Time-Series Analysis and Forecasting for US Air Pollution Data

Authors: Venkat Srinivasa Raghavan, Deepak Udayakumar, Vaishnavi Madhekar, Katyaini Raj, Rakshak Kunchum

Summary

- **Background**
  - Air pollution is a major environmental issue that affects public health and the economy, and accurate predictions of air pollution levels can help mitigate its negative impact.
  - Project utilized US pollution dataset (2016-2021) from the US Environmental Protection Agency Website.
  - Dataset includes daily values of NO2, O3, SO2, and CO.
- **Scope**
  - Our primary focus state is California because it is the home to the top five most polluted cities in the USA. Moreover, air pollution in California has been linked to the occurrence of wildfires.
- **Goal**
  - Analyze and forecast major air pollutant quantity for next 30 days.
  - **Related work:** previous research on air pollution impacts and mitigation strategies.

Summary

- **Methods**
  - pre-processing and tidying, analysis of historical data, trends, predict future trends using preprocessing, visualization and modelling.
- **Results**
  - Data follows a seasonal pattern.
  - Pollutant levels are decreasing lately.
  - Forecast shows that the pollutant levels will continue to decrease.
  - Time series models perform better than naive methods for forecasting.

Methods - Preprocessing

Data collection and aggregation:

- The raw data files were collected from the US EPA for the years 2016-2021 website and aggregated into the dataset.

Data tidying:

- **Date type** was formatted from ‘string’ type to ‘datetime’ object.

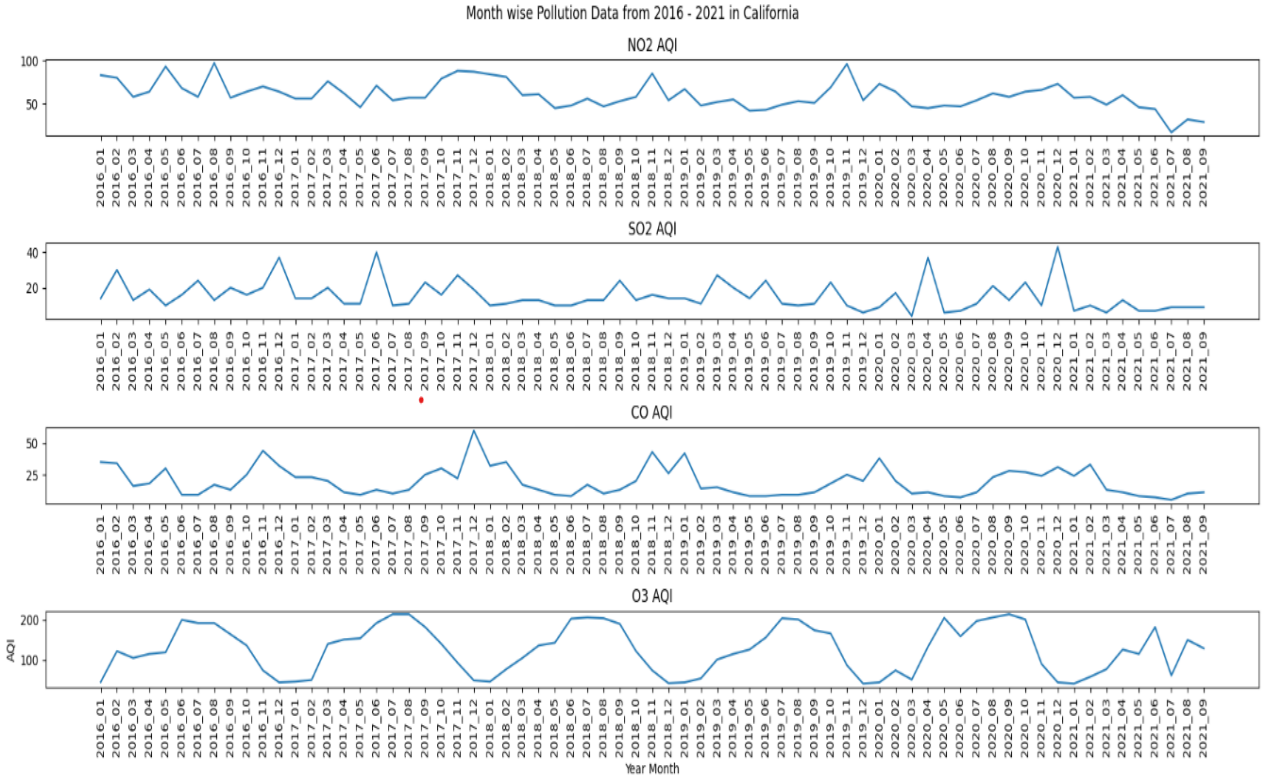
- **Data transformation for visualization:**
  Top polluters: Calculated mean of each Pollutants AQI and grouped the dataset by city to obtain the top 10 polluters.
  Confidence bands: Calculated mean and standard deviation of the pollutants distribution to obtain the upper and lower bound.

- **Data transformation for Modelling:**
  The data was not transformed much apart from selecting relevant columns and filtering the dataset scope only to california.
  The data was grouped by date, city, county which was later utilised for modelling.

Methods - Visualization

The AQI values for each pollutant seems to follow a seasonal pattern

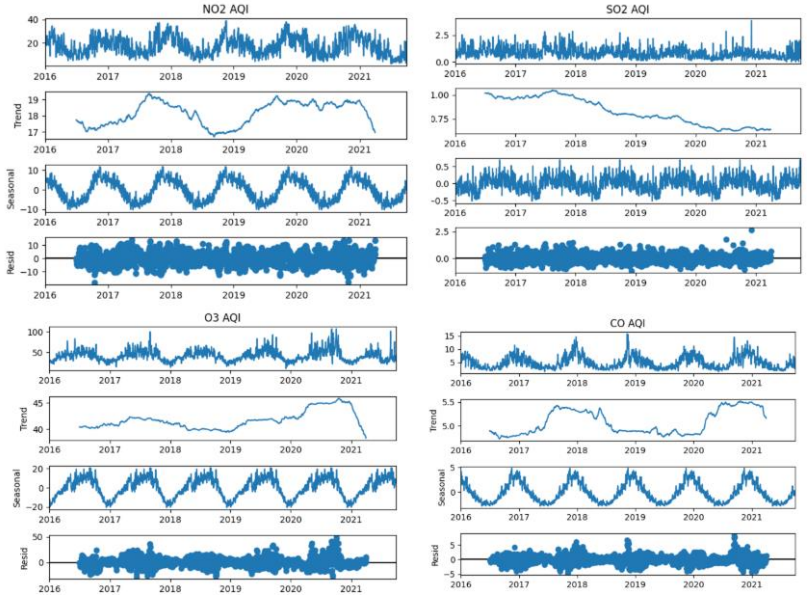
The latest pollutants indices values seem to have a decreasing trend



Methods - Visualization

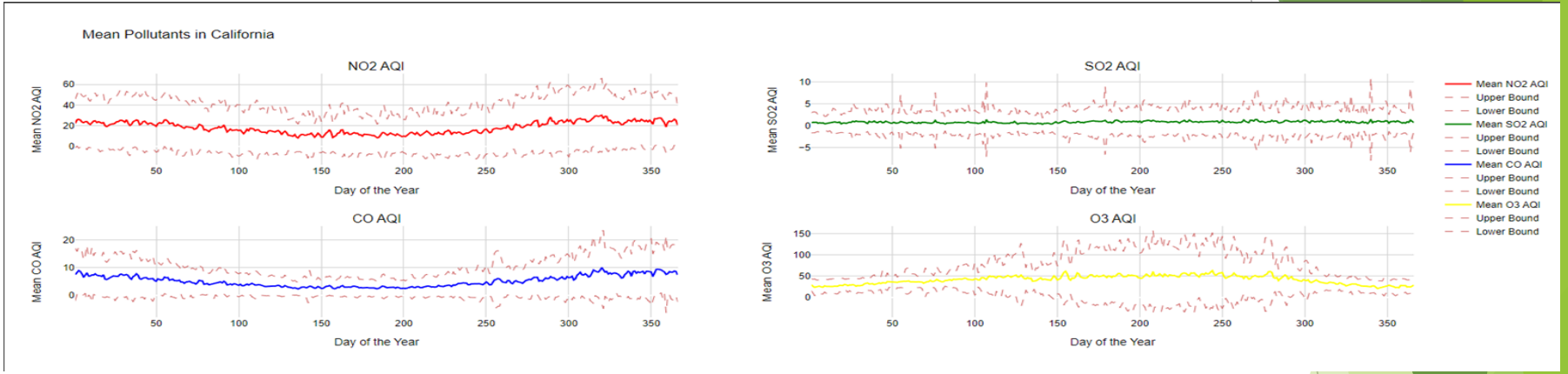
Seasonal Decomposition

- The Ozone values seem to have a steep decrease in the year 2021
- SO2 is clearly decreasing through all the years
- CO and NO2 seem to follow an alternating pattern through the years



Methods - Visualization

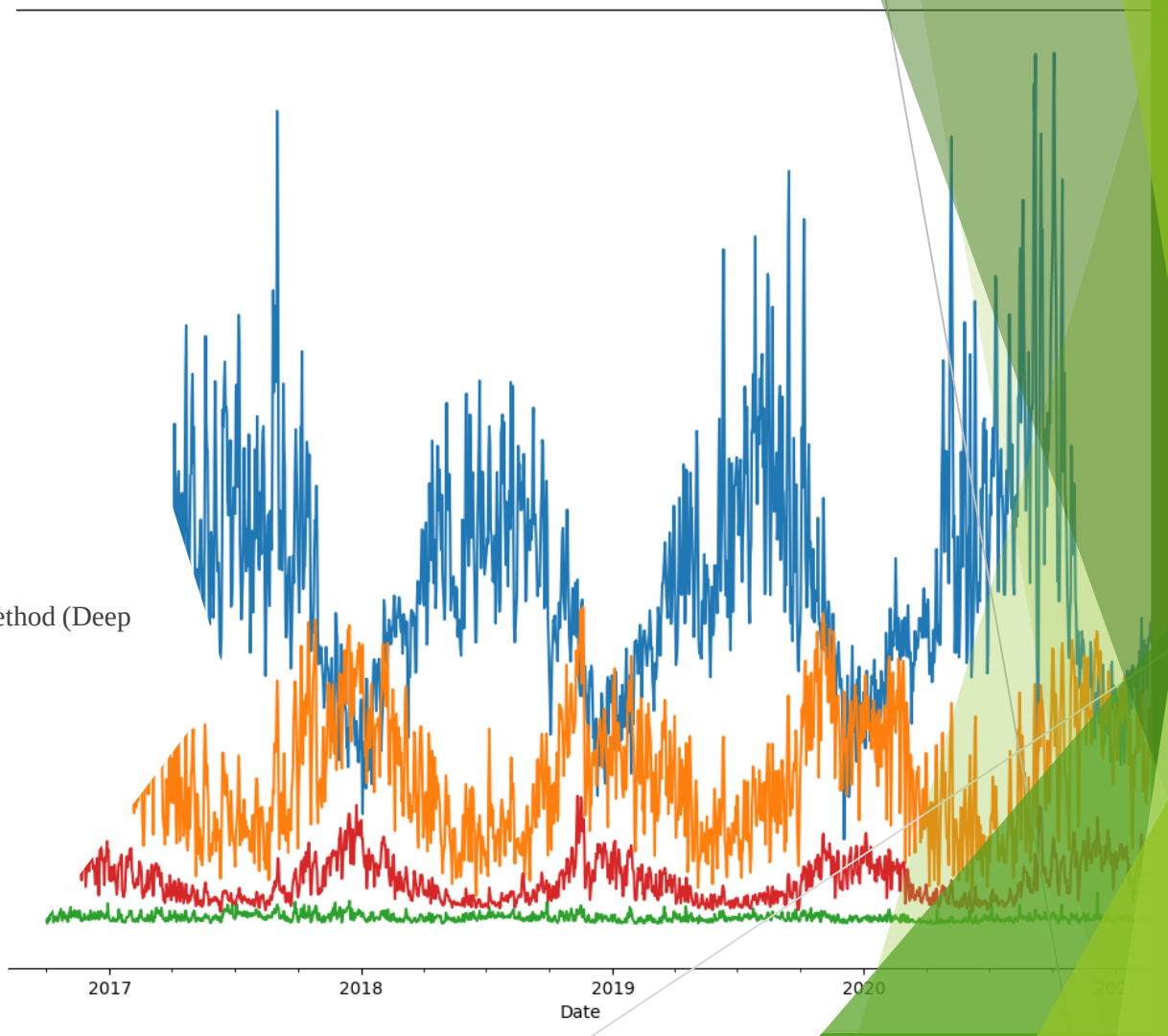
Confidence bands for Pollution data:



- Calculated based on the mean pollution index values calculated on a per-day basis.
- O3 band levels seem to rise in the middle of the year and drop in the beginning and end of a year
- CO band levels seem to follow the opposite trend.
- SO2 seem to have a stable trend throughout the year.

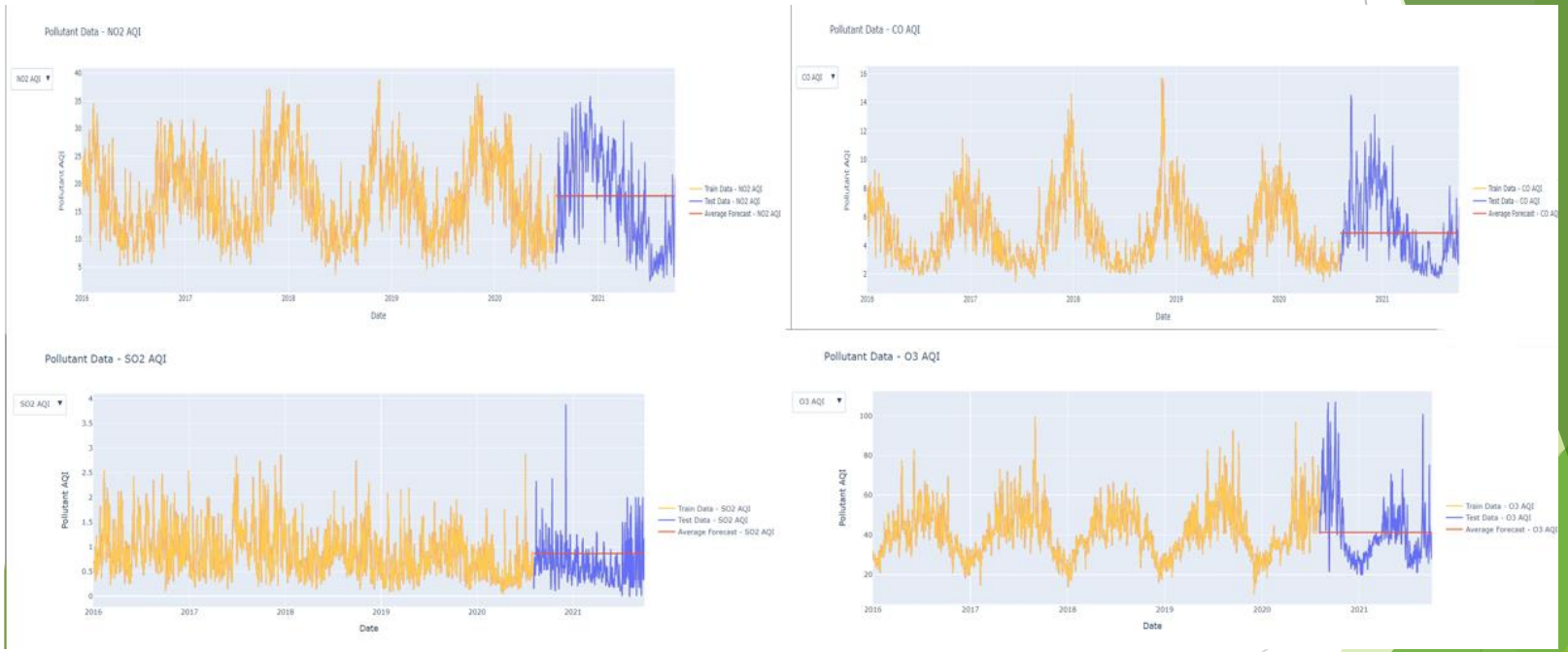
Methods - Modelling

- Average method
- Moving Average method
- Naive model
- Seasonal Naive method
- Long short term memory method (Deep learning based model)

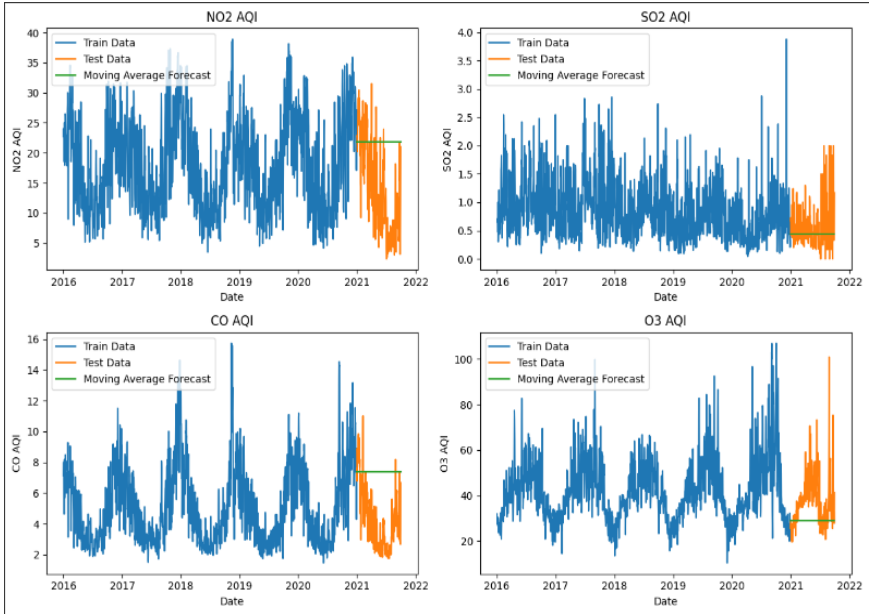


Average Method

- A simplified approach capturing temporal variations.
- Projects the average of all the pollutant AQI values as the predicted values.



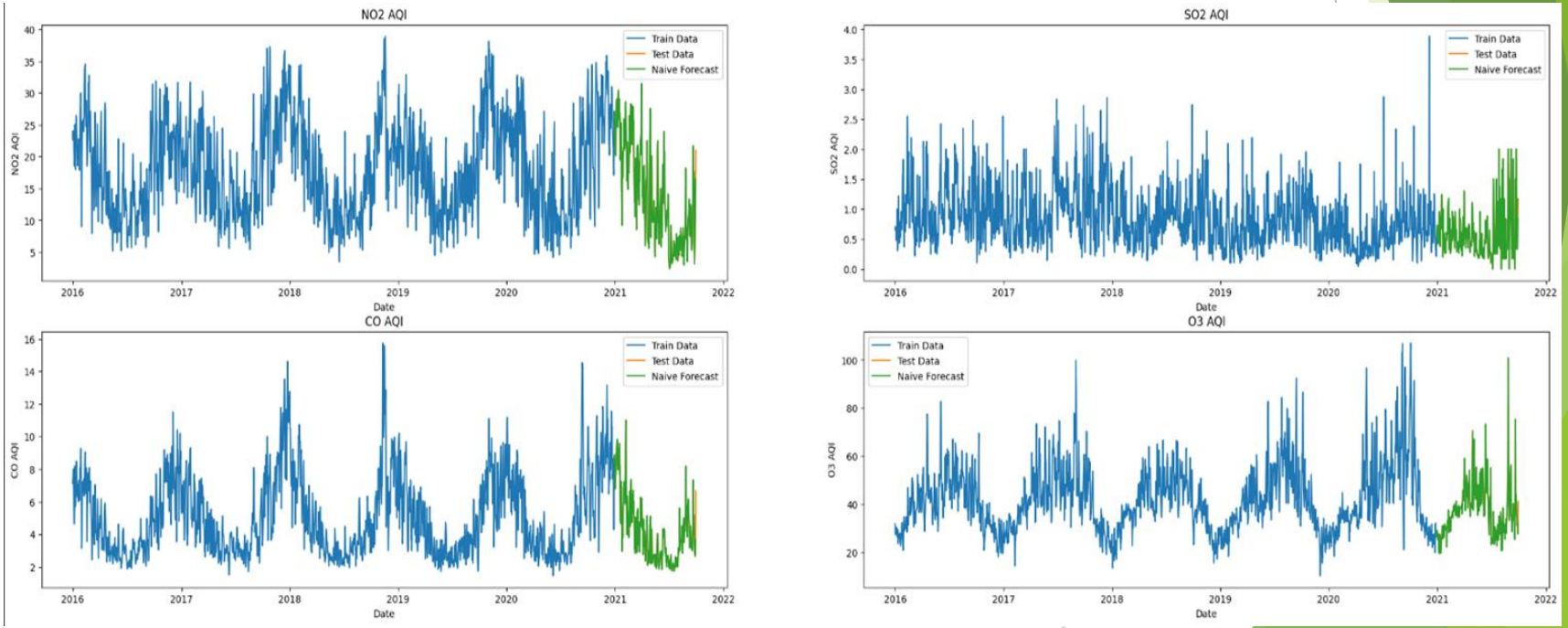
Moving Average Method



- Another simplified by dynamic forecasting method
- Projects the average of Pollutant values from a particular time window.

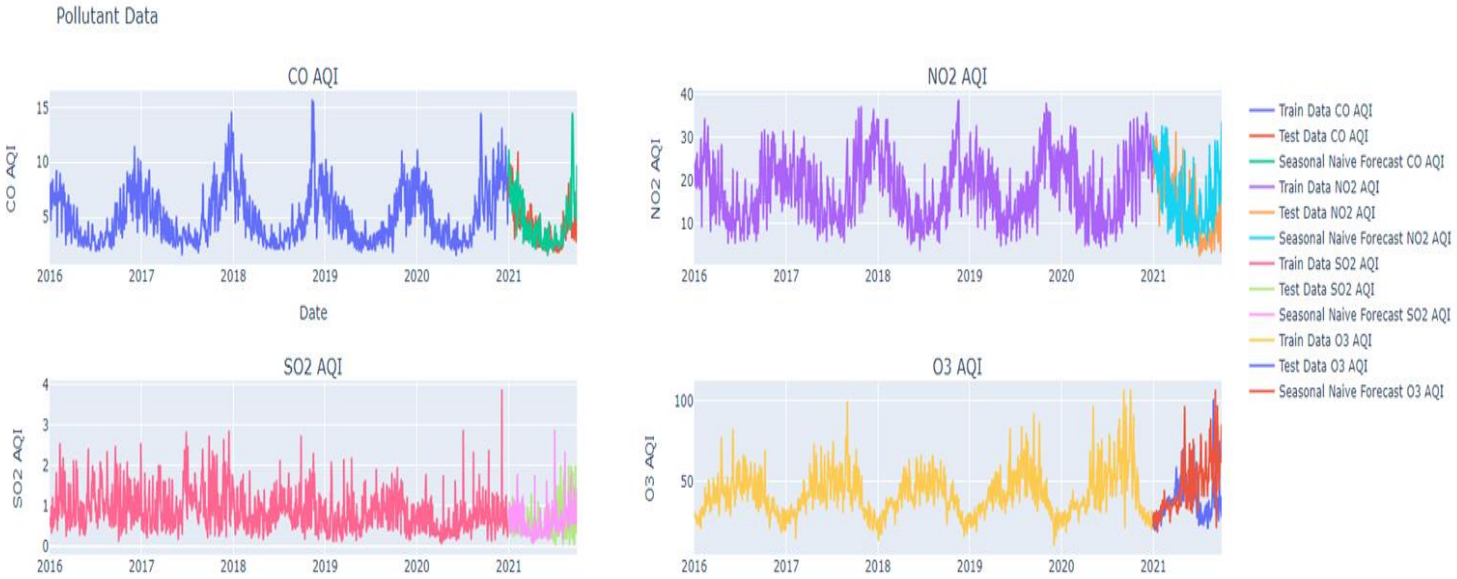
Naive Method

- Simplistic method of prediction
- Takes the previous day value as the current day value.
- Provides poor forecasts



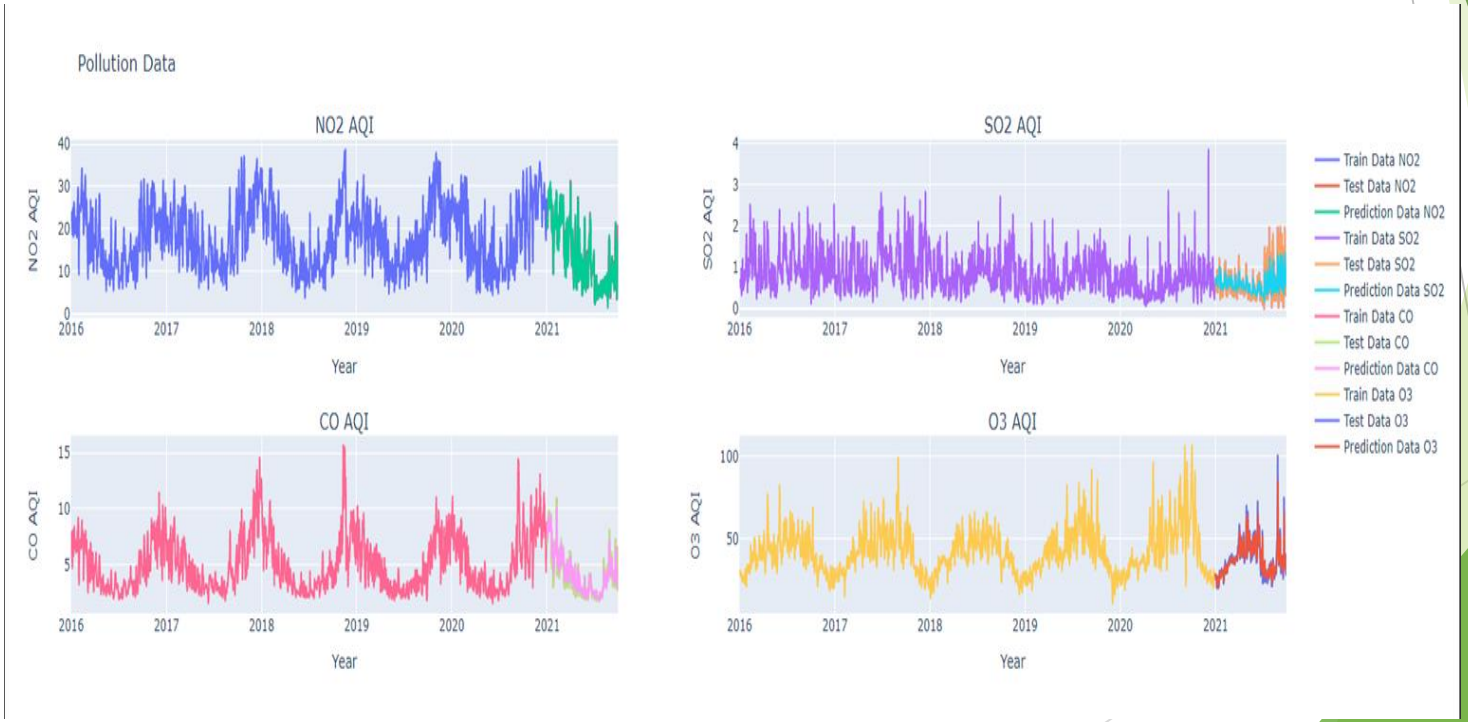
Seasonal Naive Method

- Takes seasonality into account
- Takes the value of the pollutant of the same season but at a different time.
- Prone to errors when there are erratic changes



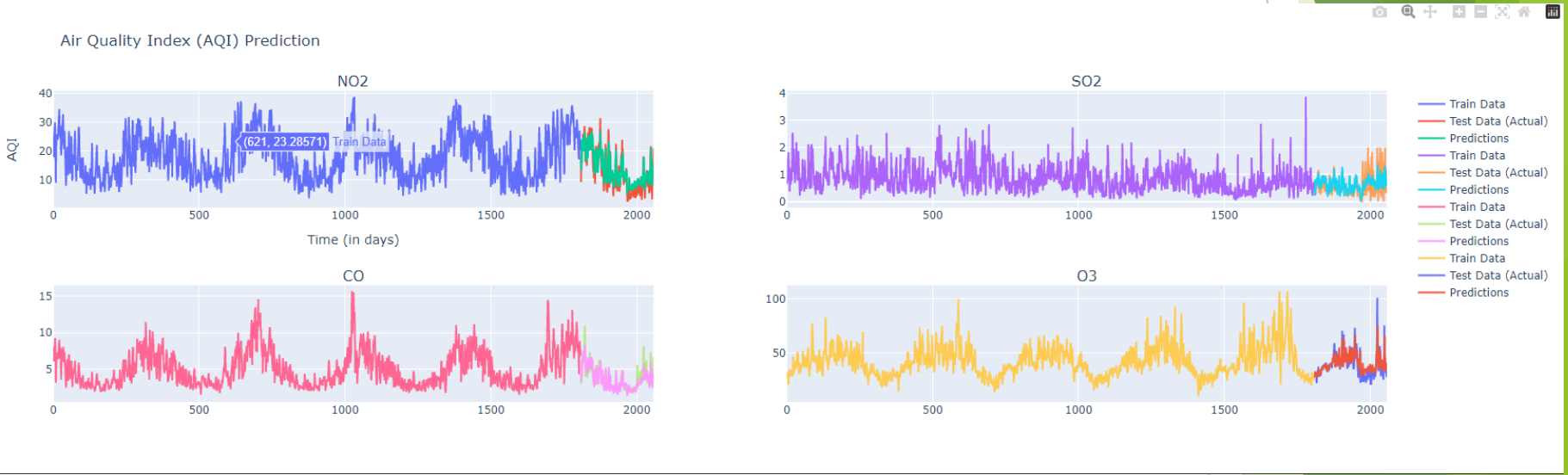
Seasonal Autoregressive Integrated Moving Average

- Useful for seasonal data such as ours
- Model contains components to account for seasonality and exogenous variables
- Could overcome the seasonality in our data by differencing.



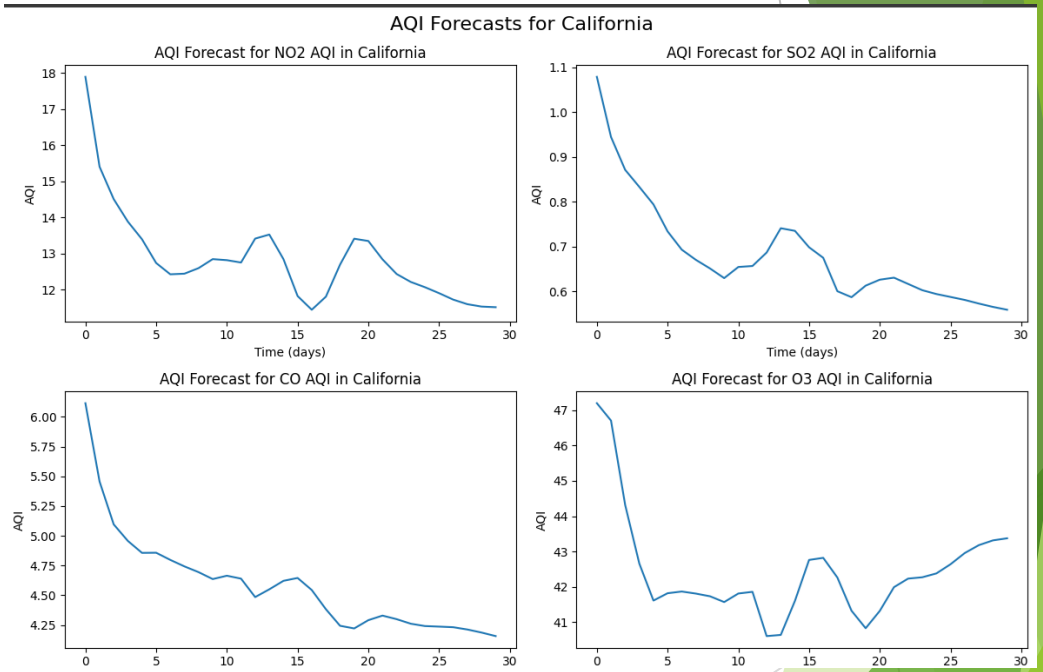
Neural Network Based Long Short Term Memory Method

- Performed well on our data as it was sequential with time.
- Model was scalable and could adapt well to changing patterns.
- Provided accurate predictions and forecasts

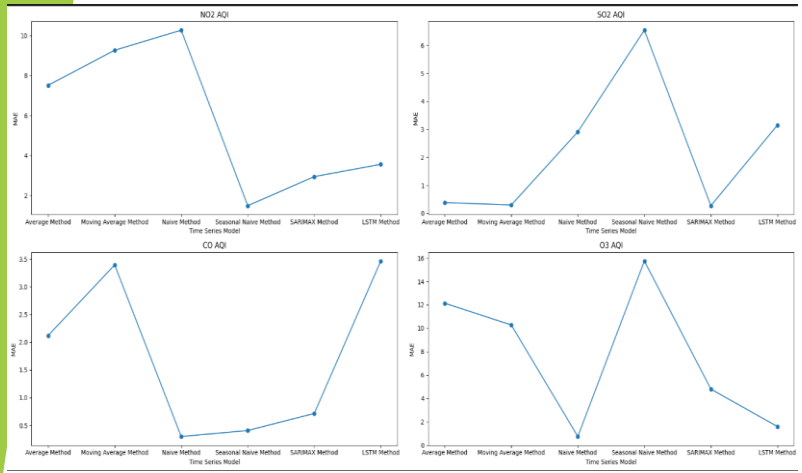


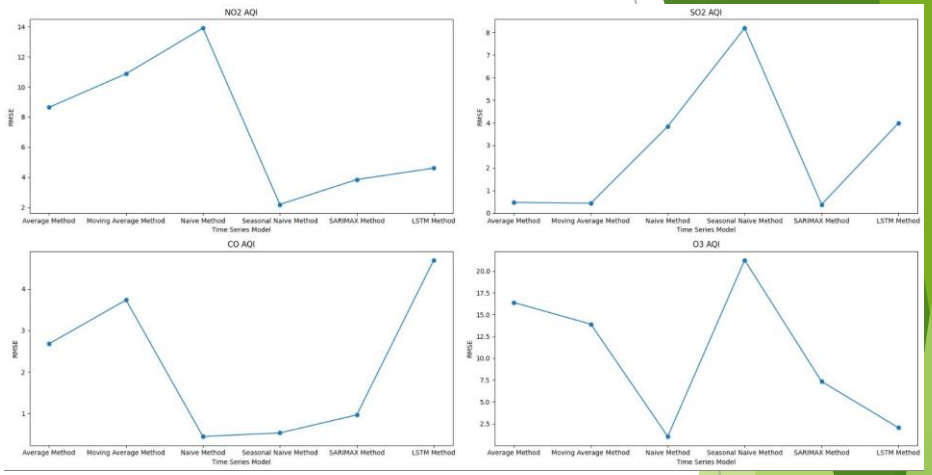
LSTM model forecasting

- O3 levels are expected to decrease initially and then increase again
- CO levels are expected to show a continuous decrease with minimal increase on some of the days.
- SO2 and NO2 show an overall decrease in the AQI values



Model evaluation





Model selection for pollutants

Nitrogen Dioxide	Sulphur Dioxide	Carbon Monoxide	Ozone
SARIMAX	SARIMAX	SARIMAX	LSTM

Discussion Impact and Future work

- **Meaning**
  - Predict the seasonality and observe future trends in pollution data.
  - Identifying top polluters in the region.
- **Impact**
  - Will help select the best time Series model for each pollutant.
  - Can facilitate taking immediate action to curb pollution levels according to seasonal changes
  - Assess the impact of pollution on human health
- **Future work**
  - Analyse more pollutants and find the pollution trend across other regions in the world
  - Advanced modeling techniques to improve pollution predictions
  - Multivariate modeling for environmental condition analysis and prediction.

References

1) https://aqs.epa.gov/aqsweb/airdata/download_files.html#Annual
2) https://otexts.com/fpp2/useful-predictors.html
3) https://machinelearningmastery.com/how-to-develop-a-probabilistic-forecasting-model-to-predict-air-pollution-days/
4) https://towardsdatascience.com/the-complete-guide-to-time-series-analysis-and-forecasting-70d476bfe775