**Air Quality Analysis of PM2.5**

**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 [cite]. In this project, we studied the statistical time series analysis of PM2.5 concentration levels in Beijing to model, forecast, and determine air quality levels.

**1. Introduction**

*1.1 What is PM2.5 and its health risks?*

PM2.5 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, PM2.5 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 PM2.5 concentrations in one year, averaged over three years, is less than or equal to 35 μg/m3”[Cite].

Primary air quality standards provide public health protection. Because of its small size, PM2.5 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. PM2.5 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 [Cite].

*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 PM2.5 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 PM2.5 concentration levels (in μg/m3) from March 2013 to February 2017.

In this report we will explore, model, and forecast daily PM2.5 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 Descriptive Statistics*

To understand the distribution of our data values, we computed its descriptive statistics. The mean was 83.47 μg/m3, representing the average of 24-hour PM2.5 concentration levels within the target population. The median was 64.63 μg/m3, 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 μg/m3. 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 PM2.5 concentration levels ranged from 4.29 to 481.29 μg/m3, which gave a difference of 477 μg/m3. Variance and standard deviation measure how dispersed the data values are around the mean. The target population had a variance of 4990.04 (μg/m3)2, with a standard deviation of 70.64 μg/m3.

**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 μg/m3. This meant that any values greater than 226.67μg/m3­ were outliers. However, **Figure 2** shows that the outliers differ by month. From months 1-3 and 10-12, outliers start around 300 μg/m3­. For months 4-9, outliers start around 200 μg/m3­. This meant that outliers vary monthly throughout the year.

**Figure 1. Density Plot of 24-hour PM2.5 ­Concentration Levels**

Chart, histogram

Description automatically generated

**Figure 2. Box and Whisker Plot of 24-hour PM2.5 Levels Indexed by Month**

Chart, box and whisker chart

Description automatically generated

*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 PM2.5 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 PM2.5 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 μg/m3­. 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 PM2.5 concentration levels. This confirmed what we presumed in our observed plot: 24-hour PM2.5 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 [cite].

**Figure 3. Time Series Decomposition of 24-hour PM2.5 Levels**

**Graphical user interface, chart

Description automatically generated**

*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 PM2.5 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.

**Figure 4. Yearly Average of 24-hour PM2.5 Levels per Season**

Chart, line chart

Description automatically generated

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 PM2.5 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 PM2.5 concentration levels just as quarter 2 and 3 remained the lowest. However, the order between the two groups switched depending on the year.

**Figure 5. Yearly Average of 24-hour PM2.5 Levels per Season**

Chart, line chart

Description automatically generated

**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 6. Monthly Rolling Statistics of 24-hour PM2.5 Levels**

Chart

Description automatically generated

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. This

**Figure 7. ACF and PACF of 24-hour PM2.5 Levels**

Graphical user interface, application, table, Excel

Description automatically generated

*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 [cite]. **Table 1** shows the results of our unit root tests.

**Table 1. Unit Root Test Results**

|  |  |  |
| --- | --- | --- |
| Unit Root Test | Augmented Dickey-Fuller | Kwiatkowski-Phillips-Schmidt-Shin |
| 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 [cite].

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 PM2.5 Levels**

**Chart

Description automatically generated**

**A picture containing chart

Description automatically generated**

**4. Modeling**

There Keep this section factual, don’t draw too many conclusion, just report results. Resampling strategy (if applicable). Data Transformation. Modeling Technique. Results.

There is a clear discussion of the methods used to analyze the data. Data transformations are discussed specifically. Results are shows in tables and graphs that are clear and answer the questions. Diagnostic methods are shown. P values and/or predictions are shown.

We want to predict grid stability. This type of a response refers to a binary response, which is a special case of the binomial that requires the response have only two levels. In this case, our two response levels are stable and unstable.

*4.1 Model Selection*

We tested three different types of time series models. 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 yearly, weekly, and daily seasonality, plus holiday effects” [cite].

Superceeds

Prophet is know it handle outliers well.

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

To model grid stability, we first created a generalized linear model (GLM) by fitting all independent attributes as predictors with num\_stabf as the response. GLM requires us to specify the family, a description of the error distribution and link function to be used in the model. In this case, we used the binomial family with a logit link function. From the summary function of the model, we can see that reaction time and price elasticity are highly significant, whereas power balance is insignificant.

We used the step function to look at each predictor in respect to the model's AIC. It removes a variable one at a time and refits the model until it produces the lowest AIC possible. This function drops all attributes for power balance. The resulting model only uses the attributes of reaction time and price elasticity, yielding an AIC value of 7835.6. To further confirm this result, we ran an anova chi square test that compared the original model to the resulting model. We see that the p-value is 0.2684, which greater than 0.05. This suggests that we can drop the attributes for power balance as predictors during the fitting of the model.

*3.2 Final Model*

Our final model includes the producer and consumers reaction times as well as their price elasticity. All of the predictors have a low p-value of 2e-16, suggesting high significance. The estimate of the intercept was 11.78026, beta\_0. This is the value of Y when all of the X\_i coefficients are zero. The beta\_i (i=1-8) represents the slope, which is negative for each beta. This means that for every unit increase in X\_i, we except stability to go down by beta\_i.



We were able to produce a 95% confidence intervals for the beta value. This means that if we repeat the experiment multiple times, we expect that 95% of the confidence intervals contain the true value of beta\_i.

Text

Description automatically generated

**4. Diagnostics**

*4.1 Residuals*

A residual is the difference between the observed value of Y and the estimated value of Y (Y\_hat). **Figure 3** depicts the Residuals vs. Fitted plot. 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. We can see that we have a relatively straight red line along the dotted line. This means that our estimated values have very little differences to the observed values.

**Figure 3. Residuals vs Fitted**

Chart, line chart

Description automatically generated

We expect the residuals to be normal. **Figure 4** 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 majority of the variance of the residuals follow the dotted line. However, we see slight deviations on 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 4. Normal Q-Q**

Chart, line chart

Description automatically generated

We also plotted the residuals of individual predictors. **Figure 5,** shows both reaction time and price elasticity against the residuals. When looking at the left plot, we see that the residuals vary as reaction time increases. Regardless, this curve is not drastic and stays relatively in a line. When looking at the right plot, we see that the residuals are consistent as price elasticity increases, reflecting a linear relationship.

**Figure 5. Predictors against Residuals**

Chart, line chart

Description automatically generated

**5. Predictions**

*5.1 Calibration Curve*

We want our model to be well calibrated. For example, if we assign 10% probability of stability and look at 100 grid frequencies, ideally 10% of them would be stable. A calibration curve plots predictions on the x-axis and true values on the y-axis. Just like our residual plots, we want our line to follow the dotted diagonal line that goes through (0,0). **Figure 6** depicts our model’s calibration curve.

For the most part, our line follows the dotted line, with a small underestimations in the middle and overestimations on the tails. For example, the actual values at 0.63 mean that a set of frequencies have a 63% of being stable. However, at 0.63, our model predicts stability at around 58%.

**Figure 6. Predictors against Residuals**

Chart, line chart

Description automatically generated

*5.2 Confusion matrix*

A confusion matrix is a table used to describe the performance of a classification model where the true values are known. It compares the true values with the predicted values from the model. **Table 1** shows the results of our confusion matrix with a cutoff of 0.5.

**Table 2. Confusion matrixc**

|  |  |  |
| --- | --- | --- |
| n = 10000 | Predicted |  |
| Actual | stable | unstable |
| stable | 2547 | 1073 |
| unstable | 778 | 5602 |

From this confusion matrix, we were able to quantify the accuracy of our model. Accuracy is the number of correctly predicted cases divided by the total number of cases. In this case, our model is 81.49% accurate. However, accuracy can be misleading in an imbalanced dataset and as we mentioned before, the percentage of stability is at 36.2%. Because of this, we can look to other classification quantifiers.

Specificity answers the question: of those that are unstable, how many were correctly predicted as unstable? Our model has a specificity of 87.81%. Sensitivity answers the question: of those that are stable, how many were correctly predicted as stable? Our model has a sensitivity of 70.36%. Positive predicted value (PPV) answers the question: of those predicated stable, how many were actually stable? Our model's PPV is at 76.6%. Negative predicted value (NPV) answers the question: of those predicated unstable, how many were actually unstable? Our model’s NPV is at 83.93%.

***Accuracy*** = (2547+5602)/(10000) = .8149.

***Specificity*** = 5602/(5602 + 778) = .8781

***Sensitivity*** = 2547/(2547 + 1073) = 0.7036

***PPV***= 2547/(2547 + 778) = 0.7660

***Negative predicted value*** = 5602/(5602 + 1073) = 0.8393

*5.3 Odds Ratio*

An odds ratio expresses relative difference, an alternative scale to probability for representing chance. If odds ratio is more than 1, there is a greater likelihood of having the outcome. If odds ratio is less than 1, there is a lesser likelihood of having the outcome. For this type of result, you need to subtract it from 1 to get the percentage. We calculated the odds ratio of the coefficients by exponentiating beta\_i, results seen in **Table 3**.

**Table 3. Odds Ratios of model coefficients**

|  |  |  |  |
| --- | --- | --- | --- |
| Coefficient | Odds Ratio | Coefficient | Odds Ratio |
| Intercept | 1.306471e+05 |  |  |
| tau1 | 7.303535e-01 | g1 | 6.790042e-02 |
| tau2 | 7.243105e-01 | g2 | 5.420788e-02 |
| tau2 | 7.272058e-01 | g3 | 4.410557e-02 |
| tau3 | 7.174766e-01 | g4 | 5.759710e-02 |

Looking at just the producer coefficients, we can see that both reaction time (tau1) and price elasticity (g1) have an odds ratio of less than 1. Once we subtracted this value from 1, we found that odds of the grid being stable is 27% less likely with every unit increase in tau1 and 93% less likely with every unit increase in g1.

We then looked to answer the question: what is the difference in odds for testing stable when producer reaction time (tau1) goes from the 1st quartile to 3rd quantile? For tau1, the difference between the 1st and 3rd quantile is 4.75. We multiplied this difference to the tau1 estimate and 95% confidence intervals (CI) and then exponentiated it to get values in an odds ratio scale. The results seen in **Table 4**.

**Table 4. Difference in Odds Ratios from 1st-3rd quantile**

|  |  |  |  |
| --- | --- | --- | --- |
| Coefficient | Conf. low | Estimate | Conf. high |
| tau1 | 0.2022387 | 0.2248071 | 0.2495402 |
| g1 | 0.2513817 | 0.2787664 | 0.3087614 |

The odds ratio estimate was 0.22 and we are 95% confident that the true value is between the CI. We subtracted the estimate from 1 to get 0.78. This means that the odds of the grid being stable is 78% less likely when tau1 goes from 1st quantile to the 3rd quantile. We executed the same procedure for g1, and found that the odds of the grid being stable is 72% less likely when g1 goes from 1st quantile to the 3rd quantile.

*5.4 Probability*

Probability reflects the chance or likelihood a particular event will occur. In this case, we looked at the event when all predictors were at their mean value except for one. The first predictor we changed was tau1. For event 1, we made tau1 at a value of its 1st quantile, 2.87. Using our model, we predicted the probability of stable for this event to be 42%. For event 2, we made tau1 at a value of its 3rd quantile 7.62. Using our model, we predicted the probability of stable for this event to be 14%.

The second predictor we changed is g1, while all other predictors were at mean value. For event 1, we made g1 at a value of its 1st quantile, 0.28, and found the probability of stable to be 39%. For event 2, we made g1 at a value of its 3rd quantile, 0.76, and found the probability of stable to be 16%.

**6. Results and Conclusion**

*6.1 Summary of Results*

All of the plots and figures from the results section are discussed and interpreted. The coefficients are interpreted. Confidence intervals/prediction intervals are shown. P value are appropriately interpreted. Diagnostic plots are discussed.

The calibration curve difference is around 5%. Though this difference maybe small, any deviations in frequency can lead to a black out.

*6.2 Conclusion and Implications*

Issues/inconsistencies that are in the data/diagnostic plots are discusses. Next steps/improvements are discussed.

Discussion: Why did you choose the model formulation you did? What did you encounter along the way to choosing it? What did you learn about the data set? Which variables are important? How could analysis be improved (e.g. more data, a different predictor, other modeling techniques?)

**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/>

[3] "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>

[4] Schäfer, Benjamin, Moritz Matthiae, Marc Timme, and Dirk Witthaut. *Decentral Smart Grid Control*. Research Paper. IOP, 2015. Web. <https://iopscience.iop.org/article/10.1088/1367-2630/17/1/015002>

[5] Arzamasov, Vadim, Klemens Böhm, and Patrick Jochem. *Towards Concise Models of Grid Stability*. IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids, 2018. Web. <http://dbis.ipd.kit.edu/english/2635.php>