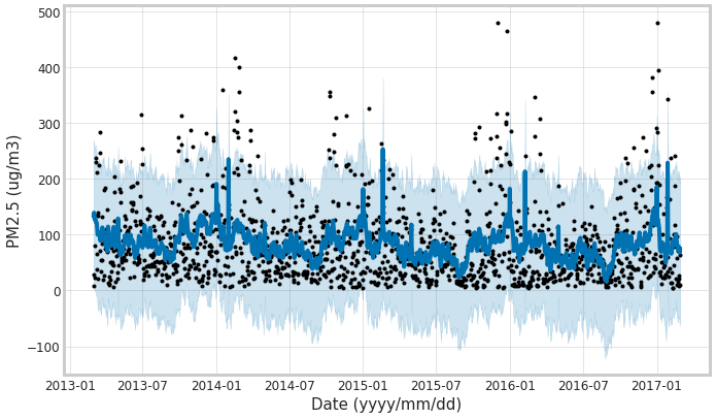
accurately fit the data and benefited from including additional yearly effects.

In addition, recall in our exploratory analysis, we observed a wide range of monthly specific outliers. Prophet is structured to handle outliers well, whereas ARIMA models have difficultly forecasting them. This may have played into the fact that Prophet superseded the other two models.

4.2 Final Model

We chose Prophet as our final model because it had the lowest AIC. **Figure 9** depicts how this model forecasted on the original data from 2013-2017. The black dots represent the actual $PM_{2.5}$ levels, the dark blue line represent Prophet's predictions, and the light blue background represent the 95% confidence interval of those predictions. We saw that prophet predicted majority of the actual values. We also saw that the high confidence interval helped predict the outliers that were not modeled.

Figure 9. Actual vs Forecasted 24-hour $PM_{2.5}$ Levels [2013-2017]



5. Diagnostics

5.1 Residuals

A residual is the difference between the observed value of Y and the estimated value of Y (Y_hat). **Figure 10** depicts the Residuals vs. Fitted plot. The black circles represent the data points. The dotted line that runs across zero on the y-axis is where we want to be. That line means that the expected value is the same as the observed value. The red line is the result of the predicted values of our model against the residuals of our model. In the beginning of the plot, we have straight red line along the dotted line. However, it starts to deviate as we move toward the right. This means that our estimated values some significant differences to the observed values.

We expect the residuals to be normal. **Figure 11** depicts a Normal Q-Q plot. This plot looks at all the residuals and compares them according to a normal distribution, estimating the variance of the residuals. If the errors are normally distributed, they will follow the diagonal dotted line. We can see that only a small measure the variance of the residuals stays close to the dotted line. However, we see majority of them deviating at the tails, which suggest abnormality. Note there is also heteroscedasticity among the data since they are not equally distributed between the top and bottom half of the plot.

Figure 10. Residuals vs Fitted

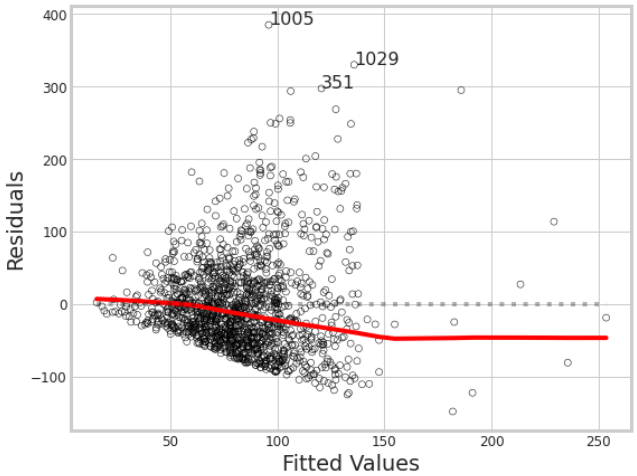


Figure 11. Normal Q-Q

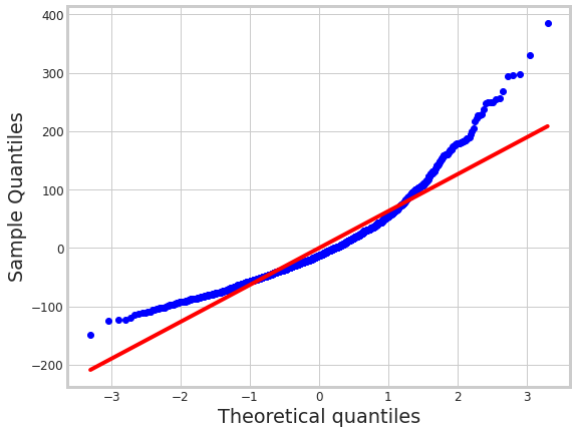
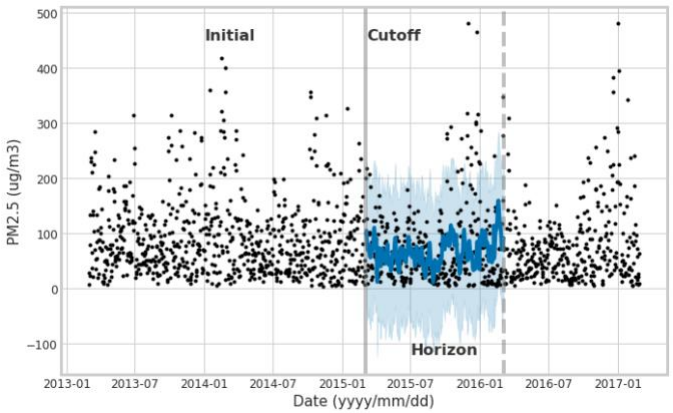


Figure 12. Prophet Cross Validation of 24-hour $PM_{2.5}$ Levels



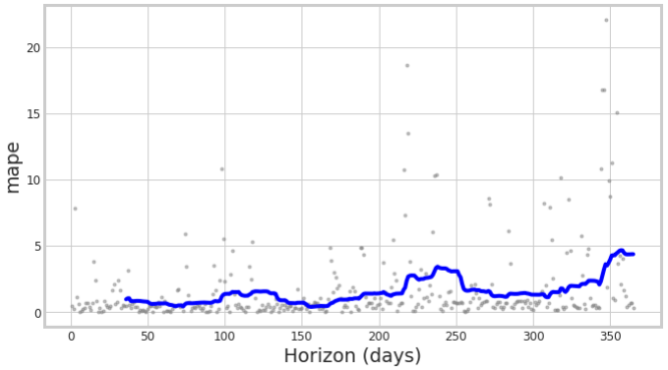
5.2 Error Metrics

Error metrics measures the performance error of a forecasting model. One type of error metric is R-squared, which measures the proportion of the variance explained by the model. Our r-squared was 0.13, which means that our model explains only about 13% of the variance of actual $PM_{2.5}$ levels.

Prophet has a built-in cross validation function which measure forecast error using historical data. To do this, we initially trained the model up until the selected a cutoff point on March 6th, 2015. Then, the model forecasted on a horizon of about a year, the cutoff date to February 2nd, 2016. **Figure 12** depicts our cross validation.

In **Figure 13**, we visualized the cross-validation performance metric MAPE. The dots show the absolute percent error for each prediction in the horizon. The blue line shows the MAPE, where the mean is taken over a rolling window of the dots [5]. For this forecast, we see that errors around 1% are typical for predictions one month into the future, and that errors increase up to around 5% for predictions that are a year out.

Figure 13. Cross Validation MAPE of 24-hour $PM_{2.5}$ Levels from 2015-2016



6. Future Forecasting

6.1 Forecasting Three Additional Years

With our final model, we made future predictions. **Figure 14** depicts how this model forecasted on the original data from 2013-2017, with three additional years from 2018-2020. The gray dashed line indicts where the future predictions start. For future $PM_{2.5}$ concentration predictions, we can see that the model forecasts remain relatively the same with a slow increase in the range of the confidence intervals.

Figure 14. Forecast of 24-hour $PM_{2.5}$ Levels from 2013-2020

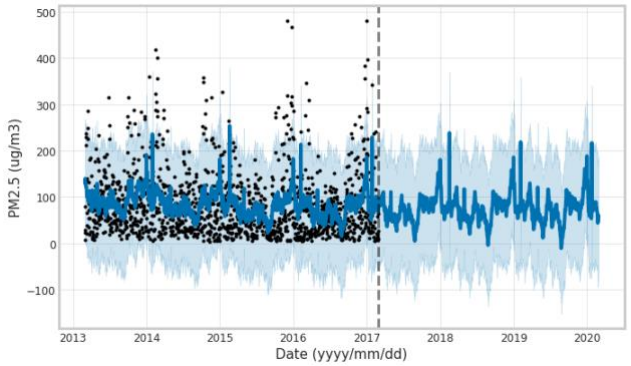
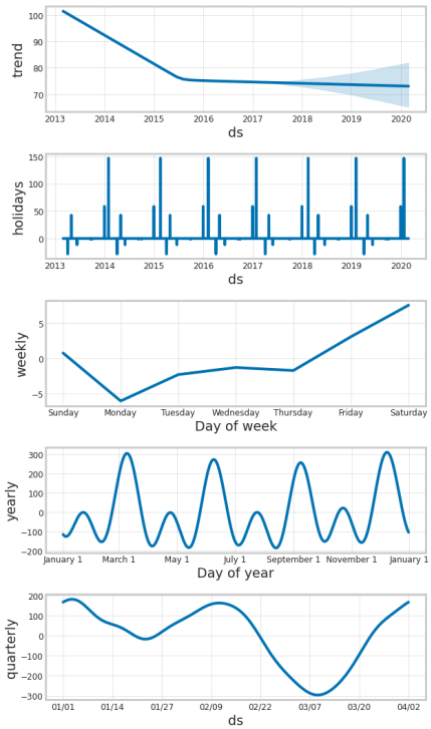


Figure 15 shows how this model's decomposition of trend, holidays, and seasonality. The trend and yearly seasonality look similar to our original time series decomposition, seen in **Figure 3**. However, the uncertainty of our model's trend increases as it forecasts on future data. We can see that weekly and quarterly seasonality have a minor, yet still important, effect in our model, while holidays have a highly a significant effect.

Figure 15. Forecasted Model's Decomposition from 2013-2020



7. Results and Conclusion

7.1 Air Quality Index

The Air Quality Index (AQI) is "a nationally uniform color-coded index developed by the EPA for reporting and forecasting daily air quality" [6]. EPA uses the air quality index to describe the levels of $PM_{2.5}$ concentration.

Figure 16 shows the association between 24-hour $PM_{2.5}$ levels in our time series and their respective AQI. Throughout 2013-2017, we can see that majority of the 24-hour $PM_{2.5}$ levels range from 55-150 and lie within the unhealthy AQI. There are a few cases where the AQI is very unhealthy and even fewer cases where the AQI is hazardous.

Figure 16. AQI of Actual 24-hour $PM_{2.5}$ Levels from 2013-2017

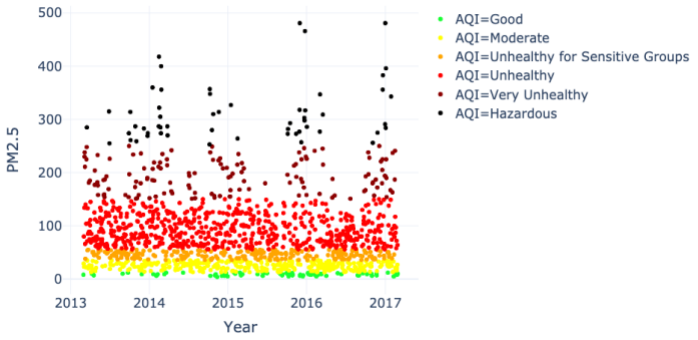


Figure 17. AQI of Forecasted 24-hour $PM_{2.5}$ Levels from 2013-2020

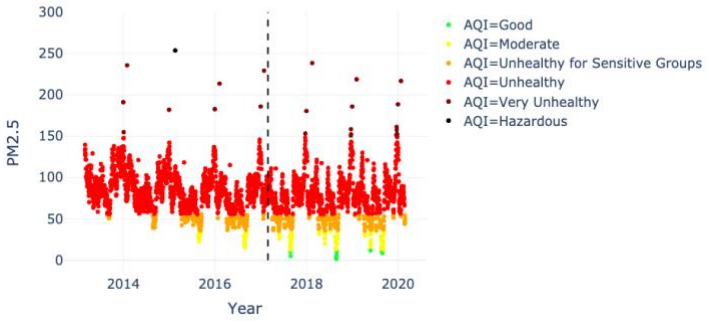


Figure 17 shows the association between our model's forecasted 24-hour $PM_{2.5}$ levels and their respective AQI. The black dashed line indicates where future forecasting begins. From years 2013-2017, our model did a decent job of predicting majority 24-hour $PM_{2.5}$ levels within the unhealthy AQI. It also predicted the few cases of unhealthy for sensitive groups and hazardous. In future forecasting, our model seemed more accurately match our original time series because it predicted a broader variety of the AQI. Therefore, our model was better at predicting more authentic 24-hour $PM_{2.5}$ levels for the future than the past.

7.2 Conclusion

Overall, we found that 24-hour $PM_{2.5}$ levels in Beijing are affected by trend, yearly seasonality, quarterly seasonality, weekly seasonality, and Chinese holidays. Because of these significant effects, we used Facebook's Prophet model. With this model, we were able to accurately forecast past and future 24-hour $PM_{2.5}$ levels.

However, since 24-hour $PM_{2.5}$ levels are dangerous are high levels, further research can be done using quantile autoregression to model the 50% and 75% qualities of the time series.

References

[1] "EPA Collaboration with China." *International Cooperation.* Environmental Protection Agency. 31 Oct. 2016. Web. https://19january2017snapshot.epa.gov/international-cooperation/epa-collaboration-china_.html

[2] Chen, Song Xi. "Beijing Multi-Site Air-Quality Data Set." *UCI Machine Learning Repository.* 20 Sept. 2019. Web. https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data

[3] "Health and Environmental Effects of Particulate Matter (PM)." *Particulate Matter (PM) Pollution*. Environmental Protection Agency. 14 Apr. 2021. Web. https://www.epa.gov/pm-pollution/health-and-environmental-effects-particulate-matter-pm

[4] Hyndman, R.J., & Athanasopoulos, G. *Forecasting: Principles and Practice (2nd Ed)*. 5 May 2021. Web. https://otexts.com/fpp2/

[5] "Saturating Forecasts." *Prophet.* Facebook. 5 Sept. 2020. Web. https://facebook.github.io/prophet/docs/saturating_forecasts.html

[6] "National Ambient Air Quality Standards for Particulate Matter." *Fact Sheet.* Environmental Protection Agency. 7 Dec. 2020. Web. https://www.epa.gov/sites/production/files/2020-04/documents/fact_sheet_pm_naaqs_proposal.pdf