Air Quality Forecasting

Joi Chu-Ketterer

DSC 680

Bellevue University

May 1, 2020

lchuketterer@my365.bellevue.edu

Instructor: Catharine Williams

Introduction

There is no dispute that air pollution negatively impacts communities. Aesthetically, having a lot of smog and pollution in the area makes a city less appealing for tourists and visitors. This can have undesirable consequencies to the city's economy, especially if they heavily rely on tourism. Furthermore, pollution and poor air quality greatly affects residents' health both in the short and long term. Some of the largest contributors to air pollution include industrial plant emissions, fossil fuel combustion, chemical farming, and natural causes (forest fires, volanoes) [Maddan 2018]. All of these are byproducts of human activity. While one could argue humans have no control over when volcanoes erupt, studies have shown about 85% of all wild fires are a result of humans poorly managing their camp fires [NPS 2018]. For this reason, being able to understand how societal trends impact air quality trends would provide companies the tools to address and mitigate their contribution, while also providing the government a platform to advocate for more sustainable policies.

One of the largest arguments for poor air quality in various cities is that humans are creating too much waste. While studies have been conducted, there has been a slow movement towards corporate, governmental, and societal change. However, as devasting as the COVID-19 pandemic is for multiple reasons, it has provided analysts with the opportunity to guage the impact humans have on earth's condition. As more people abide by the CDC's COVID-19 response suggestiosn [see Appendix A], fewer people are out on the roads and the streets. As a result, large cities have seen a drop in air pollution and improved air quality. This study aims to run several time series forcasts of air quality data for Los Angeles, CA to get a better sense of how LA's air quality is predicted to change amidst a pandamic.

Business Understanding

Air quality is a term used to quantify the number of pollutants present in the air. The main compounds that scientists measure include ["AQI Basics" *AirNow*]:

- Ground-level ozone (O_3)
- Carbon monoxide

- Sulfur dioxide (SO_2)
- Nitrogen dioxide (NO_2)

For this reason, data on O_3 and NO_2 were analyzed. These two pollutants were specifically chosen due to their origins and how they get into the air in a harmful manner [see Appendix B]. Using Time Series analysis is a beneficial approach for it's high readability, identification of seasonal patterns, and ability to create trend estimates. While there are numerous different types of time series models out there, the most common are Naïve, Exponential Smooting, and ARIMA [Burba 2019]. All three methods have their own strengths and weaknesses [see Appendix C]. This study focuses on comparing these three forcasting models using air quality data.

Analysis

Data Collection

There were two sources of data used for this study. Both NO_2 and O_3 data was retrieved from the South Coast Air Quality Management District (South Coast AQMD). Since the platform only allows data retrevial for one trace gas evey six months, multiple requests were made. Southern California COVID-19 case data was extracted from the Los Angeles Times' GitHub repository. All resources were updated frequently, providing up-to-date data for both. At the time of concluding this study, O_3 and NO2 analyzed data ranged from April 19, 2019 through April 30, 2020, and COVID-19 data ranged from January 26, 2019 through April 30, 2020.

Method

This project was divided into two data preparation stages and an analysis stage. The first data preparation stage was conducted in Python to merge all separate datasets together into one. The second data preparation stage and analysis stages were both conducted in R. R was chosen for the time series analysis given the available library packages.

Data Preparation

The second preparation stage consistened of replacing *NaN* values with *0* and removing unnecessary columns. The final dataset included:

- Datetime
- NO2 trace gas levels (pphm)
- O3 trace gas levels (pphm)
- COVID-19 confirmed case count for Los Angeles County

Results

Data Exploration

Prior to creating models, a base understanding of the current trends was required. For this reason, two separate graphs were created for trace gas levels, and COVID-19 cases.

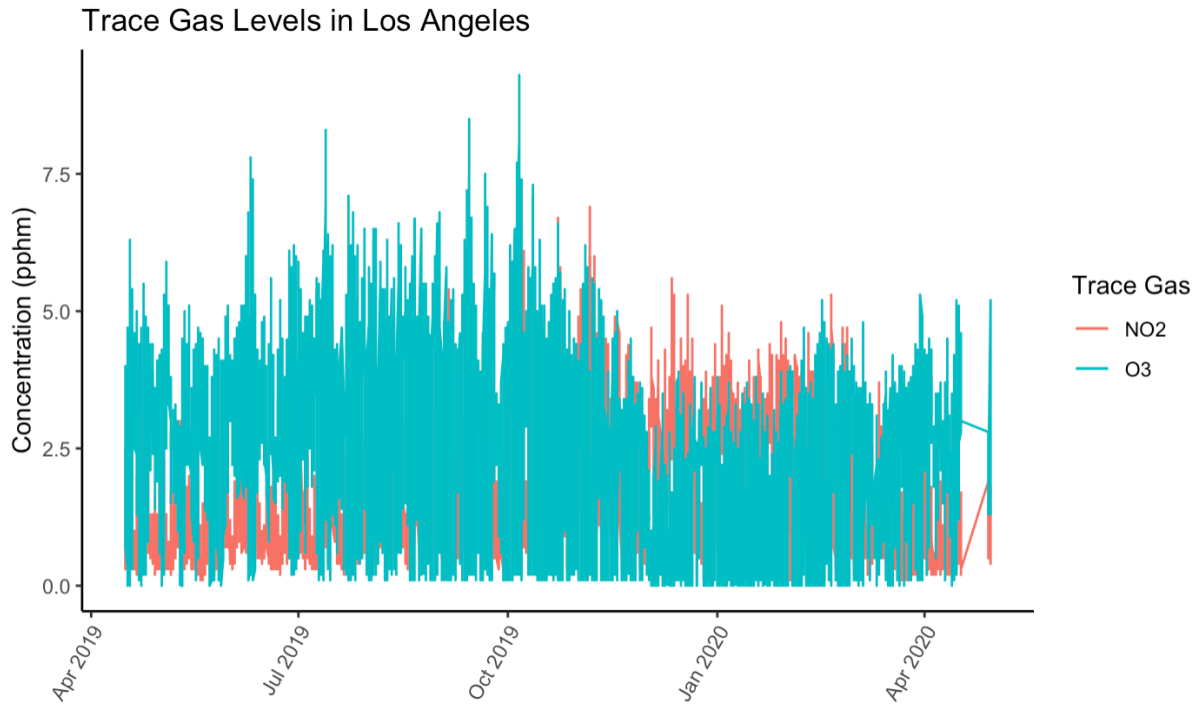

Figure 1. Trace gas levels (pphm) in Los Angeles, CA between October, 2019 and April, 2020.

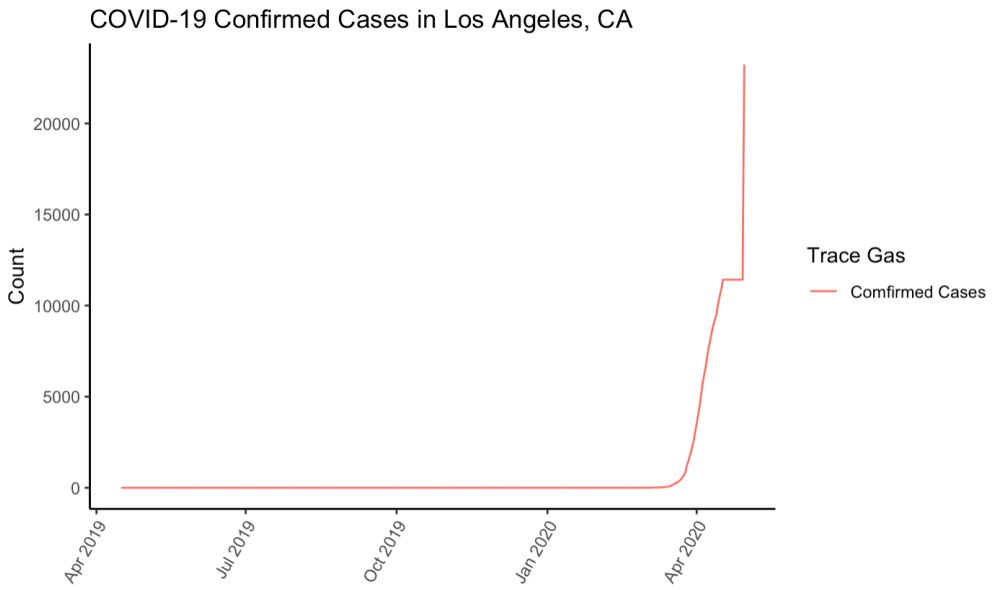

Figure 2. COVID-19 confirmed cases in Los Angeles, CA between October, 2019 and April 2020.

The trends for NO_2 levels in *Figure 1* and for confirmed COVID-19 cases in *Figure 2* show dynamic changes overtime. However, O_3 levels seem to be rather consistent through the months. For this reason, only NO_2 was selected for further analysis.

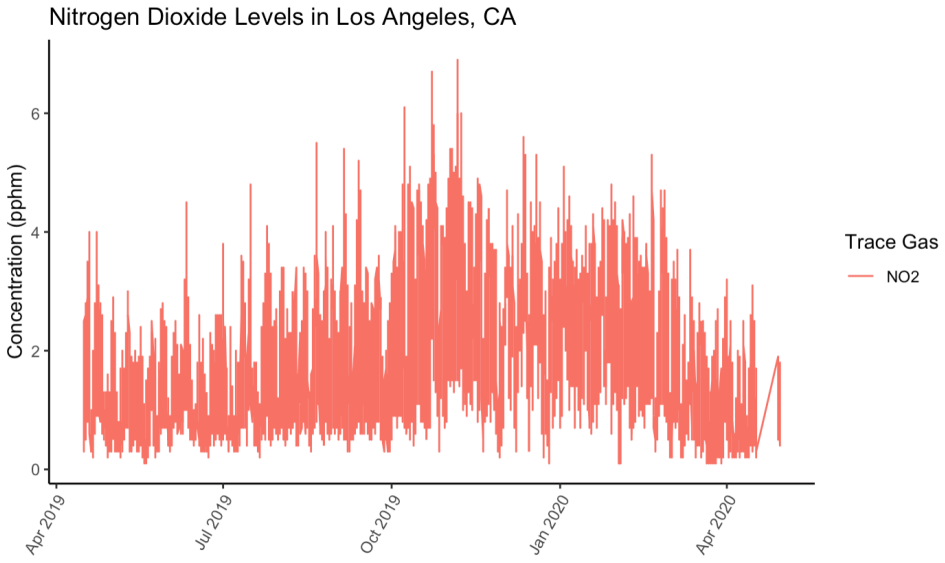

Figure 3. NO_2 levels (pphm) in Los Angeles, CA between October, 2019 and April 2020.

Figure 3 shows a clear decline in NO_2 levels, which a significant drop around February, 2020. The drop in February occurs a little earlier than the March increase of COVID-19 cases in Los Angeles, CA shown in *Figure 2*. Since there is not enough data to identify seasonal patterns, the data was transformed into stationary data to help extract possible seasonality.

Once the initial exploratory plots were created, in order to run time series analysis, the data had to be converted into a time series object. Additionally, forecasting models work btter with trend-stationary data, as it allows for true seasonal trends to be uncovered. The *diff()* method was used to transform the data. The method essentially calculates the difference between the observations rather than the observation values.

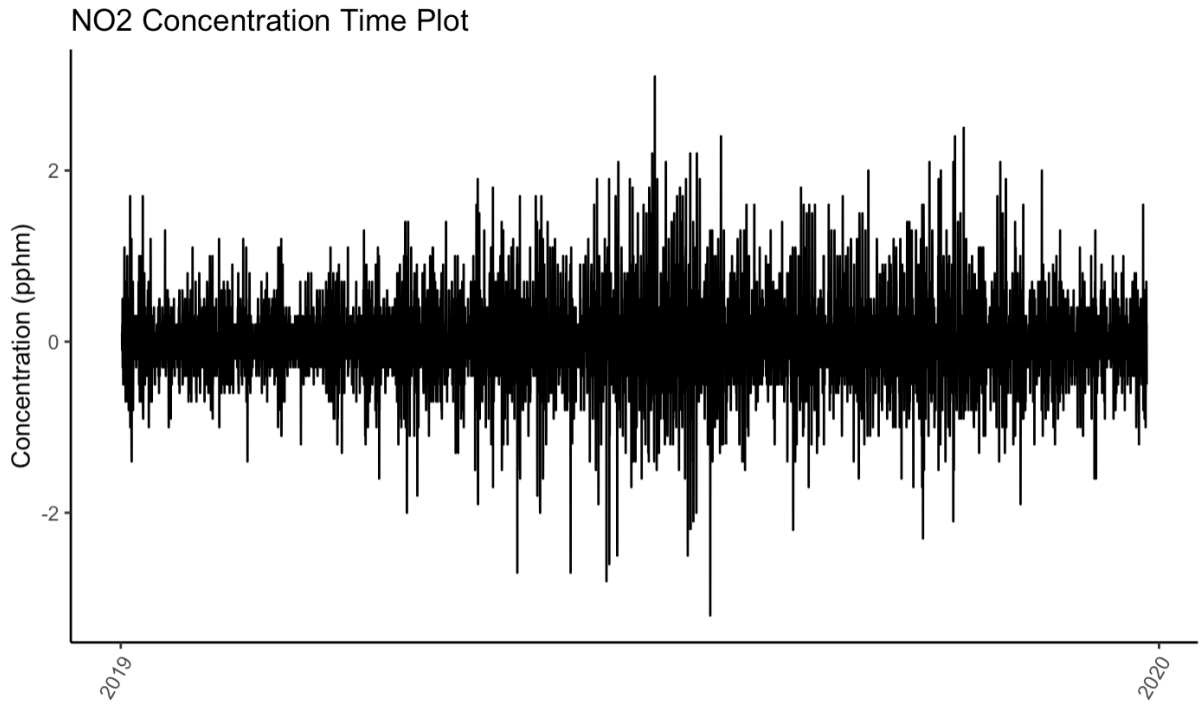

Figure 4. Trend-stationary data transformation for NO_2 levels in Los Angeles, CA between October, 2019 and April, 2020.

A stark difference between *Figure 3* and *Figure 4* is shown, demonstrating how the trend-stationary transformation helps uncover any seasonality hidden in the data. With a relatively linear pattern in *Figure 4,* it is concluded the dataset for NO_2 in Los Angeles does not have seasonality. With the data ready for

modeling, three different time series models were generated. Model performance was evaluated and compared using standard deviation error statistics and correlograms.

Model Evaluation and Comparison

For each model created, the performance was evaluated using standard deviation of residuals and correlograms. *Table 1* shows the standard deviation residual error statistics for all three models, while *Figure 5* shows the three correlograms.

Table 1. Standard deviation error values for all three time series models created

	Naive	ETS	ARIMA
Residual Standard Error	0.6248	0.429	0.610

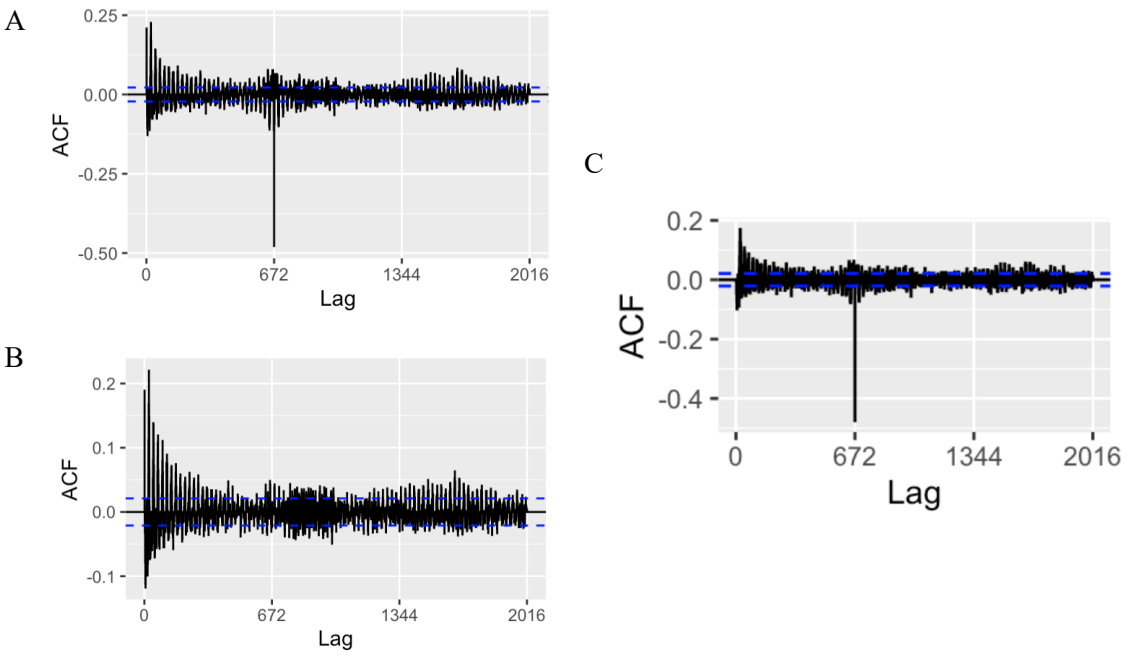

Figure 5. Correlograms for (a) Naïve, (b) Exponential Smoothing, and (c) ARIMA time series models. Blue dotted lines indicate 95% confidence intervals.

In *Table 1*, it is evident the ETS model results in the smallest error, suggesting it is forecasting the most accurately. Additionally, *Figure 5* shows the correlogram for the ETS model has the least amount of autocorrelation. Ideally, all the data points would lie within the 95% confidence interval, where any points that lie outside of these bounds suggest additional factors that were not taken into consideration are affecting the trend. For this reason, the ETS is considered the best performing model and was chosen to forecast for the next month of NO_2 levels.

Forecast

Once the final model was selected, the *forecast()* method was used to forecast the NO_2 concentrations in Los Angeles. *Figure 6* demonstrates the predicted concentrations in blue. It is evident, that Los Angeles is working towards a positive trend with NO_2 levels continuing to expect to decrease. Again, the consistent decline in NO_2 concentrations may be of consequence to Los Angeles' stay-at-home and safer-at-home orders.

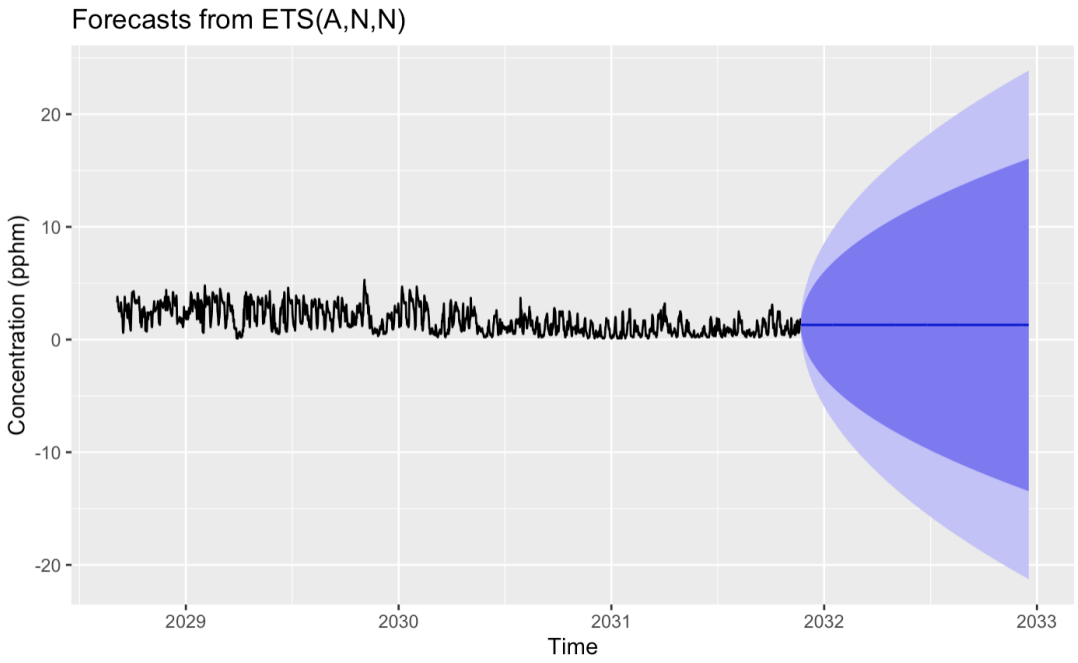

Figure 6. ETS forecast model for NO_2 concentration in Los Angeles for May 2020

Given the original data was sampled on an hourly basis, the forecasting model created also forecasts on an hourly basis. *Figure 6* shows the next month of hourly forecasts for NO_2 concentrations in Los Angeles, paired with the last three months. Given the forecasting uncertainty, the details in the patterns are lost. However, the trend does seem to be consistent with previous data, suggesting that the air quality will continue to decline.

Benefits

While the current Los Angeles stay-at-home and safer-at-home orders are not sustainable once the COVID-19 pandemic is less of a threat, understanding that having fewer residents out on a daily basis leads to higher air quality is something policy makers can not ignore. Since NO_2 is mostly a byproduct of car emissions, rather than ordering residents to stay home all the time, the city may be able to regulate the number of cars on the road. Additionally, with ongoing efforts to improve the public transportation, additional funding could be allocated to such projects to encourage more of the city to opt for that over using a personal car.

Future Work

Univariate time series forecasting focuses on generating future predictions based on past observations. While past trends can certainly help predict future ones, as shown through autocorrelation plots, there are additional factors that are not being taken into consideration. While one could speculate the impact COVID-19 regulations has on Los Angeles air quality, it was not specifically included in the time series model. For this reason, should the project continue, additional factors can be taken into account and a multivariate time series analysis is conducted. Especially in the circumstance of using such studies to advocate for sustainable policy, it would be ideal to have forecasting models that showed the importance of regulating societal behaviors.

The forecasting model was difficult to interpret due to the scale of the data, being in pphm. For this reason, the difference in data points and future forecasting may be easier to spot should the units be in

ppb. This paired with including data from further back may help demonstrate the decline in NO_2 levels better.

Conclusion

COVID-19 has provided society with an interesting ability to observe just how far society impacts the environment. While people follow safer-at-home orders, streets of large cities are found to be empty and the skies become clearer. This projects supports the observation that having fewer people on the streets has a positive impact on Los Angeles' air quality. While a little inconclusive, a forecasting model does suggest continuing to minimize the number of cars on the road would continue to improve air quality. For this reason, it would be in the city's best interest to find ways to continue regulating the number of cars on the roads every day as the threats of COVID-19 lessen and safer-at-home orders are lifted.

References

"AQI Basics", *Air Now.* Retreived from https://www.airnow.gov/aqi/aqi-basics/

Mar 15, 2020. "Interim Guidance for Coronavirus Disease 2019 (COVID-19)", *Centers for Disease Control and Prevention*. Retreived from https://www.cdc.gov/coronavirus/2019-ncov/community/large-events/mass-gatherings-ready-for-covid-19.html

November 27, 2018. "Wildfire Causes and Evaluation", *National Park Service.* Retreived from https://www.nps.gov/articles/wildfire-causes-and-evaluation.htm

October 31, 2018. "Ground-level Ozone Basics", *Environmental Protection Agency.* Retreived from https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics#formation

September 8, 2016. "Basic Information about NO_2", *Environmental Protection Agency.* Retreived from https://www.epa.gov/no2-pollution/basic-information-about-no2#What%20is%20NO2

Burba, Davide. Oct 3, 2019. "An overview of time series forecasting models", *Towards Data Science*. Retreived from https://towardsdatascience.com/an-overview-of-time-series-forecasting-models-a2fa7a358fcb

Chauhan, A. J., Krishna, M. T., Frew, A. J., & Holgate, S. T. (1998). Exposure to nitrogen dioxide (NO2) and respiratory disease risk. *Reviews on environmental health*, *13*(1-2), 73-90.

Madaan, Sonia. 2018. "Various Causes of Air Pollution", *Earth Eclipse.* Retreived from https://www.eartheclipse.com/pollution/primary-causes-of-air-pollution.html

Appendix A

Health and Safety Regulations Pertaining to COVID-19

Given the unpredictable nature of COVID-19, health and government organizations constantly update their own suggestions. At first organizations and political figures were using headcounts as a metric safety in conjunction with physical distance. Soon after, specific activities were advised against and establishments were characterized as non-essential and ordered to close. On March 11, 2020 the CDC accounced that

- Any events of 250 or more people should be cancelled
- Older persons and those with severe pre-existing health conditions are at highest risk

Around the same time, the idea of social distancing was introduced to describe keeping a distance of six feet between people. At this time, it was anticipated that closures would last a two week period until the end of March. Later on March 15, the social gathering head count was reduced from 250 to 50, and extended social distancing to eight weeks [CDC 2020]. Since then, additional regulations have been issued by state government and local officials. Also on March 15, California Govenor Newsom called for the closure of all bars and nightclubs, and reduced restuarants to take-out and delivery only [Bizjak, 2020]. Numerous states have adopted this idea since then as well as reducing people's time outside to groceries and medical appointments.

Appendix B

Nitrogen Dioxide and Ground-Level Ozone Sources

Nitrogen dioxide, NO_2, is one of the reactive gases of nitrogen oxides. It is most commonly produced by burning fuel. Specifically, it forms from the gas emissions of cars, trucks, power plants, and any fuel-powered machine [2018 EPA]. In addition to being an air pollutant, it also is able to interact with water to form acid rain, which is harmful to many ecosystems. In humans, NO_2 is linked to aggravating respiratory symptions in both those with and without underlining conditions [Chauhan 1998].

Stratospheric ozone forms in Earth's upper atmosphere, and functions as a protective layer for the planet. However, tropospheric ozone, or ground-level, is harmful. Ground-level ozone is a mixture of nitrogen oxides and volatile organic compounds, VOCs. While O_3 is the result of nitrogen oxides, VOCs, and sunlight reacting with one another, it is still prevelant during winter months. Similar to NO_2, O_3 can cause respiratory issues in humans.

Appendix C

Time Series Models

Seasonal Naïve models assume the current value holds the same value as the previous observation plus an artibuartiy amount [Burba 2019]:

$$\hat{Y}(t+h|\,t) = Y(t+h-T)$$

Exponential smoothing models generate predictions using a weighted average of past observations. Each observation that is further back in time is assigned decreasing weights [Burba 2019].

$$\hat{Y}(t+h|\,t) = \alpha y(t) + \alpha(1-\alpha)y(t-1) + \alpha(1-\alpha)^2 y(t-2) + \cdots$$

ARIMA, AutoRegressive Integrated Moving Average, models generate predictions based on a linear combination of past observations [Burba 2019]. Similar to ETS models, ARIMA models use weighted averages of past observations to predict future observations. The differences lies in that ARIMA models also take into consideration weighted averages of the errors as well. For this reason, ARIMA models are considered to be more sophisticated and accurate compared to other forecasting models.