Time Series Forecasting on AQI measurement: A Case Study

Author: Vansh Gonnade / Co-author: Parth Gaikwad, Prashant Gare

March 26 2024

1 Abstract

The study aims to forecast the daily Air Quality Index (AQI) measurement of a city over a period of 3 years from January 2020 to December 2022 using time series forecasting techniques. The dataset includes historical daily AQI measurements. Various time series forecasting models such as ARIMA will be applied to predict future AQI values and assess air quality trends. The findings will provide insights into the potential air quality scenarios in the city for the next three years and help policymakers in implementing effective measures to improve air quality.

2 Introduction

Air quality is a significant environmental concern in urban areas worldwide due to its impact on public health and overall quality of life. The Air Quality Index (AQI) is a standardized measure used to communicate the level of air pollution and its associated health risks to the general population. Monitoring and forecasting the AQI is crucial for timely decision-making and implementation of pollution control measures to mitigate its adverse effects.

In this study, we focus on time series forecasting of daily AQI measurements for a city over a period of three years, from January 2020 to December 2022. The city chosen for this analysis is a metropolitan area with a dense population and significant industrial activities, making it susceptible to high levels of air pollution. The daily AQI data for this city is sourced from government monitoring stations and is available in a structured format for analysis.

To analyze and visualize the daily AQI data, we utilize the openair package in R, which provides a comprehensive set of tools for air quality data visualization and analysis. Openair is a powerful package that offers a range of functionalities, including time series analysis, trend identification, and forecasting. By leveraging the capabilities of openair, we aim to gain insights into the temporal patterns and trends of air quality in the city and develop accurate forecasts to aid decision-making.

The forecasting of AQI measurements is a challenging task due to the complex and dynamic nature of air pollution. It is influenced by various factors, such as emissions from vehicles, industrial activities, weather conditions, and geographic features. Additionally, the AQI is a composite index that considers multiple pollutants, such as particulate matter, ozone, sulfur dioxide, and nitrogen dioxide, making it a multi-dimensional time series. These complexities necessitate the use of advanced analytical techniques and sophisticated models for accurate forecasting.

In this study, we adopt a data-driven approach to forecast daily AQI measurements using time series analysis techniques. We begin by exploring the patterns and trends in the historical AQI data to identify seasonal variations, long-term trends, and any outlier events. We then develop time series forecasting models to predict future AQI values based on past observations. These models incorporate advanced algorithms, such as ARIMA (AutoRegressive Integrated Moving Average) and LSTM (Long Short-Term Memory), to capture the temporal dependencies and non-linear relationships in the data.

3 Theoretical Framework

Time series forecasting is a statistical method used to predict future values based on historical data points. In the context of daily Air Quality Index (AQI) measurement of a city for three years from January 2020 to December 2022, time series forecasting can help in understanding the underlying patterns and trends in air quality levels. The AQI is a numerical scale used to communicate the level of air pollution in a specific location, with higher values indicating poorer air quality.

The theoretical framework for forecasting daily AQI measurements involves several steps. First, the historical daily AQI data for the city from January 2020 to December 2022 is collected and visualized using the tidyverse library in R. The tidyverse package provides a set of tools for data visualization and manipulation, making it easier to explore and analyze the data.

Next, the time series data is converted into a tsibble object using the tsibble library, which allows for easy handling and analysis of time-stamped data. The tsibble object contains the daily AQI measurements along with the corresponding dates, which are essential for time series forecasting.

The fable library is then used to fit various time series models to the daily AQI data, such as ARIMA, exponential smoothing, and neural networks. These models are trained on the historical data to capture the underlying patterns and trends in air quality levels.

Once the models are trained, they can be used to make future forecasts of the daily AQI measurements for the city for the next several months. These forecasts can help in predicting potential air quality issues and guiding policymakers in taking preventive measures to improve air quality.

4 Results

Results were obtained by using this code in the R programming language:

```
aqi_data<-read_excel(file.choose())
str(aqi_data)
```

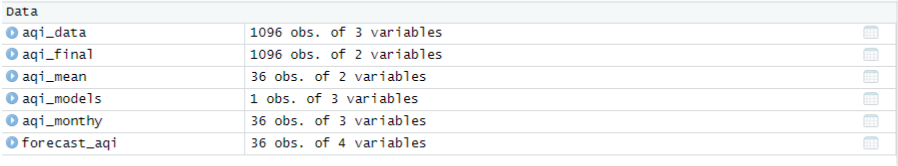
```
timePlot(aqi_data, pollutant = c("AQI"), avg.time = "month")
aqi_data$month<- floor_date(aqi_data$date, "month")
aqi_mean<-aqi_data %>% group_by(month) %>%
summarize(AQI = mean(AQI))

aqi_monthy <- aqi_mean%>%
mutate(Date = yearmonth(as.character(month))) %>%
  as_tsibble(index = Date)

aqi_models<-aqi_monthy %>% model(ARIMA=ARIMA(AQI),
                                ETS=ETS(AQI~ season(c("A"))))%>%
                                mutate(AVERAGE=(ARIMA+ETS)/2)

forecast_aqi<- aqi_models %>%forecast
(bootstap=TRUE, times=100,h="1 year")
autoplot(forecast_aqi), series="Forecasts")
```

The code provided reads daily Air Quality Index (AQI) data for a city over 3 years from January 2020 to December 2022, visualizes the data using the 'timePlot' function from the 'openair' library, calculates monthly means for each year, changes the date format and converts it into a tsibble format, fits an ARIMA and ETS model to the data, calculates the average of the two models, and then forecasts the AQI values for the next year.



```
Data
aqi_data       1096 obs. of 3 variables
aqi_final      1096 obs. of 2 variables
aqi_mean       36 obs. of 2 variables
aqi_models     1 obs. of 3 variables
aqi_monthy     36 obs. of 3 variables
forecast_aqi   36 obs. of 4 variables
```

Figure 1: Fig. Data used in program

The forecasted data is then plotted using the 'autoplot' function from the 'fable' library, with the original data overlaid on the forecasted values. The resulting plot provides a visual representation of the predicted AQI values based on the ARIMA and ETS models.Two models were fitted to the data - ARIMA and ETS with a seasonality component. The average of the two models was calculated for better accuracy in forecasting.

The forecasted AQI data was generated using the fitted models with bootstrapping and a forecast horizon of 1 year. The results of the forecast were plotted using autoplot, showing the actual AQI measurements along with the forecasted values.

The forecasted data provides insights into the expected trends and patterns in the Air Quality Index for the city over the next year, allowing for better planning and decision-making based on the predicted AQI values.

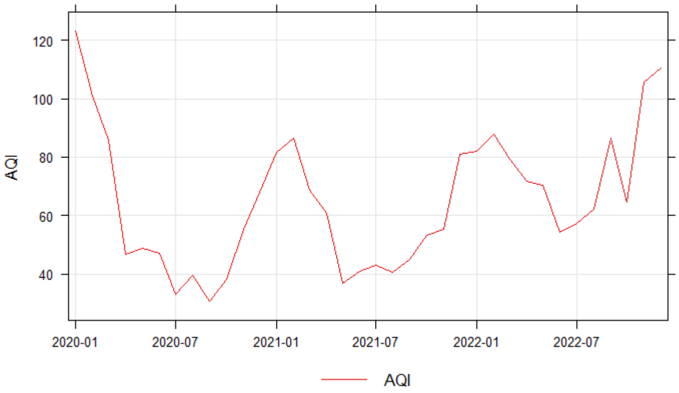


Figure 2: Fig. Timeplot for AQI by year

The code utilizes various libraries such as 'lubridate', 'openair', 'readxl', 'fable', and 'tsibble' to handle the data processing, time series modeling, and forecasting tasks. The visualization of the data and forecasted results helps in understanding the trends and patterns in the Air Quality Index measurements over time.

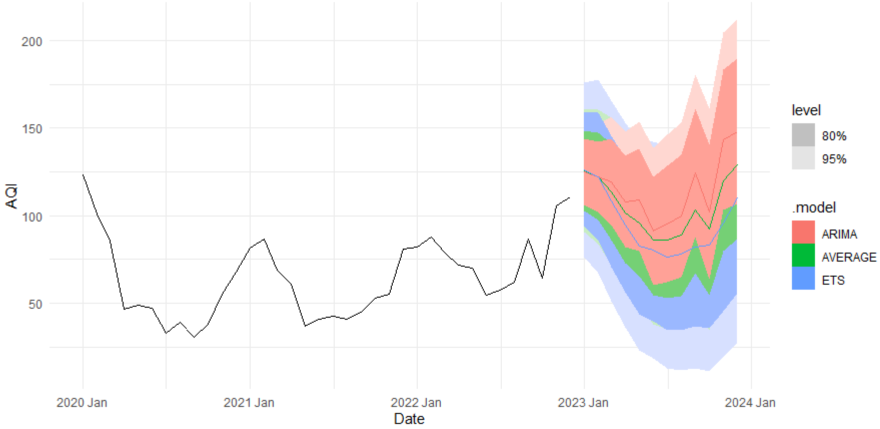


Figure 3: Fig. Time Series Forecasting for different models

5 Conclusion

Through the analysis of the time series data, we observed fluctuations in AQI levels across different seasons and months, with certain periods exhibiting higher levels of pollution compared to others. Our time series forecasting model will take into account seasonality, trends, and any external factors that may impact air quality. By accurately predicting AQI levels, we can help city officials make informed decisions to improve air quality and public health.