Exploratory Data Analysis and Forecasting of US and California Air Quality using SARIMAX and TBATS Time Series Approach

Angela Cao

Briarcliff High School

Acknowledgements

I would like to thank to my mentor Mr. Johnny Ma who was willing to provide assistance for my research when I needed help the most after losing my scheduled summer laboratory internship due to the Pandemic. I would also like to thank Mrs. Annmarie O'Brien, the best science research teacher who has supported me throughout the process.

Table of Contents

Table of Figures

Abstract

Air pollution has been a persistent problem impacting citizens' health and the environment. Therefore, the ability to model, predict, and monitor air quality is relevant and necessary, especially in urban cities. This study investigated sixteen years of US air quality data first through an exploratory data analysis and then focused on data from the state of California only, building and comparing models for air quality forecasting. The exploratory data analysis built up a basic understanding of the dataset and identifies general and interesting patterns. Multiple naïve and advanced time series forecasting methods were used to provide a comprehensive forecast of the levels of four air pollutants (SO_2, NO_2, O_3, CO).

It was found that over the sixteen years, the CO, SO_2, and NO_2 pollutants had downward trends while O_3 fluctuated at a high level. There was a 0.67 correlation between NO_2 and CO levels in California, and these two pollutants were most correlated in the winters and least in the summers. Yearly seasonality was observed for each pollutant, with O_3 having a seasonal pattern with peaks opposite of the rest. Both advanced models of SARIMAX and TBATS fitted well with SARIMAX fitting better for NO_2 and O_3, and TBATS fitting better for CO and SO_2. The applications, limitations, and future potentials of the two models are also discussed. For future research, a plan is to incorporate data of meteorological/human-caused parameters from the National Climate Data Center or California Transportations to compare the magnitudes of influences and improve the accuracy of the models. Predicting air quality and better understanding air pollution are complex yet necessary tasks that require sustained attention from scientists around the world for the sake of humanity and environment.

1. Introduction

Despite the establishment of the United States Environmental Protection Agency (EPA) in 1971 and its constant effort to combat all types of pollution, air pollution in the United States continues to harm the citizens' health and environment. According to the EPA, air pollution levels in many areas exceed the National Ambient Air Quality Standards (NAAQS) established under the authority of the Clean Air Act for at least one of the six common pollutants. These six "criteria air pollutants" include SO_2, NO_2, O_3, CO, lead, and particulate matter.[3]

SO_2 and NO_2 are two pungent gases that are associated with increased respiratory symptoms, diseases and even premature death. Both can irritate the respiratory pathway and are most dangerous towards people with asthma or similar pre-existing conditions. Elevated CO levels outdoors are of particular concern for people with some types of heart disease, leading to reduced oxygen to the heart, angina (chest pain), and even death. Ground level Ozone (O_3) is different from the protective ozone layer in the upper atmosphere; it is the main ingredient in "smog" and can lead to numerous respiratory diseases. All four of the air pollutants mentioned also affect sensitive vegetations and ecosystems, leading to leaf damage and increased susceptibility to diseases.[4] This study does not involve lead and particulate matter (PM) data, but these air pollutants are also immensely harmful as well. To put this problem of air pollution into perspective, it was estimated that between 90,000 to 360,000 deaths per year in the US are linked to air pollution and that air pollution-related illnesses cost approximately $37 billion each year in the US, with California alone costing $15 billion.[5,6]

The main general causes associated with air pollution are the burning of fossil fuels, agriculture, exhaust from factories and industries, residential heating, and natural disasters.[7] For the western US states such as California, there are a few more specific factors such as Asian air pollution emissions, wildfires, and methane. Air pollution from China, India and several other Asian countries has wafted across the Pacific Ocean over the past years, drastically increasing levels of "smog" in the western states such as California, the state the latter half of this study focuses on. In fact, rising Asian emissions contributed to as much as 65% of the western US O_3 increase from 1980 to 2014. Other factors such as wildfires and methane from livestock contributed much less with wildfire emissions supplying less than 10% and methane supplying about 15% during these 34 years[8,9]. In more recent times, California wildfires are becoming more threatening. Furthermore, as the most populated state in the US with the fifth largest economy in the world, California faces significant sources of air pollution such as automobile traffic and industrial pollution. In addition, the local topography in the state and its warm weather trap pollution within valley walls, increasing the levels of O_3. All of these factors combined put California consistently

at the top of the list in the American Lung Association's State of the Air Reports as the worst air quality in the country.[10,11]

In the face of increasingly serious air pollution problems, scholars have conducted a substantial amount of related research using statistical models and analysis, ARIMA forecasting, Box–Jenkins time series models and nonlinear regression.[12,13,14,15,16] In many of these previous studies, the levels of air pollution are measured through Air Quality Index (AQI), the standard indicator of air quality that allow comparison between different pollutants. The ability to accurately forecast the levels of air pollutants continues to increase in importance. It has a crucial role in understanding the change in air quality over time, measuring the effectiveness of federal and regional environmental policies, and managing air quality to protect both human health and the environment.

The present research investigates sixteen years of US air quality data first through an exploratory data analysis and then focuses on air quality data from California, building and comparing models for air quality forecasting. The exploratory data analysis includes a brief correlation analysis looking for possible correlation patterns between different air pollutants, which is something barely studied before. Multiple naïve and advanced time series forecasting methods are used to provide a comprehensive forecast of the four pollutants' levels, allowing comparisons between different methods and more insights to be derived from such comparisons.

2. Statement of Purpose

This study will investigate US air quality data first, and then narrow down to focus on air quality data from California only. As suggested above, California has one of the worst air qualities out of all the states. The first half of this study aims to identify patterns within the US air quality data and derive insights through an exploratory data analysis, while the second half aims to build models for air quality forecasting for the state of California, using and comparing between three naïve methods, an improved version of the Seasonal ARIMA model, and a TBATS model that was introduced to R in 2011 and to Python in 2019.

Three Groups of Research Questions:

1. Preliminary Research: Is the overall US air quality worsening? If so, which air pollutants are to blame? Are there any seasonality patterns regarding US air quality and California air quality?
2. Correlation Analysis: Is there a correlation between the levels of any two of the four air pollutants studied? If so, is there a seasonality pattern of the said two correlation levels (tested through correlation coefficient)?
3. Time Series Analysis: What are the trend and seasonal components of each of the four air pollutants? What insights can be drawn from them?

3. Materials and Methods

3.1 Dataset[1,2]

The dataset used in this study was scrapped from the US EPA's Air Data by a creator on Kaggle, an online community of data scientists and machine learning practitioners. The creator Brenda So gathered information of the four major pollutants (SO_2, NO_2, O_3, CO_2) for every day from 2000 to 2016 and stored the information in a CSV file. There was a total of 28 fields. The four pollutants each had 5 specific columns, one column being each pollutant's AQI, which is the measurement of air quality used in this study. Observations totaled to over 1.4 million rows.

3.2 Tools

The programming language used in this study was Python through the Anaconda distribution and package management platform. All of the coding was done on Jupyter Notebook, a web-based interactive data science environment that acts as an Integrated Development Environment and allows data visualizations. Some Python packages involved in this study were Numpy, Matplotlib, Pandas, SciPy, statsmodels, pmdarima, and TBATS.

3.3 Data Cleaning

The dataset contained raw data with unnecessary attributes and null values, aspects that should be checked before the exploratory data analysis. First, it was checked that the data columns and entries are the expected numbers. Then, columns not relevant to this particular study such as "State Code" and "Address" were dropped. Additional statistical estimators and attributes such as "modes" and "duplicates" were checked with a custom function. It was then found that about half of the dataset had null or missing values for two variables—SO_2 and CO AQI. This was because NO_2 and O_3 AQI values were collected four times a day, while SO_2 and CO AQI values were collected twice a day. The rows without SO_2 and CO AQI values were dropped because the study focused on all four AQIs, and this did not disturb the dataset's overall stability, allowing the coding to continue.

All the attributes were checked once more to achieve greater familiarity with the dataset and remove certain confusing aspects. For example, it was found that California had the most data out of the whole dataset according to the percentages of data from each state. Also, "The Country of Mexico" was dropped from the "states" in the dataset.

Assuming that the data is drawn from a normal distribution—an assumption made because visually the shapes looked relatively normal, the outliers lower or higher than the 5^{th} or 95^{th} percentile

respectively can be removed to restrict analysis to the most common portion of data. The distributions looked approximately normal, with the CO and SO_2 AQI curves looking slightly more skewed to the right than the other two curves.

The last step was to prepare the Pandas data frame for analysis. The date columns were changed from strings into a specific, usable format called "datetime." Year and month columns were created.

3.4 Exploratory Data Analysis

An exploratory data analysis was conducted to build up basic understanding of the dataset for more advanced research as well as answer the preliminary research questions. Data visualizations were key to summarizing the main characteristics of this dataset. The medians of each date's and each year's observations were taken for all four AQIs and first plotted against states, then against time to visualize the change of the four AQIs over time. Median, instead of mean, was used as the main point estimator throughout this research because it was more resistant to skewness and knowing that CO and SO_2 AQI distributions were slightly skewed, a more conservative approach was taken. Heatmap pivots with states as rows, years as columns, and the four AQIs as values were constructed to summarize the historical, regional changes of the US air quality data over time.

In addition to data visualizations, a correlation analysis was performed to see if there was a possible correlation between the levels of any two of the four AQIs, something that may suggest some common factors influencing two AQIs at once. A general correlation matrix was first constructed; then, potential correlations between CO and NO_2, SO_2 and NO_2 respectively were tested for by further controlling elements such as time and/or region. In this study, the region for correlation analysis was restricted to California. Graphs were also created to look for patterns of any change of correlation over time, with correlation measured through correlation coefficients.

Potential seasonality (yearly, monthly, and weekly) was also checked for in order to enable time series analysis.

3.5 Time Series Analysis

After an exploratory data analysis, it was decided that the more advanced time series analysis should focus only on data from California because it had the most amount of and most complete data over the 16 years, as well as the most serious air pollution problem. An analysis of the four AQI time series was necessary before building forecasting models. Prediction is an expectation for a combination of predictors and forecasting is a special type of prediction based on previous values. The data was first resampled to have one value only for every day, which is the median of each day's observations. This way, the four AQI data became time series, a sequence where a metric is recorded over a regular, daily interval. After

reformatting the data frame to have only four AQI columns and the date column, it was decomposed with an additive model to derive insights on each AQI time series' trend, seasonal, and residual components and prepare for a time series forecasting.

3.6 Time Series Forecasting

A 15-year train set (2000-2015) and a 1-year test set (2016) were created in order to train a model that can eventually predict a year into the future. A custom function was used to directly calculate and store accuracy measures such as mean squared errors for each method into a neat data frame. A second function was used to plot the 1-year forecasts for each AQI onto the actual values of 2016 to see how successful each forecasting method was and derive insights.

The naïve methods of forecasting were tested first because the most simplistic methods can work the best in some scenarios and the results of naïve methods allowed for comparisons with the more sophisticated methods later. The three naïve forecasting methods used were the average method in which the forecast value is simply the average of past values, the median method in which the forecast value is simply the median of the past values, and the seasonal naïve method in which the forecast value is the same value as the corresponding value in the last seasonal period.

Then, two advanced models for time series forecasting were applied and their accuracy measures were stored in the same data frame. The first was an improved seasonal version of a commonly used statistical model called ARIMA, short for "Auto Regressive Integrated Moving Average." An ARIMA "explains" a given time series based on its own past values i.e., its own lags and the lagged forecast errors. It is characterized by 3 terms, p as the order of the "Auto Regressive" (lag) term, q as the order of the "Moving Average" (forecast errors) term, and d as the number of differencing required to make the time series stationary. A time series is stationary when its values are not a function of time, meaning that the statistical properties of the series such as mean, variance, and autocorrelation are constant over time. It is necessary to make the time series stationary because linear regression statistical models work best if the predictors—lags of the series—are nearly independent. The equation for an ARIMA model is as below:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$

In words, an ARIMA model is: Predicted = Constant + Linear Combination Lags of Y up to p lags + Linear Combination of Lagged forecast errors up to q lags.[17]

However, general Seasonal ARIMA (SARIMA) models are not designed to work on data with yearly seasonality patterns. Even though the arima( ) function can allow a seasonal period up to m=350 in theory, it will usually run out of memory whenever the seasonal period is more than about 200 in practice. In order to include a long seasonal period of 365, a possible way is the fourier series approach where the seasonal pattern is modelled using Fourier Terms with short-term time series dynamics allowed in the

error.[18] Fourier Terms is based off Joseph Fourier's idea that adding a certain amount of different sine and cosine functions can compose any wave function. The equation is as below:

$$y_t = a + \sum_{k=1}^{K} [\alpha_k \sin(2\pi kt/m) + \beta_k \cos(2\pi kt/m)] + N_t,$$

where N_t is an ARIMA process and K is chosen by minimizing the Akaike information criterion (AIC), a mathematical indicator that evaluates how well a model fits the data it was generated from

Therefore, the model used in this study was an altered version of SARIMAX (the X representing an inclusion of exogenous factors), since it used Fourier Terms as the exogenous factors.

The second model used in this study was a newer, rougher, and more flexible model called TBATS. Its name includes acronyms for key features: Trigonometric seasonality, Box-Cox transformation, ARMA errors, Trend and Seasonal components. This model is developed by De Livera, Hyndman, and Snyder. It takes its roots in exponential smoothing methods and considers these various alternatives.[19, 20] It can be described by the following equations:

Model:

$$y_t^{(\lambda)} = l_{t-1} + \phi\, b_{t-1} + \sum_{i=1}^{T} s_{t-m_i}^{(i)} + d_t$$

$$l_t = l_{t-1} + \phi\, b_{t-1} + \alpha\, d_t$$

$$b_t = \phi\, b_{t-1} + \beta\, d_t$$

$$d_t = \sum_{i=1}^{p} \varphi_i d_{t-i} + \sum_{i=1}^{q} \theta_i e_{t-i} + e_t$$

Seasonal part:

$$s_t^{(i)} = \sum_{j=1}^{(k_i)} s_{j,t}^{(i)}$$

$$s_{j,t}^{(i)} = s_{j,t-1}^{(i)} \cos(\omega_i) + s_{j,t-1}^{*(i)} \sin(\omega_i) + \gamma_1^{(i)} d_t$$

$$s_{j,t}^{*(i)} = -s_{j,t-1}^{(i)} \sin(\omega_i) + s_{j,t-1}^{*(i)} \cos(\omega_i) + \gamma_2^{(i)} d_t$$

$$\omega_i = 2\pi j / m_i$$

Where:

$y_t^{(\lambda)}$ - time series at moment *t* (Box-Cox transformed)

$s_t^{(i)}$ - *i*th seasonal component

l_t - local level

b_t - trend with damping

d_t - ARMA(*p*,*q*) process for residuals

e_t - Gaussian white noise

Model parameters:

T - Amount of seasonalities

m_i - Length of *i*th seasonal period

k_i - Amount of harmonics for *i*th seasonal period

λ - Box-Cox transformation

α, β - Smoothing

ϕ - Trend damping

φ_i, θ_i - ARMA(*p*, *q*) coefficients

$\gamma_1^{(i)}, \gamma_2^{(i)}$ - Seasonal smoothing (two for each period)

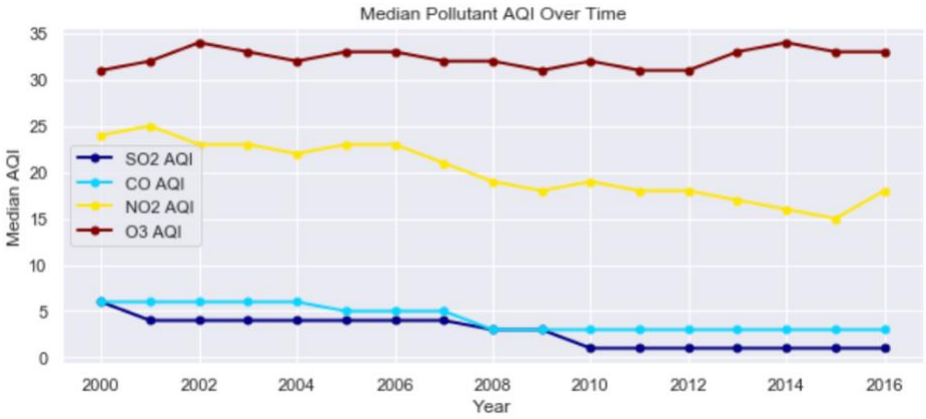
4. Results and Discussion

4.1 Data Visualizations and Patterns

The median values grouped by each year for the four air pollutant AQIs changed over time (See Figure 1 below). O_3 AQI fluctuated but remained within the range of 30 to 45 values in the 16 years. NO_2

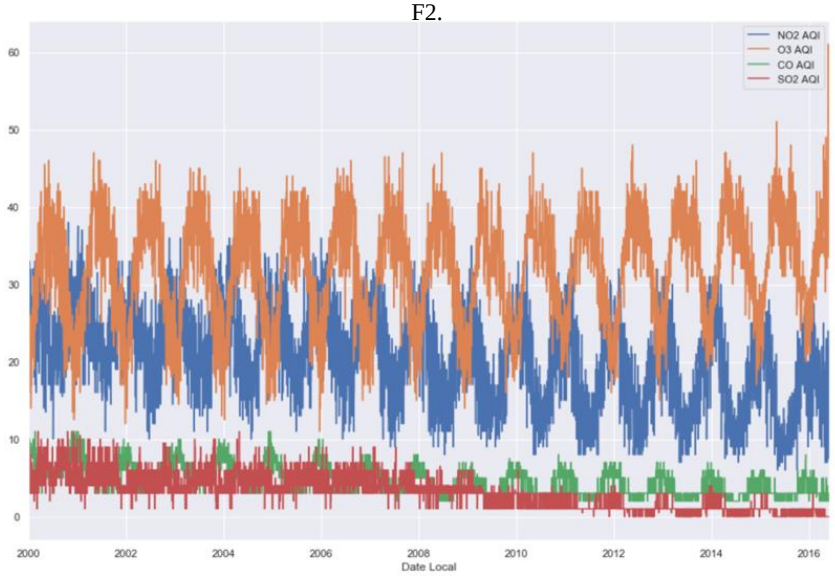
AQI decreased from about 24 values to about 18 values. CO AQI decreased from about 6 values to about 3 values. SO_2 AQI decreased from about 6 values to about 1 value. In a general and national sense, the AQIs of CO and SO_2 pollutants had clear downward trends and both maintained a small value close to zero in around 2008 to 2010. NO_2 had been decreasing but seemed to be possibly switching directions in 2016, something that can be easily confirmed or disproved with additional data from 2016 to 2020. O_3 AQI remained at a high level and had no decreasing trend, which confirmed previous statistics showing that O_3, along with PM that is not part of this study, is responsible for a large portion of air-pollution relevant deaths.

F1.



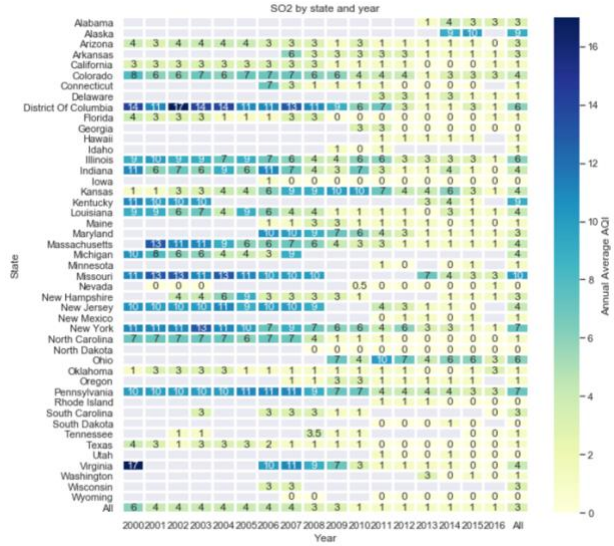
Regarding seasonality, CO, NO_2 and O_3 AQIs clearly had yearly patterns, while SO_2's yearly pattern looked noisier (See Figure 2 below). O_3 AQI peaked in July and August likely because O_3 is produced by chemical reactions involving sunlight and is correlated to traffic volumes, which are often highest in the summer. O_3's pattern behaved in the opposite way of CO and NO_2. Monthly and weekly patterns were also checked. There seemed to be a slight monthly pattern in CO. Other than that, minimal monthly and weekly patterns were observed through graphs.

F2.

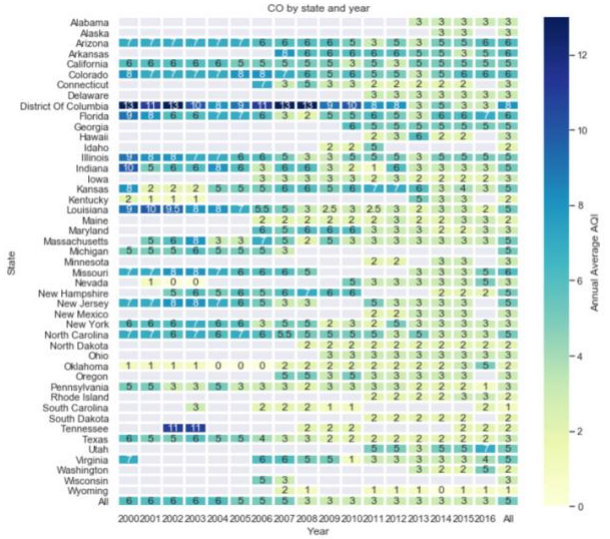


Heatmaps below showed the changes of median values of the four air pollutant AQIs in each state (See Figures 3-6 below). For example, Florida's NO_2 level decreased from about 23 to about 14 through the sixteen years.

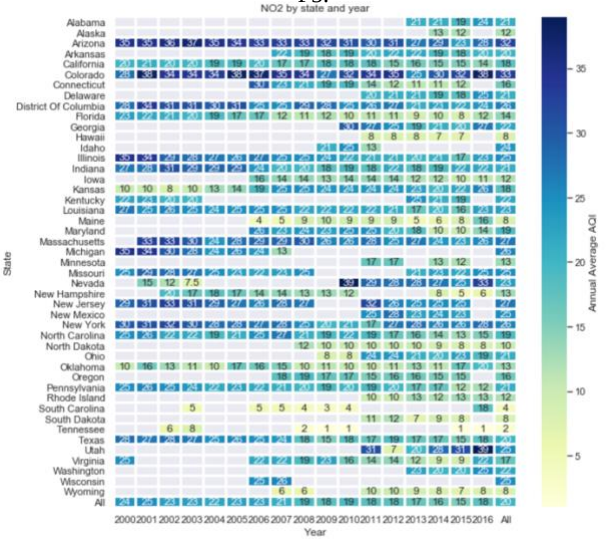
F3.



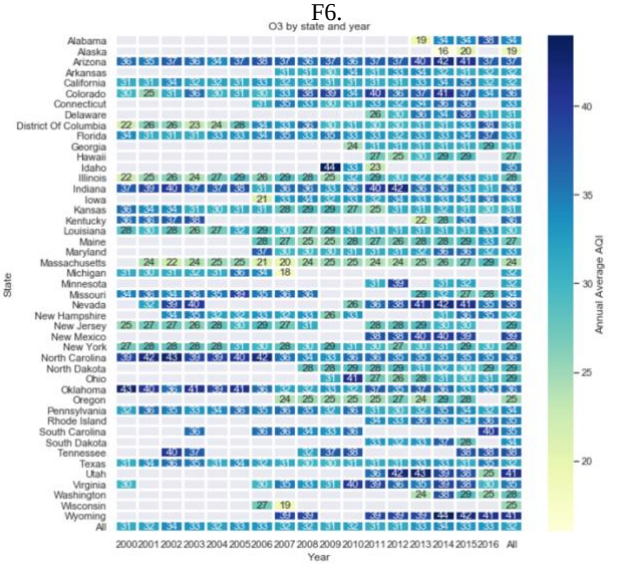
F4.



F5.

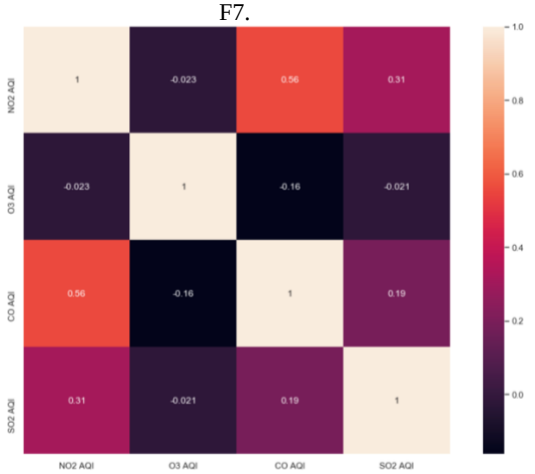


F6.

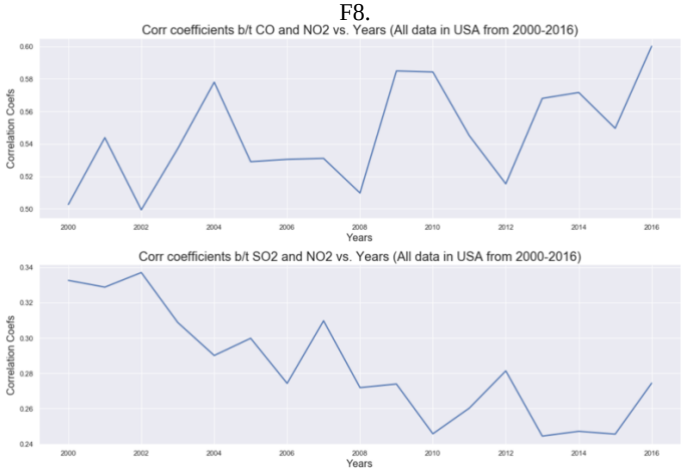


Using the data of all sixteen years and from every state in the US, it was found that there is a 0.56 national correlation between NO_2 and CO and a 0.31 correlation between NO_2 and SO_2 (See Figure 7 below, on the left). It was suspected that the actual correlations between these air pollutant levels (in the same region at the same time) are higher than these observed values due to the fact that the data has not controlled elements such as time and region. As per the general correlation matrix in Figure 7, the correlations between any two other AQI levels are likely not high enough to derive any insights. This is confirmed later as even when the data used is from the same month in the same state, none of the other correlations raised to 0.30 or above.

F7.



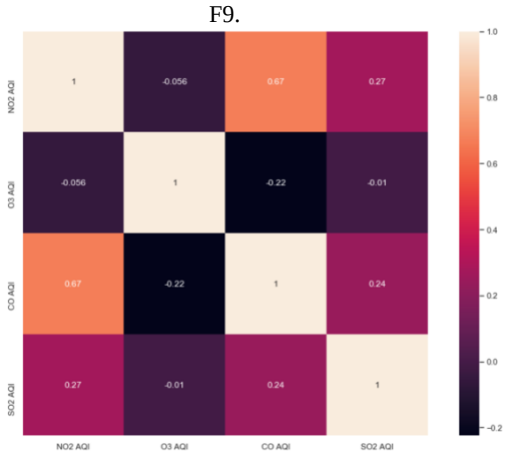
F8.



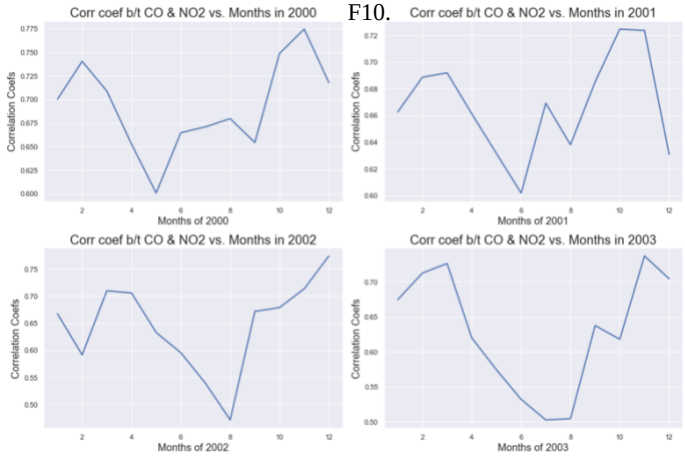
It was found that the correlation between NO_2 and CO increased from about 0.50 to 0.60 through the sixteen years while the correlation between SO_2 and NO_2 decreased from about 0.34 to 0.27 (See Figure 8 above, on the right).

After further restricting the data to those from the state of California, it was found that there was a 0.67 regional correlation between NO_2 and CO (See Figure 9 below, on the left). This correlation confirmed a previous study by Stieb et al, saying that the strongest correlation was between NO_2 and CO.[19] It can be caused by CO influencing the oxidation of NO to NO_2, or some other common influencing factors.

F9.



F10.

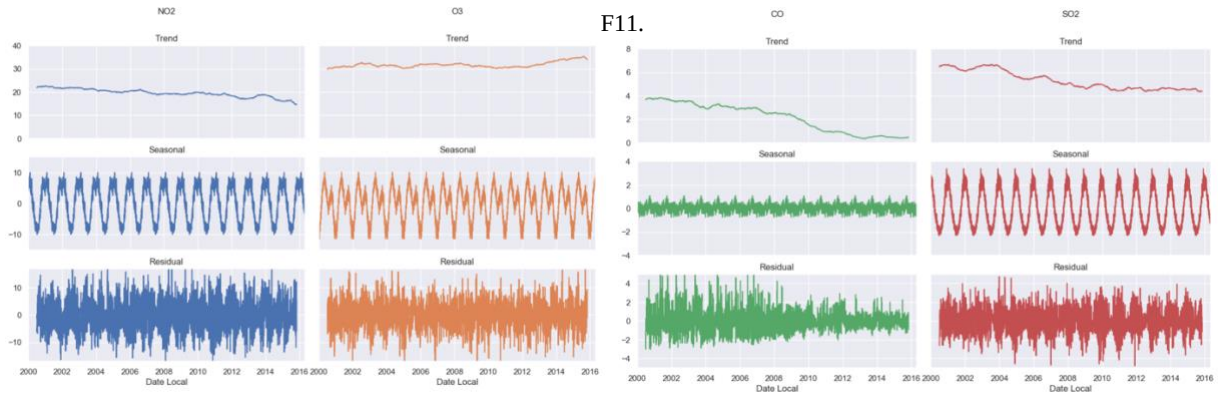


When calculating the correlation coefficient between NO_2 and CO_2 for each month using data restricted to only California, it was found that there seemed to be an interesting pattern (See Figure 10 above and on the right for a few years of example). The two air pollutant levels are more correlated from September to March and less correlated from May to August, with October being the most correlated and July being the least correlated. This pattern could possibly be explained by the fact that summer is usually the most volatile season with all types of influencers.

4.2 Results of Time Series Analysis

After decomposition, the trend of NO_2 AQI decreased from about 23 to 15, O_3 AQI increased from about 30 to 35, CO AQI decreased from about 4 to 0, and SO_2 AQI decreased from about 6 to 4 (See Figure 11 below). From the seasonality graph, it could be observed that CO had the lowest seasonal pattern amplitude while NO_2 and O_3 both had high pattern amplitudes. Confirming previous insight, O_3's seasonal pattern was opposite of the rest. The residuals were relatively high, likely indicating that there were quite a few other outside variables/predictors of each exact day (temperature, air humidity, traffic, fires, etc) influencing the air quality data other than the extracted general trend and seasonality pattern. This was part of the limitation of performing time series analysis and forecasting on air quality data because time series are usually based solely on previous values and patterns. All but the CO AQI residual seemed completely scattered. There was some fanning in the CO AQI residual plot, likely due to the fact that the CO AQI level decreased to about 0 in the last few years, causing all patterns and influences after around 2009 to be less prominent.

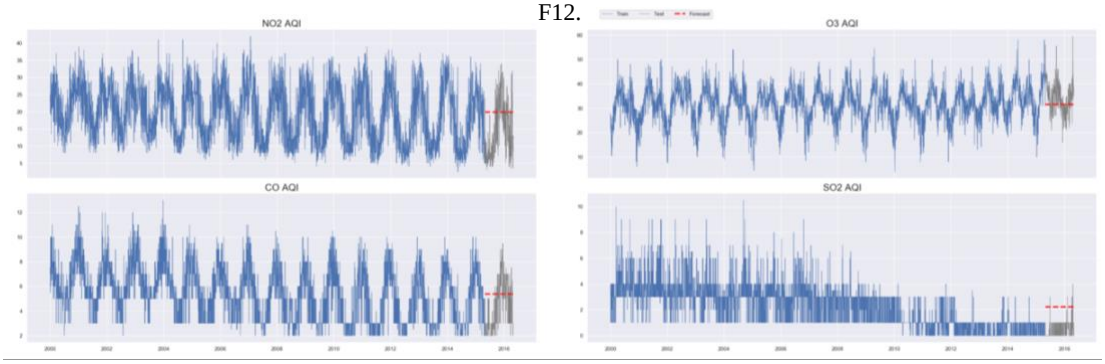
F11.



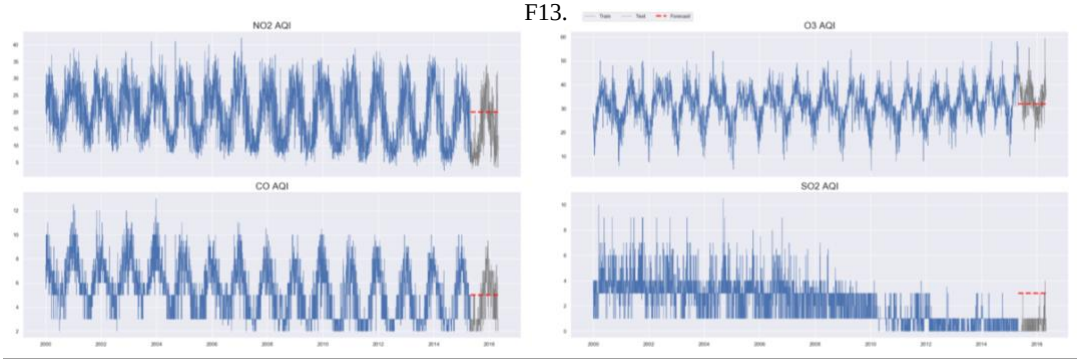
4.3 Fitting of Naïve Forecasting Models

Regarding the naïve methods of forecasting, the average and median methods both led to a linear line of forecast. They performed very similarly for all AQI time series but SO_2's as SO_2's data distribution was the least normal, with a slight skew. According to both the visual depictions (See Figures 12 and 13 below) and the accuracy measures of Root Mean Squared Error (RMSE) and Mean Absolute Error (MSE) (See Figure 17 below), these two methods were clearly unfit for forecasting.

F12.

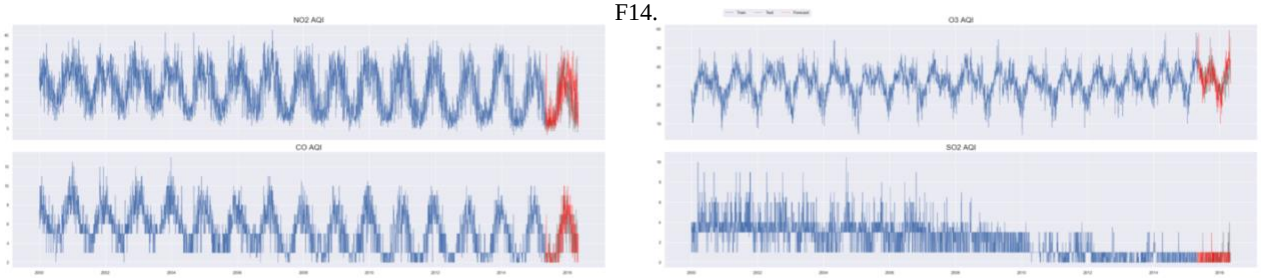


F13.



The seasonal naïve method worked the best out of the three naïve methods according to both the plots and the accuracy measures of RMSE and MAE, indicating that the AQI values of the year 2016 were extremely similar to that of the previous year (See Figure 14 below). This again confirmed the seasonality patterns for the AQIs. Though the seasonal naïve method worked the best, its accuracy measures for O_3 and CO AQI time series were only slightly better than those of the average and median methods (See Figure 17 below). The fact that the average and median methods had relatively low errors suggests that O_3 and CO AQI values have been quite consistent in its past years.

F14.



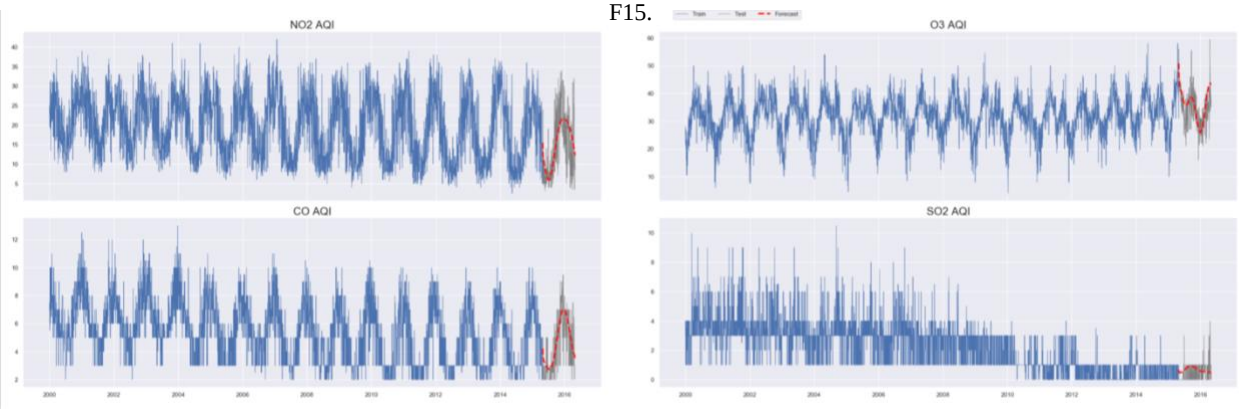
4.4 Fitting of Advanced Forecasting Models

The SARIMAX models fitted well visually overall (See Figure 15 below). A small issue was that the model predicted a smaller amplitude than the actual one for NO_2 AQI. The accuracy measures confirmed that when comparing to the naïve models, the SARIMAX models worked better in every case except one, which will be discussed below.

The SARIMAX models of NO_2 and O_3 AQIs yielded higher AIC (definition on Page 7). while those of CO and SO_2 AQIs yielded much lower AIC, meaning that NO_2 and O_3 AQIs had worse fits. The SARIMAX model of SO_2 had insignificant p-values for some of its variables—especially the Fourier Term variables—likely because the pattern of SO_2 AQI values deviated from the seasonal cycles as the values experienced a drop in 2008. All other p-values for the four SARIMAX models were significant.

According to MAE only, the Seasonal Naive model worked slightly better for SO_2 AQI time series (See Figure 17 below). However, RMSE didn't agree with MAE as it showed that the SARIMAX model fitted better instead. Since the seasonal naive method completely depended on the last year before the test year and SO_2 AQI historical values have been very different from its more recent values, it made sense that the seasonal naive method worked well for SO_2 AQI.

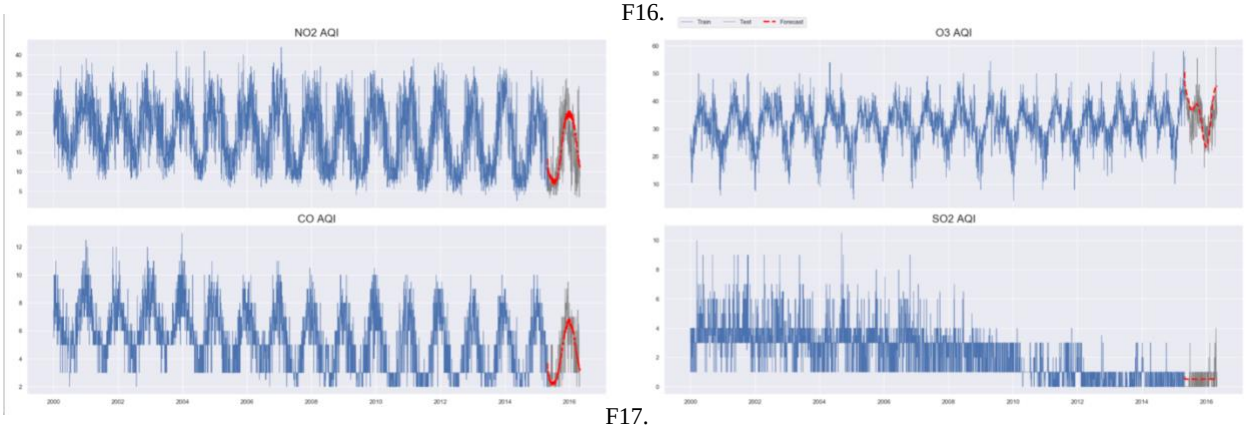
F15.



The TBATS models also provided a decent fit with its ability to capture dynamic and complex seasonality, meaning that it can account for the slight changes in seasonality patterns (See Figure 16 below). However, it performed worse than the SARIMAX model for NO_2 and O_3 AQIs time series as per the accuracy measures (See Figure 17 below). The first TBATS round of model used in this study was designed for data with both yearly and monthly seasonality, and minimal monthly pattern was observed for NO_2 and O_3 AQIs. It performed better for SO_2 and CO AQIs for likely two reasons. First, a slight monthly seasonality was observed in CO AQI earlier. This reason might be true for CO AQI, since an additional round of TBATS testing (TBATS2 in Figure 17) with only yearly seasonality yielded slightly higher errors than the first round. Second, the TBATS model is more flexible and can put more weight on values closer in time to the test set of 2016 and thus could perform a better job when there was a drastic change in values and patterns sometime within the 16 years.[20] This reason is likely true for both SO_2 AQI and CO AQI, since both experienced a decrease in 2008 and the first performed similarly for both rounds of TBATS (TBATS likely picked the same model). For NO_2 and O_3, the second round of TBATS without monthly seasonality performed better than the first round of TBATS with monthly seasonality. A limitation to note was that the TBATS model was done through a completely automated manner. As with

any automated modelling framework, there may be cases where it gives poor results with large residuals. It also ran a lot slower and might be crude for complicated processes.[20, 21]

F16.



F17.

		NO2 AQI	O3 AQI	CO AQI	SO2 AQI
RMSE	Average	9.38819	6.84743	2.1554	1.86375
	Median	9.46645	6.69784	2.01459	2.60057
	Seasonal_Naive	7.94802	6.4829	1.7216	0.818718
	SARIMAX	5.86568	5.3431	1.22328	0.697068
	TBATS	6.16912	5.93251	1.21408	0.641165
MAE	Average	8.14482	5.20569	1.85638	1.79049
	Median	8.19755	5.10354	1.72752	2.53134
	Seasonal_Naive	5.76431	5.00409	1.28474	0.544959
	SARIMAX	4.54948	4.12856	0.961896	0.560748
	TBATS	4.72799	4.73584	0.942076	0.524515
RMSE	TBATS2	6.028461	5.552590	1.224744	0.641165
MAE	TBATS2	4.566134	4.275004	0.949578	0.524515

5. Conclusion and Future Research

Predicting the air quality is a complex task due to the dynamic nature, volatility, and variability in space and time of air pollutants. Yet it is necessary to keep improving this ability to model, predict, and monitor air quality.

This study worked on a dataset with sixteen years of air quality data of four air pollutants, all extracted from the US EPA. In addition to confirming the answers to some preliminary questions, it also derived interesting insights about the correlation patterns that could be useful in future research aiming to find the most influential factors of air quality that influence the levels of multiple air pollutants at once. The insights about correlation were not previously studied and could be potentially useful for supporting city administrators in decision making.[22] While SARIMAX and TBATS models both performed well in fitting, they also helped provide a better understanding of these 16 years of air quality data.

This study certainly had lots of areas of improvements. It could have benefited from more research of literature and analysis, yet time for reading previous research and coding of this study was lacking due to unfortunate circumstances caused by the Pandemic. However, an extension of this research would be pursued by the author in the following summer.

The dataset was both the center of this study and the core limitation. For future research, data closer in time should be included and recent factors, such as 2020's wildfires, should be taken into considerations as well. In California only, there was a total of 10.8 million acres burned from 2011-20, with 3.2 million acres burned just in the year 2020 alone, covering cities with an orange haze.[23]

This study could have benefited greatly if its dataset incorporated data about other predictors such as temperature and traffic levels, possibly from the National Climate Data Center. A future study could compare the magnitude of influences of predictors and of previous values and then combine both patterns to hopefully greatly improve the accuracy of the models. Another future study could focus on connecting the changes in air quality with corresponding environmental laws in order to see "how great of a job" humans are doing so far combating air pollution. Nevertheless, this is a long battle that all people—scientists, politicians, and everyone else—need to face for the sake of humanity and the environment.

6. References

1. So, B. (2016, November 4). U.S. Pollution Data. *Kaggle*. https://www.kaggle.com/sogun3/uspollution.
2. AirData Website File Download Page. (2020, May 19). *Environmental Protection Agency*. https://aqs.epa.gov/aqsweb/airdata/download_files.html.
3. Air Topics. (2020, May 15). *Environmental Protection Agency*. https://www.epa.gov/environmental-topics/air-topics.
4. Ground-level Ozone Basics. (2020, September 10). *Environmental Protection Agency*. https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics.
5. Dedoussi, I. C., Eastham, S. D., Monier, E., & Barrett, S. R. H. (2020, February 12). Premature mortality related to United States cross-state air pollution. *Nature News.* https://www.nature.com/articles/s41586-020-1983-8.
6. Holmes-gen B., Barrett W. (2016). Clean Air Future, Health and Climate Benefits of Zero Emission Vehicles, *American Lung Association.*
7. Castelli, M., Clemente, F. M., Popovič, A., Silva, S., & Vanneschi, L. (2020, August 4). A Machine Learning Approach to Predict Air Quality in California. *Complexity*. https://www.hindawi.com/journals/complexity/2020/8049504/.

8. Lin, M., Horowitz, L. W., Payton, R., Fiore, A. M., & Tonnesen, G. (2017, March 1). US surface ozone trends and extremes from 1980 to 2014: quantifying the roles of rising Asian emissions, domestic controls, wildfires, and climate. *Atmospheric Chemistry and Physics.* https://acp.copernicus.org/articles/17/2943/2017/.
9. Rice, D. (2017, March 3). Air pollution in Asia is wafting into the USA, increasing smog in West. *USA Today*. https://www.usatoday.com/story/weather/2017/03/02/air-pollution-asia-wafting-into-usa-increasing-smog-west/98647354/.
10. State of the Air. (2020). *American Lung Association.* https://www.stateoftheair.org/.
11. Rosenfeld, J. (2019, October 17). The Top Seven U.S. States with the Worst Air Quality. *Molekule Blog.* https://molekule.science/the-top-seven-u-s-states-with-the-worst-air-quality/.
12. Goyal, P., Chan, A. T., & Jaiswal, N. (2006). Statistical models for the prediction of respirable suspended particulate matter in urban cities. *Atmospheric Environment*, 40(11), 2068–2077. https://doi.org/10.1016/j.atmosenv.2005.11.041
13. Kumar, U., & Jain, V. K. (2010). ARIMA forecasting of ambient air pollutants (O_3, NO, NO_2 and CO). *Stochastic Environmental Research and Risk Assessment*, 24(5), 751–760. https://doi.org/10.1007/s00477-009-0361-8
14. Liu, P. W. G. (2009). Simulation of the daily average PM_{10} concentrations at Ta-Liao with Box–Jenkins time series models and multivariate analysis. *Atmospheric Environment*, 43(13), 2104–2113. https://doi.org/10.1016/j.atmosenv.2009.01.055
15. Nickerson, D. M., & Madsen, B. C. (2005). Nonlinear regression and ARIMA models for precipitation chemistry in East Central Florida from 1978 to 1997. *Environmental Pollution*, 135(3), 371–379. https://doi.org/10.1016/j.envpol.2004.11.010
16. Tripathi, O. P., Jennings, S. G., O'Dowd, C. D., Coleman, L., Leinert, S., O'Leary, B. (2010). Statistical analysis of eight surface ozone measurement series for various sites in Ireland. *Journal of Geophysical Research*, 115, D19302.
17. Prabhakaran, S. (2019, February 18). ARIMA Model - Complete Guide to Time Series Forecasting in Python. *ML+*. https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/.
18. Hyndman, R. J. (2010, September 29). Forecasting with long seasonal periods. *Hyndsight.* https://robjhyndman.com/hyndsight/longseasonality/.
19. Stieb, D. M., Burnett, R. T., Smith-Doiron, M., Brion, O., Shin, H. H., & Economou, V. (2008). A New Multipollutant, No-Threshold Air Quality Health Index Based on Short-Term Associations Observed in Daily Time-Series Analyses, *Journal of the Air & Waste Management Association*, 58:3, 435-450, DOI: 10.3155/1047-3289.58.3.435.

20. De Livera, A. M., Hyndman, R. J., & Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. *J American Statistical Association*, *106*(496), 1513–1527. https://robjhyndman.com/publications/complex-seasonality/
21. Hyndman, R.J., & Athanasopoulos, G. (2018) *Forecasting: principles and practice*, 2nd edition, *OTexts:* Melbourne, Australia. OTexts.com/fpp2. Accessed on November 22, 2020.
22. Cagliero, L., T. Cerquitelli, S. Chiusano, P. Garza, G. Ricupero and X. Xiao, "Modeling Correlations among Air Pollution-Related Data through Generalized Association Rules," *2016 IEEE International Conference on Smart Computing (SMARTCOMP)*, St. Louis, MO, USA, 2016, pp. 1-6, doi: 10.1109/SMARTCOMP.2016.7501707.
23. Krishnakumar, P., & Kannan, S. (2020, September 15). 2020 California fires are the worst ever. Again. *Los Angeles Times*. https://www.latimes.com/projects/california-fires damage-climate-change-analysis/.
24. Rawat, S. (2020, February 2). India - Air Pollution Data Analysis. *Medium.* https://towardsdatascience.com/india-air-pollution-data-analysis-bd7dbfe93841.
25. Zanella, F. (2019, August 4). SQL, Tableau, and Forecasting on US Pollution Data. *Kaggle.* https://www.kaggle.com/fredzanella/sql-tableau-and-forecasting-on-us-pollution-data.
26. Park, J. (2017, January 17). Animation, Basemap, Plotly for Air Quality Index. *Kaggle.* https://www.kaggle.com/jaeyoonpark/animation-basemap-plotly-for-air-quality-index.
27. Roller, J. (2020, July 4). Data analysis & interactive visualizations (plotly). *Kaggle.* https://www.kaggle.com/janroller/data-analysis-interactive-visualizations-plotly.