DSU Covid-19 Environmental Impact Project

The Data Science Union at UCLA

Project Leads: *Helen Coffman, Audrey Cabrera*

Project Members: *Ryan Ohlinger, Arnav Gangal, Debbie Ryu, Hayden Souza, Ishaan Shah, Josh Chan, Renzo Tanaka-Wong, Vince Front, Grace Panos*

Introduction:

In March 2020, multiple counties in California issued a lockdown in response to Covid-19. People halted their everyday activities, industries and travel paused their productions, and businesses shut down due to loss of income. As a result, the air quality improved, consumption of fossil fuel and nonrenewable energy sources was reduced, and recovery of the ecosystems. These positive effects on the environment demonstrated a solution for society to reduce the rate of climate change. Our project aimed to explore the impact of the pandemic on the environment, more specifically, how the pandemic impacted air pollution in California. We chose to investigate the change of three air pollutants - Sulfur Dioxide (SO_2), Carbon Monoxide (CO), and Nitrogen Dioxide (NO_2). We gathered past records of air pollutant levels from 2015 to 2020 to create prediction models that estimate future air pollutant levels under pandemic conditions and applied exploratory analysis to demonstrate a change in air pollutant levels. In addition, we built an interactive dashboard model of all the contributors to climate change and the impact the lockdown measures had across California.

Data Collecting:

Our data sources were the U.S. Energy Information Administration (EIA) site, which we used to obtain energy and fuel consumption data, and the Bureau of Transportation of Statistics (BTS) to acquire travel statistics. To gather information on our air pollutant levels of Sulfur Dioxide, Nitrogen Dioxide, and Carbon Monoxide, we used daily data from the United States Environmental Protection Agency (EPA). We scraped our data from each of these sites and put it together on a master spreadsheet.

a. Air pollution:

To analyze how levels of Carbon Monoxide, Sulfur Dioxide, and Nitrogen Dioxide changed over the course of the pandemic, we sourced data from EIA. Our data consisted of concentration values and Air Quality Index (AQI) values for all three pollutants above, taken from various tracking sites across the state of California daily. Data from 2015 to 2020 was our main focus, and this data was modified to monthly concentration and AQI values by year for each pollutant for us to best understand how pollutant levels changed in this time period. Columns regarding tracking site location, state name, latitude, longitude, observation count and codes were not relevant to our area of exploration and were dropped.

b. Energy & Fuel Consumption:

To examine the environmental impacts during the pandemic lockdown we investigated the electricity consumption generated by all fuels, such as petroleum, coal, natural gas, and more. We found California's electricity net generation consumption from the EIA site. The dataset is organized by sectors of commercial, industrial, electric utility, independent producer, and one column that contains the total value across all sectors. We decided to combine commercial and industrial columns into one to simplify the dataset.

c. Travel & Transportation:

Data relating to travel and transportation includes domestic flight departures from California's top ten popular airports, California's total gasoline and jet fuel sales in thousand gallons per day. The total gasoline and jet fuel sales were acquired from the EIA database. For the gasoline dataset, we discovered missing values for some months of 2015 and all of 2016. To resolve this problem, we decided to impute missing values with previous values for that specific month. For example, to impute the missing value in January 2016, we totaled up all of the January data recorded in the other years and averaged those values to replace January 2016. As for airline travel data, we researched the top ten most popular airports in California and used the BTS as a directory to search for flight statistics at each airport. To simplify our flight dataset, we decided to focus on departures because the emissions originated in California.

Data Analysis:

Carbon Monoxide:

First, we explored how Carbon Monoxide (CO) levels changed over time through 2015 to 2020. Atmospheric CO levels were compared with electric, natural gas, solar, geothermal, and other energy production in order to determine if there was any correlation between the two.

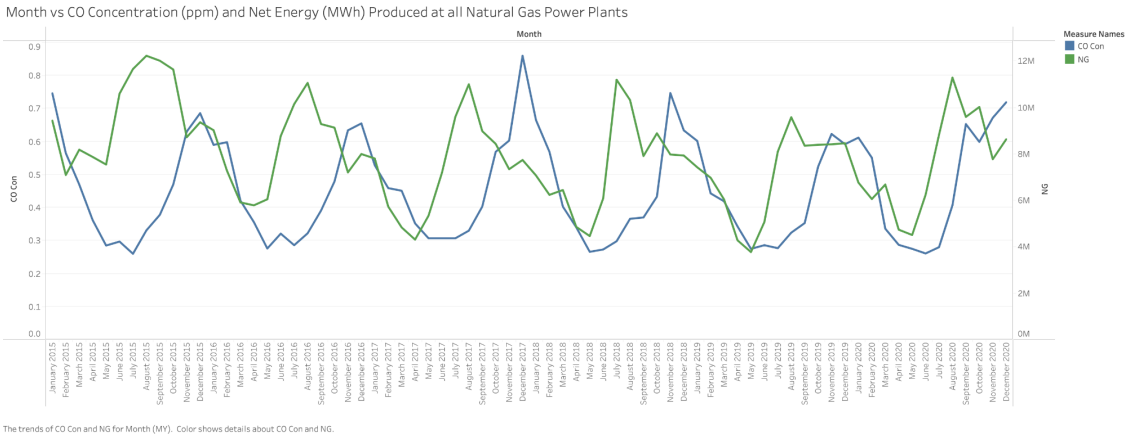


Figure 1

Figure 1 depicts a time series for CO concentration and net energy produced at natural gas plants (NG_PL), who appear to have similar trend lines over the period of 2015 to 2020. This implies a correlation between the two, and that natural gas plays a role in

atmospheric CO levels, however further data suggest otherwise. For example, Figure 2 below is a heatmap that tells us that atmospheric CO concentration appears to have a slight positive correlation with industrial coal production (Tot_Ind_Coal) and miscellaneous power plant energy (OT_PI). However, the relationship between CO and natural gas plant energy production appears to be very weak, only amounting to a value of 0.096, which goes against our findings from Figure 1.

Figure 2

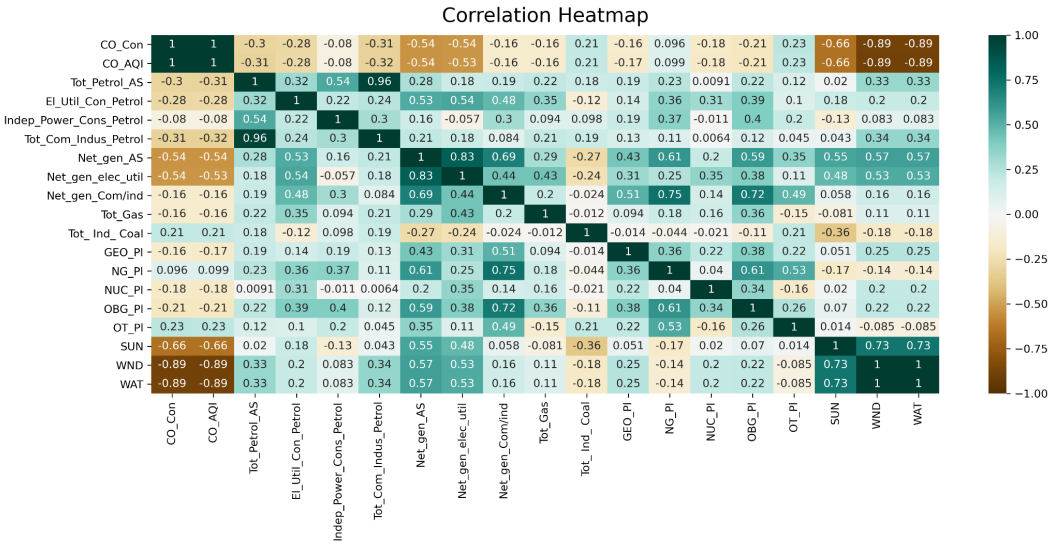
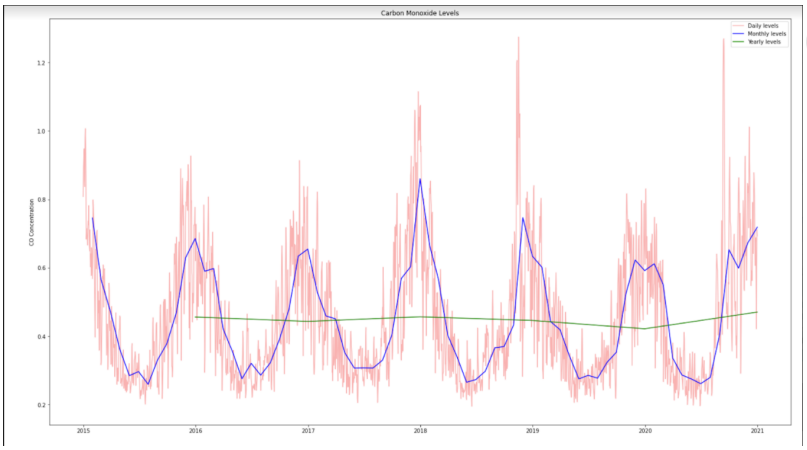


Figure 3, on the other hand, shows overall carbon concentration from 2015 to 2021. We can see that the year 2020 has a lower peak for atmospheric CO concentration than the other years, and as this is the year associated with the lockdown, we can infer that perhaps there is some truth to the lockdown reducing pollution emission. Thus, our figures unfortunately seemed to have some mixed results in regards to carbon concentration during the time period of focus.



(Figure 3)

Sulfur Dioxide:

The second pollutant that we researched and did our data analysis on is Sulfur Dioxide (SO_2). Just as with the other pollutants we researched, we compared SO_2

concentrations in the atmosphere to energy production to analyze any possible trends between the pollutant's atmospheric concentration and fossil fuels/renewable energy.

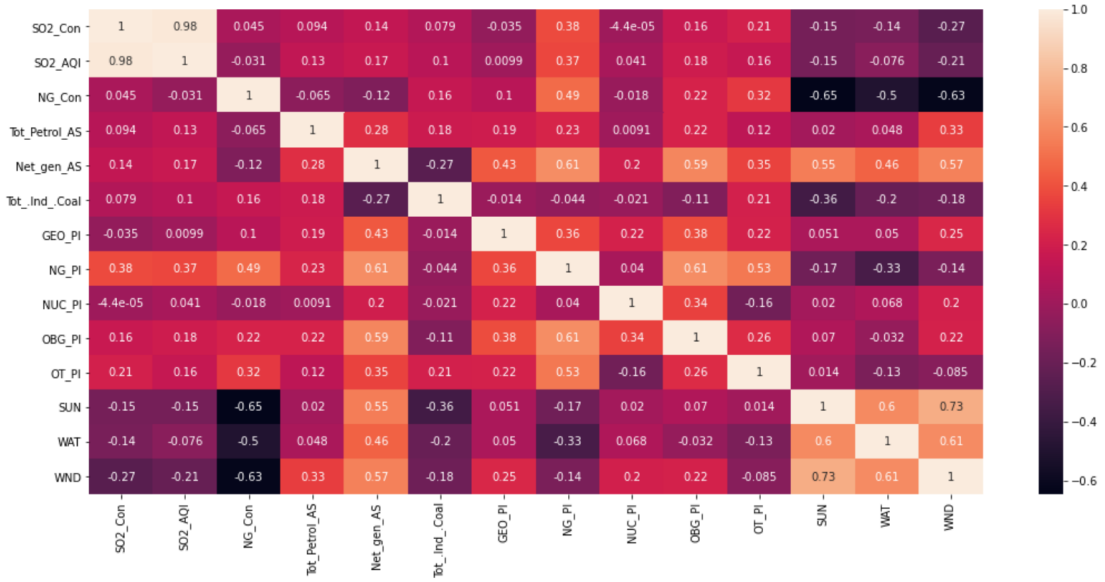


Figure 4

Figure 4 is a correlation heat map that compares the average atmospheric SO_2 concentration, in parts per million, to various predictor variables. The strongest relationships we found were between SO_2 and the powerplant data, the predictors with the highest correlations being natural gas (r = 0.38) and wind energy (r = - 0.27). Notably, all of the nonrenewable energy sources (biomass gas, natural gas, other nonrenewables) were positively correlated with SO_2 concentration, while renewable sources (solar, water, wind) were all negatively correlated with SO_2 concentration. From here, we decided to explore more about the relationship between power plant data and sulfur dioxide with a subsetted correlation plot.

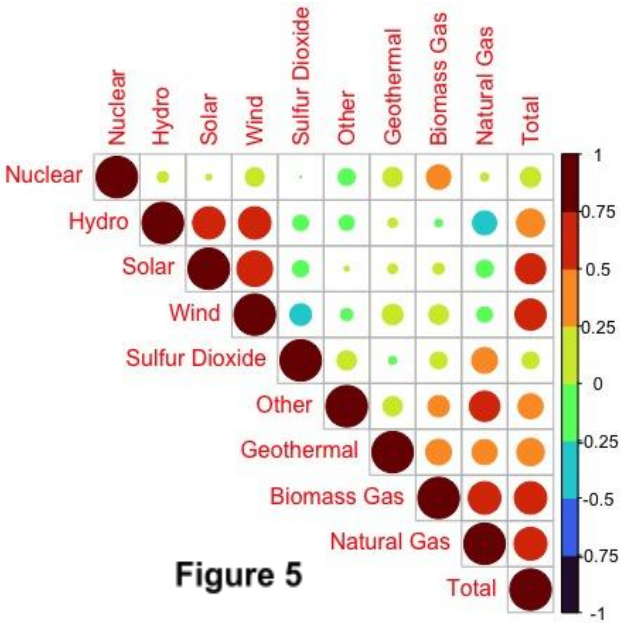


Figure 5

Figure 5 graph shows the relationship between the predictors and just the powerplant data, larger dots represent a higher correlation, and smaller dots represent a lower correlation. The scale for Figure 5 is a more accurate representation of the strength of the relationship, where the scale in the Figure 4 heatmap only goes down to about -0.2 instead of the minimum possible r-value of -1. This simplified visual shows that among all the power plant predictors, natural gas is the only one with a moderate relationship to the SO_2 variable. Natural gas has a positive relationship with SO_2 which suggests that an increase in monthly natural gas production is moderately related to an increase in monthly SO_2 concentration.

To explore this further, we introduce Figure 6 which visualizes the time-series relationship between SO_2 and natural gas. The line plot below shows standardized changes in natural gas production seem to be positively related to standardized changes in SO_2 concentration.

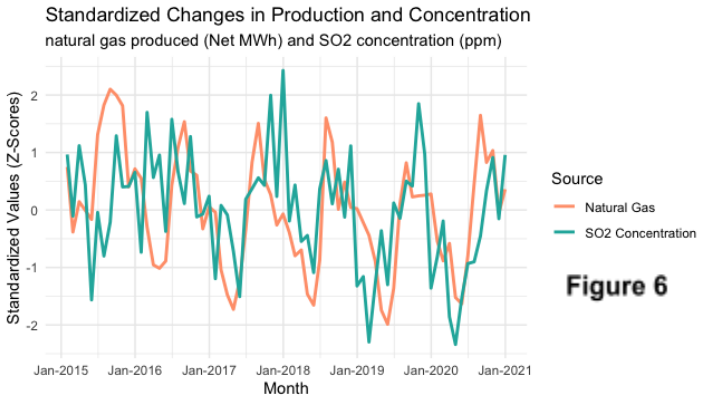
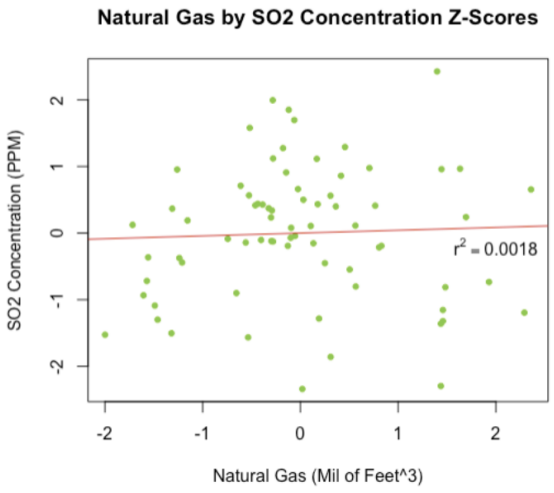


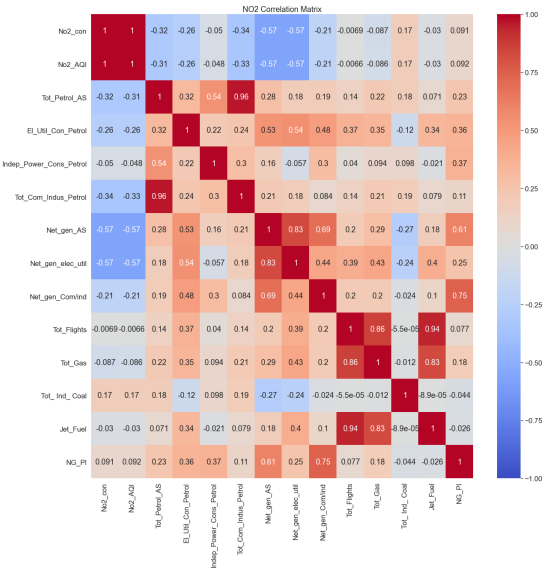
Figure 6

Figure 7



We also wanted to explore the NG_Con data from the original heatmap. NG_Con measures the amount of natural gas delivered to consumers. Our Figure 7 that shows standardized changes in SO_2 in relation to standardized changes to natural gas production has a correlation coefficient of 0, which suggests that there is no relationship between natural gas delivered to consumers and SO_2 concentration. This was intriguing considering that natural gas power plants had a moderate correlation, but this can probably be attributed to the volume of natural gas used in the electricity production process. Power plants produce more energy than home usage of natural gas, meaning that the volume used by consumer households is not enough to cause an increase in atmospheric SO_2 concentration.

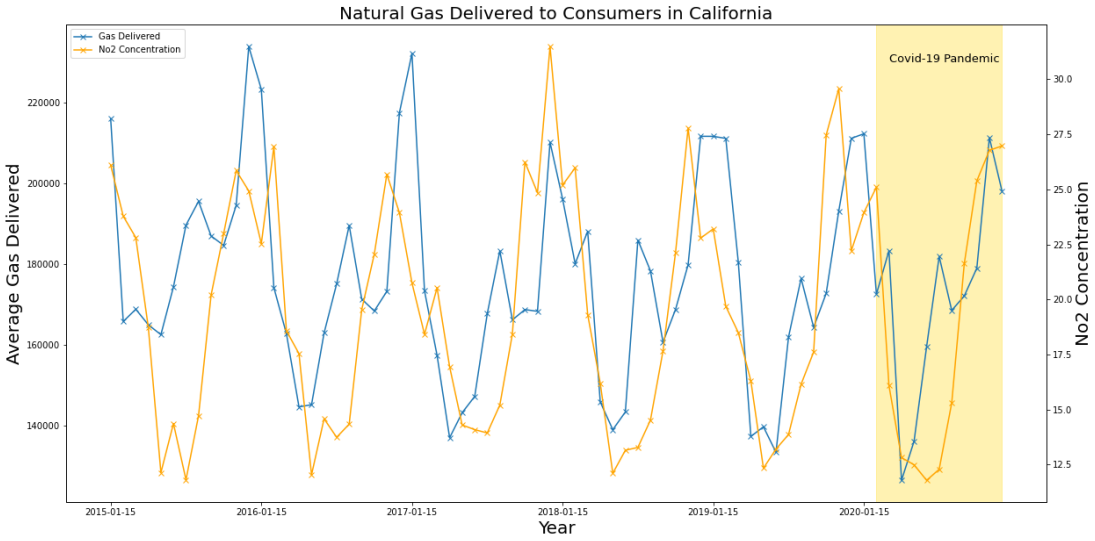
Our exploratory analysis of sulfur dioxide led us to a few interesting conclusions about the data. First of all, none of the predictors we were working with had a strong relationship with sulfur dioxide concentration. The highest correlation was between natural gas production and SO_2 concentration at r = .38. Secondly, all of our nonrenewable sources had a positive relationship with SO_2 concentration, while our nonrenewables had a negative relationship with SO_2 concentration. This suggests that switching to renewable energy generation is a potential solution to decrease SO_2 concentration in the atmosphere. These data-driven insights align well with what previous research on SO_2 has suggested and have helped us gain a greater understanding of how data is able to capture ecological phenomena.



<u>Nitrogen Dioxide:</u>

From the correlation heatmap on the right, we can see that there is little to no correlation between the concentration of nitrogen dioxide and many of the key indicators. This is

surprising, as we expected there to be stronger correlation between some variables, particularly between nitrogen dioxide concentration levels and the total number of flights (or other metrics that represent transportation). Additionally, we observed that there was a reasonably strong negative correlation between electricity generation by utility companies and production of nitrogen dioxide. This was as expected, as most electricity generation processes do not produce nitrogen dioxide.



This model above compares the concentration of nitrogen dioxide and natural gas delivered from 2015 to 2020, with the period of the pandemic highlighted in yellow. Both nitrogen dioxide and natural gas levels experience a drastic decrease at the start of 2020 and slowly begin to increase throughout the rest of the year. This aligns with California's stay-at-home order, which was issued in March 2020 and lifted in June 2020. While the decrease in gas levels cannot be directly attributed to the stay-at-home order, we can see that nitrogen dioxide and natural gas levels declined during the start of the pandemic.

Interactive Dashboard: (https://datastudio.google.com/reporting/a0a98a30-8302-44c7-9260-3c088fcf4897)
We used Google Data Studio software to create an interactive website for users to interact with our visuals. Our visuals compare each air pollutant to their contributors through 2015 - 2020. We hope users will gain a better understanding of the impact of the pandemic on fuel and energy consumption sources.

Model Analysis:

Time Series Model (ARIMA)

A time series model is a series of observations taken at specified times basically at equal intervals. It is used to predict future values based on past observed values. In our case it was the concentrations of three pollutants - Carbon Monoxide, Sulfur Dioxide, and Nitrogen Dioxide. To forecast values, we will use an ARIMA model (AutoRegressive Integrated Moving Average). We will forecast the concentration of each pollutant by

looking at the concentration of the previous few days. Our goal is to predict future pollutant levels in 2021.

To use such a model, we need to ensure our time series is stationary. In this case, stationary implies the series is mean-reverting. This means that a stationary time series is one whose statistical properties don't depend on time. This is important as an ARIMA model uses its own lags as predictors. Hence, if the underlying distribution of the time series is changing, it is tough to make predictions.

An ARIMA model has 3 parameters : p, d, and q.

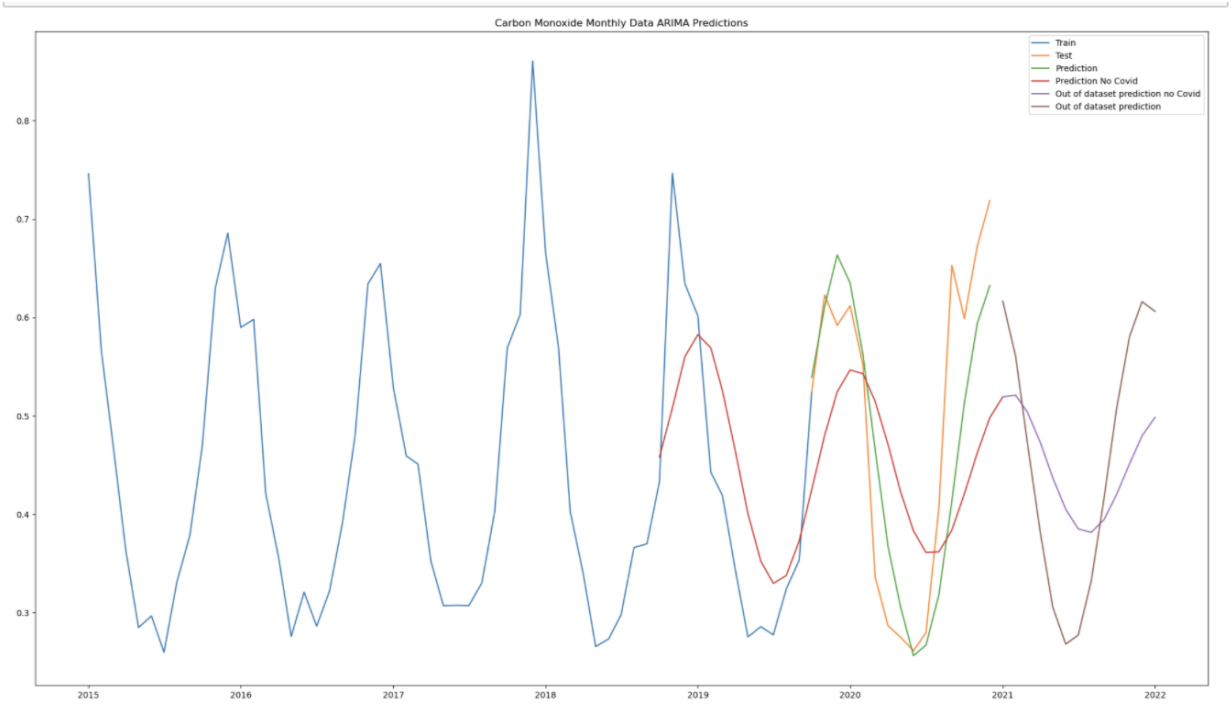
- p is the number of lags to use a predictor.
- d is the number of orders of differencing. Differencing once is subtracting the current value from the previous value. It is done to convert a non stationary time series into a stationary time series.
- q denotes the lag of the error component, where the error component is a part of the time series not explained by trend or seasonality.

The stationary analysis was done for each of the three components and appropriate values of p,d, and q were estimated.

Carbon Monoxide:

For carbon monoxide, daily values of the last 5 years were used (2015 - 2020). Looking at the autocorrelation and partial autocorrelation graphs, we chose p = 3 and d = 0 and q = 0. However, results were inadequate and r-squared was negative. We then tried to use auto ARIMA to find the appropriate values and it chose p = 10 and d = 0 and q = 0. No progress was made. This is due to the model overfitting the training data on account of daily fluctuations. Hence, it predicts poorly on the test data. As the ARIMA model learns on its own (i.e. the past values), one bad prediction can cause several bad predictions. We then used monthly averages with the ARIMA model. As a result, the data was smoother and received better results with a r-squared of 0.6. The model could sufficiently predict values until 2022 as well. However, when the pandemic lockdown year (2020) was removed, the model became worse due to the training set being too small. The final results for 2021 CO levels using ARIMA can be seen below. The model produced unsatisfactory results because of the high amounts of fluctuations which are causing overfitting on the data. As mentioned before, ARIMA uses older values to predict new values, one bad prediction can lead to multiple bad predictions. In the future, to improve this we would try to use Facebook Prophet for time series forecasting

as it is a more powerful and updated library.

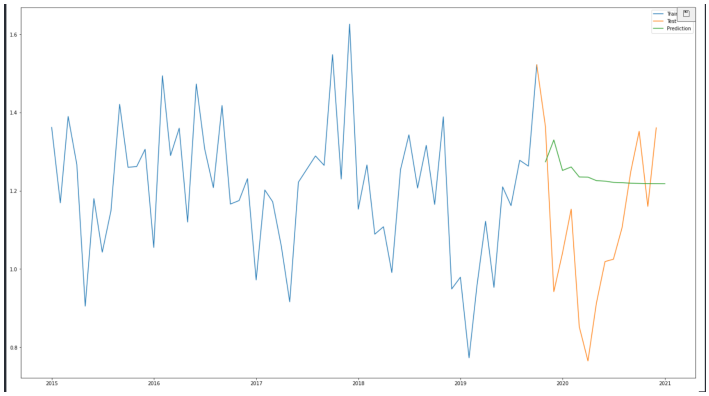


Sulfur Dioxide:

To develop an ARIMA-based predictive model for SO2 concentrations, we used data from 2015 to 2020 (inclusive). Before building the model itself, we tested the data for stationarity. A plot of concentration values over time showed no trend/seasonality of emission levels; that is, the fluctuations seemed very arbitrary and likely that the data was not stationary. We applied an Augmented Dickey-Fuller Test (ADF) to confirm this hypothesis.

The ADF test is a popular and powerful tool to determine if data is mean-reverting/stationary. It makes use of a hypothesis test, where the null hypothesis is that the given data is “non-stationary”. The alternative hypothesis is that the data is “stationary”. Python provides a simple package to implement this test without having to get mired in the complex technicals of the test. We used this package to deploy the test, doing so against a significance level of 0.05 (the recommended significance level value for ADF tests). To reject the null hypothesis, the test is required to return a p-value > 0.05, thereby providing evidence that the data is “stationary”. We obtained a p-value of 0.1, implying that we could not reject the null hypothesis. We then tried building a model using this fact. Non-stationarity needs to be accounted for when building an ARIMA model; however, one caveat of the ARIMA model is that it automatically becomes unreliable when data is non-stationary. *(Note: We did not check for patterns in the autocorrelation graph since this ADF test achieves the same end of establishing any trends/stationarity.)*

We used the auto_arima function to gauge the optimal p, d, q values for the model, and got p = 2, d = 0, q = 0 to be the best fit values. We used data from 2015-2019 to train the model and use the model to predict values for 2020. The model compares the predicted values against the already known 2020 data. There was a huge discrepancy in the predicted and testing values, hinting that this model was a terrible fit. We received an r^2 value of -66. This implies that our model fit is worse compared to a horizontal line of regression.

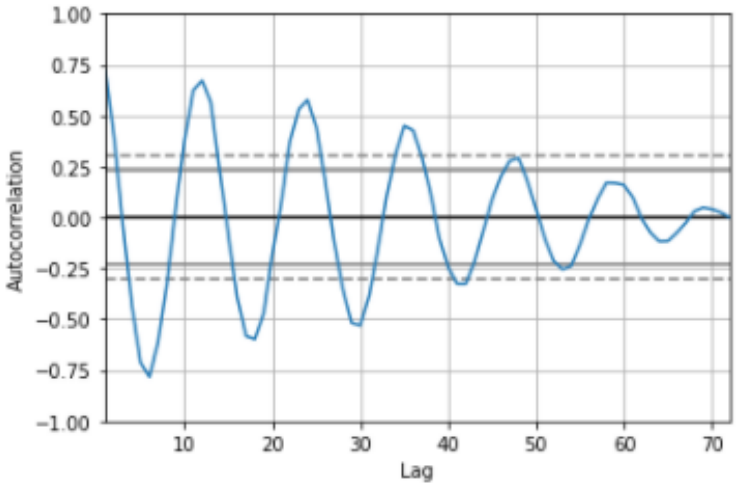


Graph comparing predicted SO2 values for 2020 using the ARIMA model vs actual data

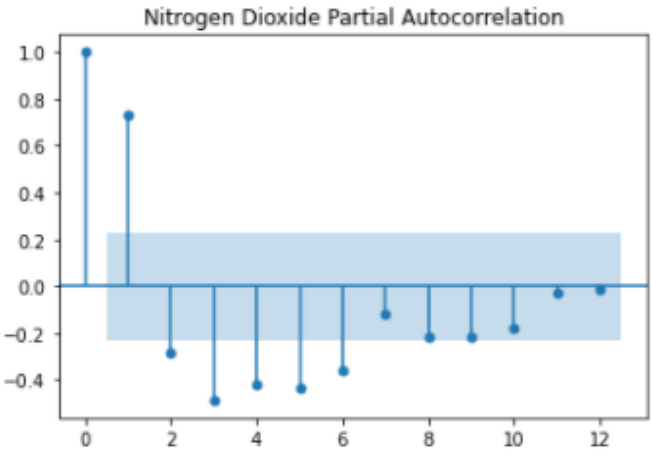
Two factors that led to such an inaccurate prediction model were 1.) the non-stationarity of data, and 2.) the limited quantity of data points. Using only 5 years of data (= 60 data points) on readings with seemingly arbitrary fluctuations severely limits the accuracy of such a model. Perhaps using SO2 concentration values from more years + a multivariate analysis (through SVR, for example) may create better predictions.

Nitrogen Dioxide:

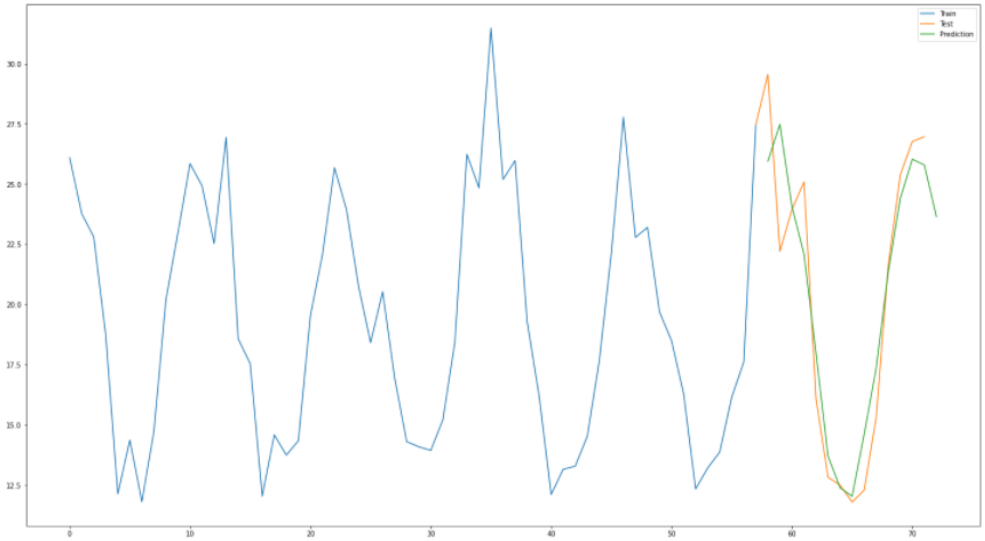
We decided to use an ARIMA model to predict NO2 concentration because the Earth's weather patterns are relatively stable and we don't expect drastic change in the near future. Our dataset used NO2 concentration data from the years 2015-2020 (including the effect of COVID). First, we tested the NO2 dataset for stationarity and graphed the autocorrelation.



There is a repeating pattern in the autocorrelation graph above about every 12 months, implying a yearly trend and constant pattern. This further provides evidence for stationarity. Next we test the dataset for partial autocorrelation in order to find a p value for the ARIMA model. A partial autocorrelation plot depicts the correlation between two observations that the shorter lags between those observations do not explain.



Here we see that lags of order 0-1 are highly statistically significant, which will be our p value. We can use this to create our ARIMA model later on. An ARIMA model uses 3 main parameters: p, d, and q values. For the sake of this model, our initial estimations were P = 0-1, d =0 (due to the stationarity of the dataset), and q = 0 (due to the absence of lag of the error component). However, after using the auto_arima function from pmdarima with the initial estimates, we found p=6, d=0,q=0 to be the best values. After plotting the ARIMA model using those parameters, our predictive model produced the following results with an R squared value of .6 and an mse of 3.27.



Support Vector Regression (SVR) Model

Using pollutants' previous concentrations

Introduction:

In order to predict the concentrations of our pollutants, we utilized the data of each pollutant's previous average monthly concentrations. We did so by using the Radial Basis Function (RBF) kernel with the support vector regression (SVR) model in the Scikit-learn (sklearn) machine learning library in Python. SVR is a type of supervised machine learning model that analyzes data for regression analysis and attempts to find a hyperplane in an n-dimensional space that best fits the data. With SVR, we can choose a margin of error that we tolerate, therefore being able to minimize error. RBF is the default kernel in sklearn and uses two main parameters, gamma and C. The gamma parameter defines how far into the data one training point can affect and impact, while the C parameter directs SVR on the size of the margin regarding misclassification. A smaller value of C allows for a larger margin and leads to more misclassification by the hyperplane. In our project, we found and used the best values for the parameters gamma and C in order to allow SVR to find a hyperplane that fits the data best. This helped us minimize error and perform more successful regression analysis.

Why we chose SVR Model:

We decided to try out this model because a SVR can be very advantageous due to its simplicity. By using only two variables, date and pollutant average monthly concentration, we would be able to run the model with minimal pre-preparation and obtain accurate results.

Cleaning and preparing the data:

The data preparation process for the SVR model began with converting the data columns to interpretable integer data types for machine learning. This encoding step is a necessary pre-processing step in machine learning because models can not interpret any other data type. Since the pollutant data consists of pollutant concentration and aqi, which are floats, and date which is an object, we proceeded by converting the date column to its intergerized equivalent.

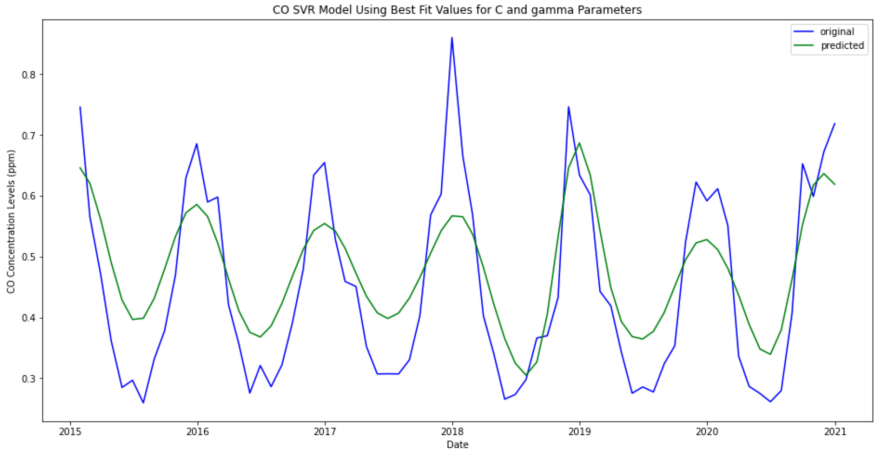
First, we converted the date column to a datetime object, a Python module used for manipulating dates and time. Next, we applied the toordinal() function which returns the proleptic Gregorian ordinal of a date. The resulting column was a numeric column with encoded dates.

After this step our data was converted to the proper shape: The X-values for data with n columns must be of the shape (n,1) and the Y-values must be a list of n objects. We then used the train_test_split function from the sklearn modules to split our data into training and testing sets at the proportion 70:30. We split the data accordingly in order to have a set of data which would be used to train our models and a set of data to determine the accuracy of our model and determine necessary adjustments to our model's parameters.

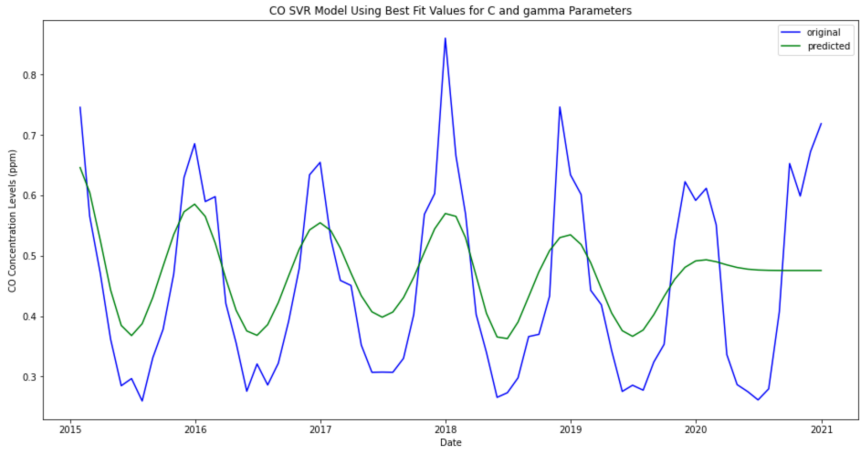
Results and findings for each pollutant:

Carbon Monoxide (CO):

When working with carbon monoxide (CO) from 2015 to 2020, we used a random train test split on the data and utilized the grid search method to find the best values for C and gamma, which came out to be gamma = 0.0001 and C = 1. Using these parameters, the root mean square error (RMSE) was 0.109, the coefficient of determination R^2 was 0.745, and the percent error was 4.7%.

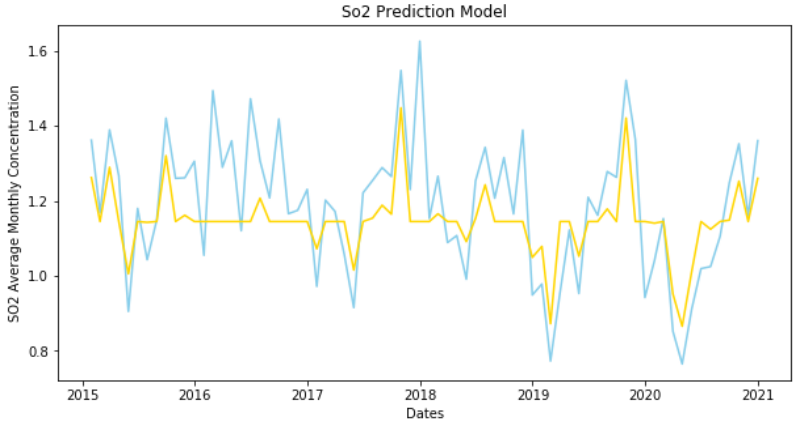


Next, we used a similar method as the previous approach, but attempted to use the 2015 to 2019 CO concentrations data to predict future levels and to predict the concentrations of 2020. We used a random train test split again and trained on the 2015 to 2019 data and tested on the 2020 data. From the grid search method, we used gamma = 0.0001 and C = 1. This gave us a percent error of 25.4%, RMSE of 0.119, and R^2 of 0.691. Seeing as the data from 2020 involved concentration levels impacted by Covid-19, the error was clearly greater and made this model more inaccurate with its predictions for 2020. Unfortunately, when using this approach for the model, we came across an error with the 2020 predictions. The predicted values for that year evened out to be the same value, as seen in the graph below, due to an insufficient amount of training data to predict from.

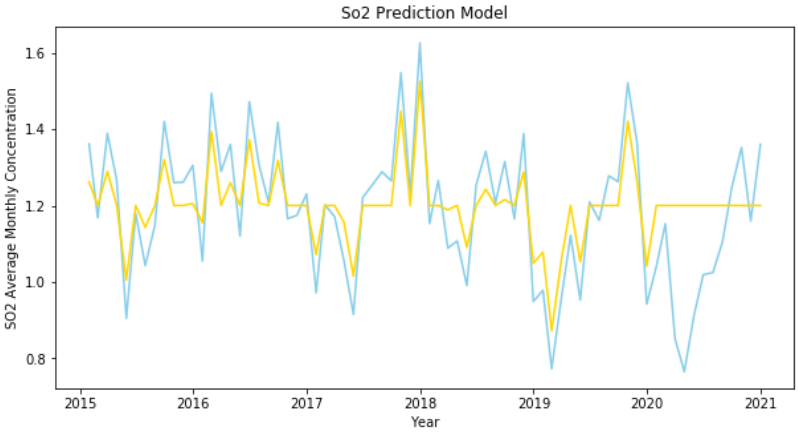


Sulfur Dioxide (SO2):

For sulfur dioxide we were able to obtain an R-Squared of .76 by training the model on randomly split training test data. However, we noticed that this did not account for the time series component of the prediction model that we were trying to solve. By randomly splitting the data, the model would only be predicting values that were already within the range of dates we were training it on.



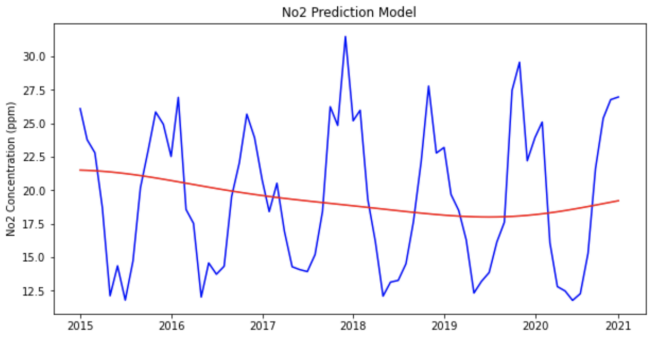
We tried to amend this problem by splitting the data and training it on 2015-2019 sulfur dioxide concentrations and testing it on our 2020 data, but after January of 2020 the model outputted the same average monthly sulfur dioxide prediction for every month until the end of the year.



This is because the model was only predicting on the previous point in the training data, and by only using two predictors it would be impossible to create an accurate model for 2020 data with this method, there needed to be more predictors to build a working SVR forecasting model.

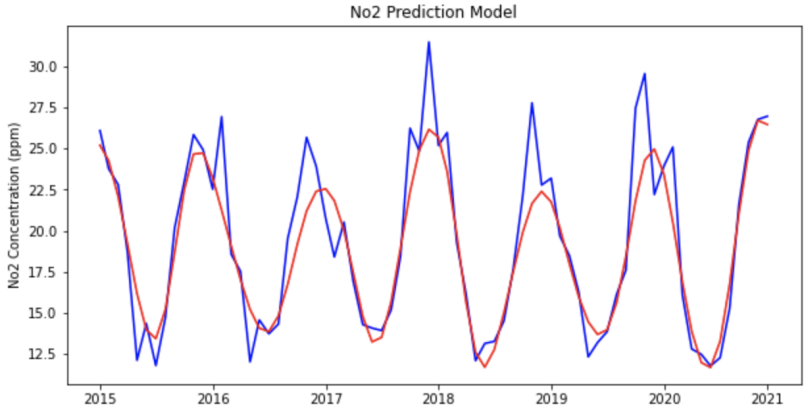
Nitrogen Dioxide:

The Nitrogen Dioxide SVR model was created and trained in the same fashion as the above two models. After pre-processing the date column, a default model - with parameters kernel = ‘rbf’, gamma = ‘scale’, C = 1.0, epsilon = 0.1 - was trained on the training data. The resulting model, represented by the red line, was a very flat model that failed to capture the fluctuations in the data.



The grid search method was then used to iterate through various values for the listed parameters to determine a set of optimal parameters as follows: kernel = ‘rbf’, C = 10, epsilon = 0.5, and gamma = 0.0001. The result of this model was a smoothed model

that captured all of the movements in our data. The r-squared score was 0.911, and the RMSE was 3.068, indicating a highly successful and predictive model.



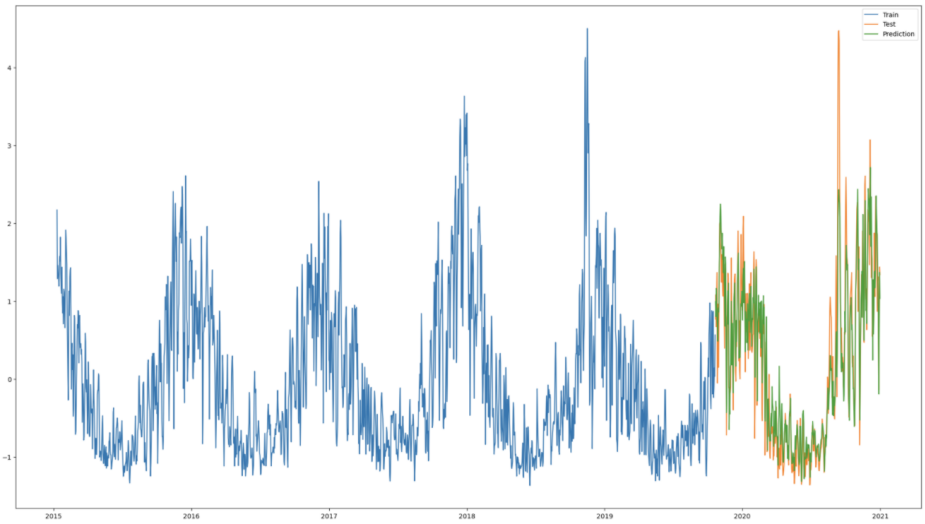
The train-test-split on the No2 data was performed at random, meaning that this model was successful in predicting No2 concentration levels at random intervals in our data, however, it was not necessarily strong at predicting concentration levels for consecutive future dates. To do this we discovered that we needed a different approach to training our model which involves offsetting our data with a lag in our input values. This method would allow us to predict future concentration levels based on past levels, but we were not able to complete this due to time constraints.

Using weather data

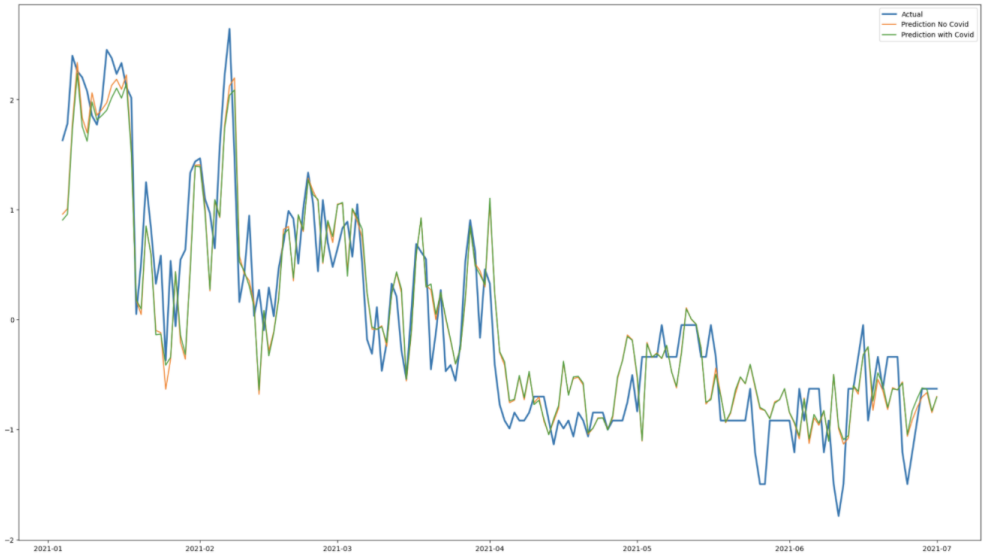
After unsatisfactory results with the time series analysis and the original SVR models, we decided to use an SVR with weather data to help predict CO concentrations. SVR is good at predicting non-linear data and hence was the best choice for this given problem. Weather data was chosen for this analysis because several processes in the atmosphere contribute to the creation and spread of these pollutants. We focused on humidity, precipitation, pressure, temperature, and wind speed as our variables.

After collecting data, we created new variables to the dataset which could help better predict the CO concentrations. We added an ‘is_weekend’ variable which denotes if a given day is a Saturday or Sunday. We also added a variable which signified which season it is given in a particular month. Finally, we added 3 lag variables which signified the concentration of CO for the past 3 days. Finally, a standard scaler was used to scale all of the variables.

We got great results, with a R-squared value of 0.93 on the training set and 0.86 on the test set. Given that the data was at a daily level, these results are amazing in our opinion.



We then trained a model without taking 2020 into account. Below are the results for 2021 using the actual data, COVID data (including 2020) and Non COVID data (not including 2020). Both models did a fairly good job of predicting the 2021 values and were almost identical to each other. Both models had a R-squared value of 0.84 signifying that the models are good at predicting future CO concentrations in California.



Conclusion:

Our exploratory data analysis results provided insight that the pandemic did not cause a significant change in pollution levels. We believe that the pandemic lockdown period between March 2020 to December 2020 was too short to reduce the pollution levels. However, our data analysis showed dramatic changes in travel and transportation (i.e., flights, jet, and vehicle fuel consumption) after March 2020. Even though the fuel consumption was reduced, it was not enough to lower the air pollution levels. Nevertheless, we investigated if the current fuel and energy consumptions would be

enough to lower air pollutant levels in the future. When implementing our machine learning models (ARIMA and SVR), we discovered there was little to no correlation between the contributors and air pollutants. As a consequence, we could not accurately predict future pollutant levels due to fluctuating values and lack of data and predictors. Despite this problem, we used the weather data SVR model to predict future air pollutant levels and received more accurate results. In the end, we were able to explore the air pollutants' concentration levels using machine learning models. The remainder of our dataset, consisting of air pollutant contributors, proved to be poor and insufficient for predicting air pollutant levels. In the future, when tackling a similar project, we believe we would need more data that is stronger and able to support our efforts in predicting air pollutant levels with SVR and ARIMA machine learning models.