Forecasted AQI as a sustainable Policy Making tool

Statistical Applications of R

Group 28

2016IPM004	ADITYA BADONIA
2016IPM070	OSHIN SINGAL
2016IPM094	SHASHANK GIRI
2016IPM118	WELLINGTON DANIEL

Forecasted AQI as a sustainable Policy Making tool

Introduction:

Climate change is a very big problem in contemporary times, and India, as the second most populous country in the world, contributes significantly to air pollution with many cities in the top 50 most polluted cities in the world. But the extent of pollution is not properly studied and used in decision making in India. Air quality monitoring stations are present in many cities but the data isn't used for predictions in most cases, but rather as an observation. Another problem is that these AQI monitoring stations aren't present in small cities. The project aims to build a forecasting model to predict AQI. The project focuses on the city of Delhi, Mumbai and Chennai which are major cities in India.

Literature Review:

Currently similar models developed do not clearly explain and predict AQI, and research in this area has only picked up pace in the past few years according to our literature review. On Zhijun Yan and Yan Liu's study of seasonal effects of seasonal changes and AQI in Beijing, they only had an adjusted R squared value less than 0.2 while Arun Solanki's paper in the Journal of Xi'an University of Architecture & Technology shows better results with a SARIMAX model. Since Temperature, Humidity and other meteorological variables are measured in almost every city and are easily available, we will work towards building a model which can use these data to predict AQI. In this project, we build a model that can use past research on AQI and easy to access data to predict the quality of air. Through our research we found some research done with deep learning models and Neural networks (Du, Zhehua & Lin, Xin) but we have not used those techniques in our project due to the lack of high computing power.

Data:

Meteorological data: This data is collected through a web scraper built in R from the website " https://en.tutiempo.net/ ". The web scraper was built by us and will scrape day wise measures of meteorological variables.

Air Quality Data: This data is downloaded from Kaggle which is originally from the Pollution Control Board of India (CPCP) " https://www.kaggle.com/rohanrao/air-quality-data-in-india ".

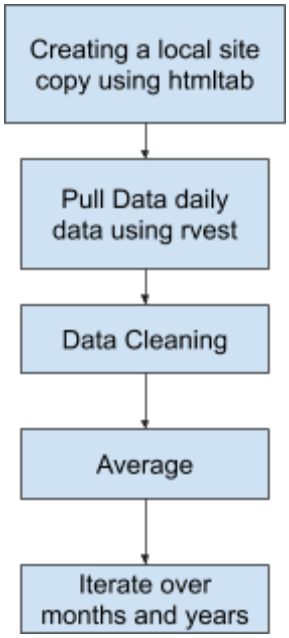
World pollution data: This data was downloaded from WHO Global Ambient Air Quality Database - update 2018.It had location wise pollution data for countries. "https://www.who.int/airpollution/data/cities/en/ ".

Latitude Longitude countries: This Data was downloaded from Kaggle.Since goggle's ggmap is no more free. This was a better option than buying an API.
"https://www.kaggle.com/max-mind/world-cities-database?select=worldcitiespop.csv ".

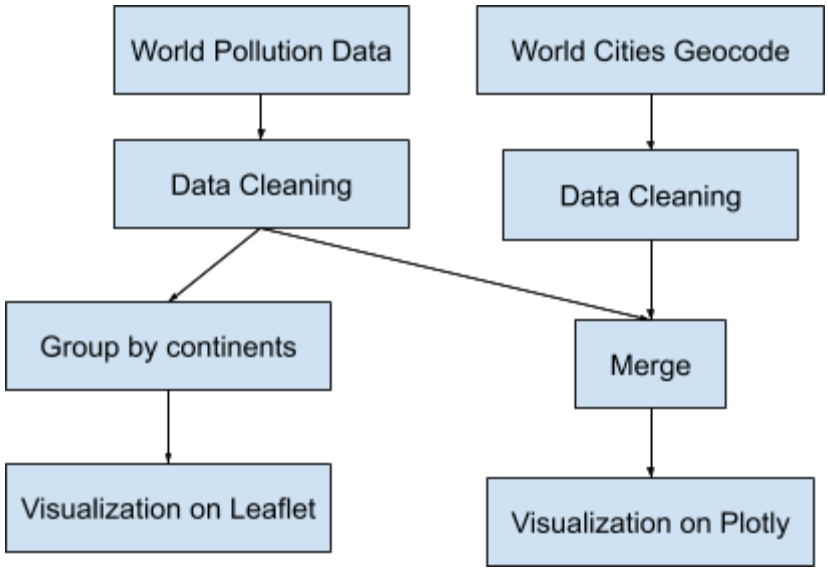
Latitude Longitude Indian cities: This data had location (geocode) of different Indian cities. This is matched with existing pollution data for visualizations.
"https://simplemaps.com/data/in-cities ".

The Process:

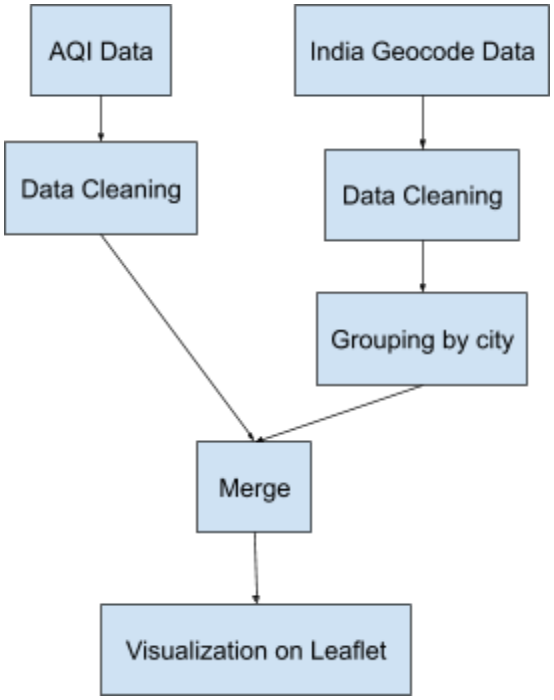
Web Scraping



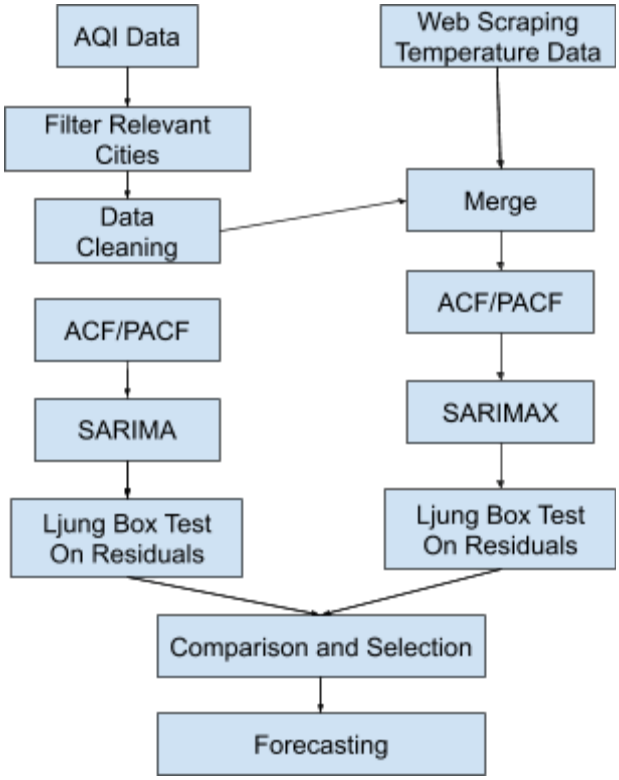
Visualization 1



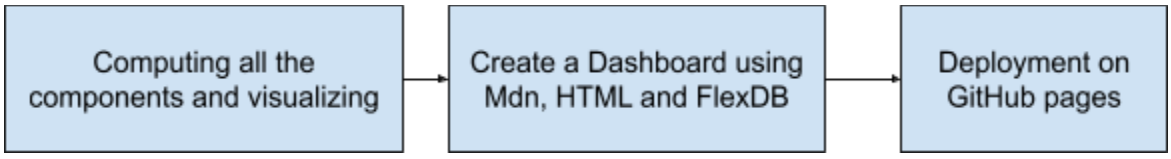
Visualization 2



Model Fitting



Visualization and Deployment



Statistical and Mathematical Concepts behind the forecasts

Autocorrelation Function:
The Autocorrelation function is the correlation between a Random time varying variable and its own value in the past. The ACF plot can help understand the nature of the variable with respect to time. With the ACF plot we can understand if the variable has a Moving Average component i.e., the variable is dependent on its errors in the past.

Partial Autocorrelation Function:
The Partial Autocorrelation function helps understand if a variable has an Autoregressive component. It is the coefficient of regression plotted against the value at present and value in the past**.**

SARIMA:
SARIMA is a model which has both Autoregressive and Moving Average components and also has a trend (Integrated). The difference between ARIMA and SARIMA is that the SARIMA model has a seasonal component as well. The seasonal component is calculated similar to the ARIMA components but with differencing done in the length of the seasonality.

Notation: SARIMA (p,d,q)(P,D,Q)
P,p: AR level for season and daily values respectively.
D,d: MR level for season and daily values respectively.
Q,q: Integration (I) level for season and daily values respectively.

SARIMAX:
SARIMAX is a model with all the components of SARIMA but it also includes exogenous components. In this project the exogenous component X is taken as the meteorological components. This will ideally give a better result but not if there is high variability in the exogenous data itself.

Ljung - Box Test on residuals:

Ljung Box test on Residuals check if, after fitting the model, the residuals are just white noise i.e. no predictability.

The Ljung Box test on the Residuals squared helps us to understand if there are conditional heteroskedasticity.

ARCH / GARCH:

ARCH/GARCH models help in understanding the uncertainties in the predictions. They fit a parametric model on the variances of the past. In our project we observed our final model to be conditionally heteroscedastic but we did not fit a GARCH model as it would further complicate the model and it will be hard for anyone to interpret.

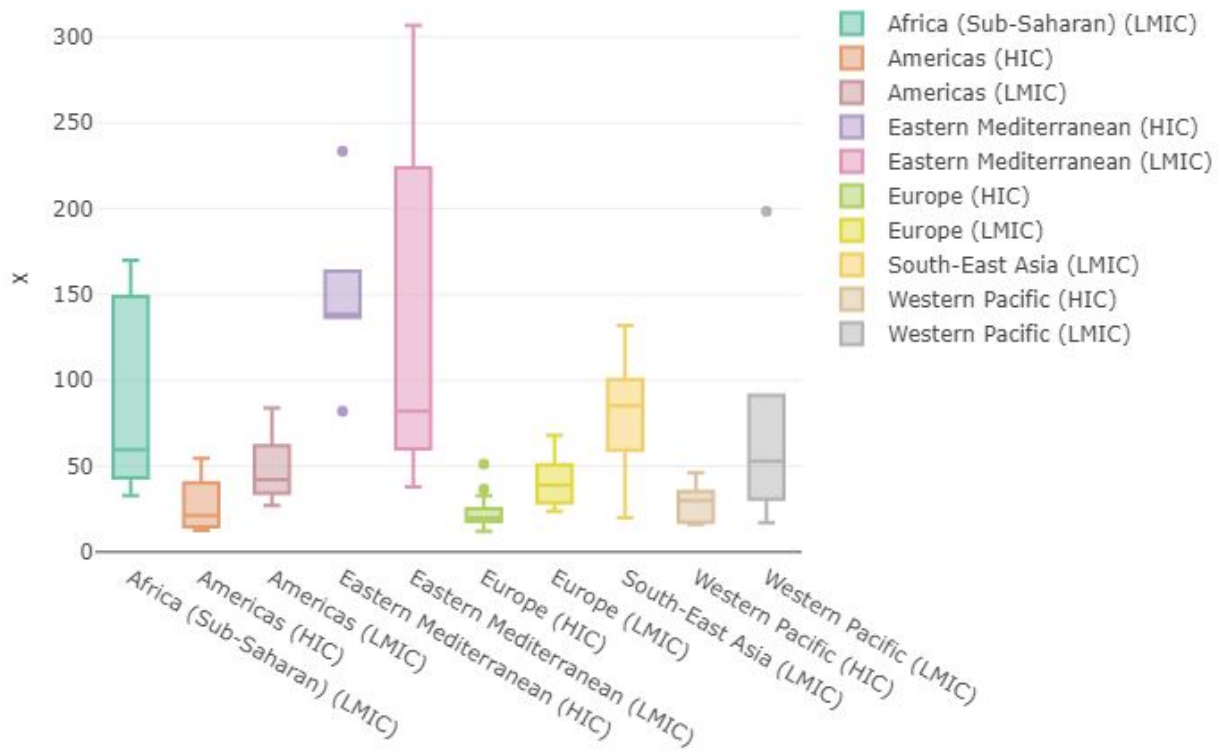
Mean Absolute Percentage Error:

MAPE is an error measure that we will be using to choose the best model among the models. We also have robustness of the model as a criteria. Therefore we will be looking for that as well.

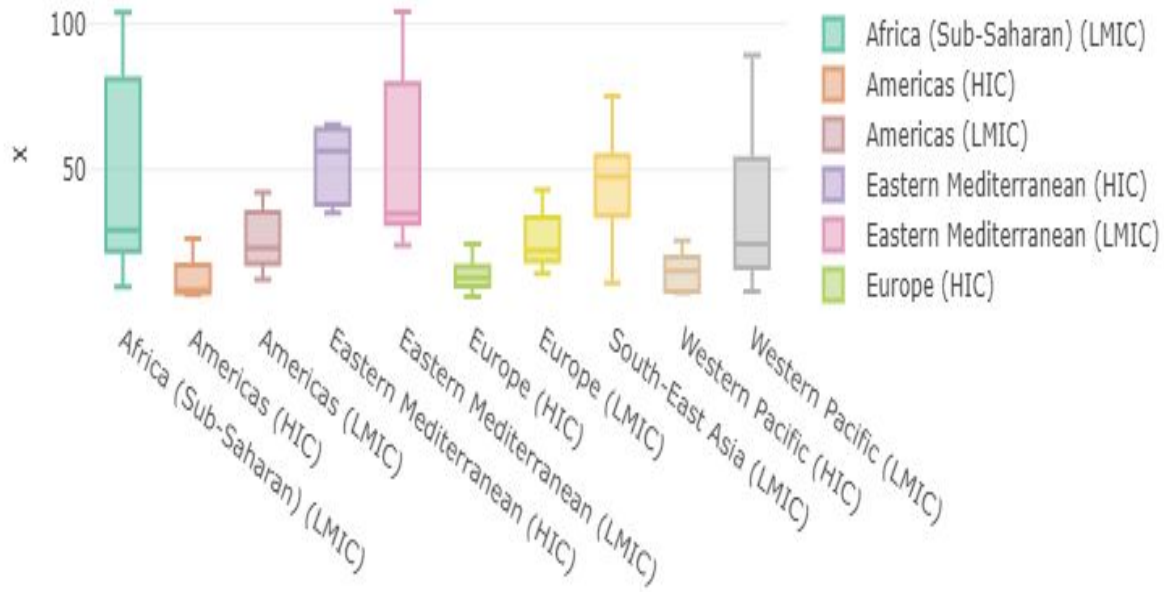
Visualizations and Results:

We have listed below some of the many visualizations we have done. For all the visualizations visit, https://lwellingtondaniel7.github.io/SAR-Project/#overview

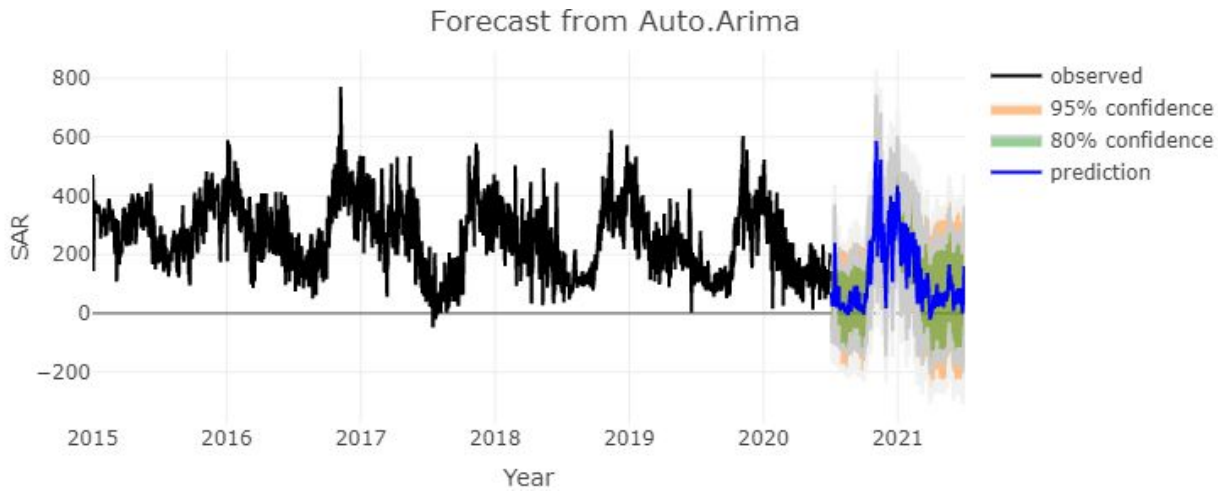
PM 10 pollution across continents



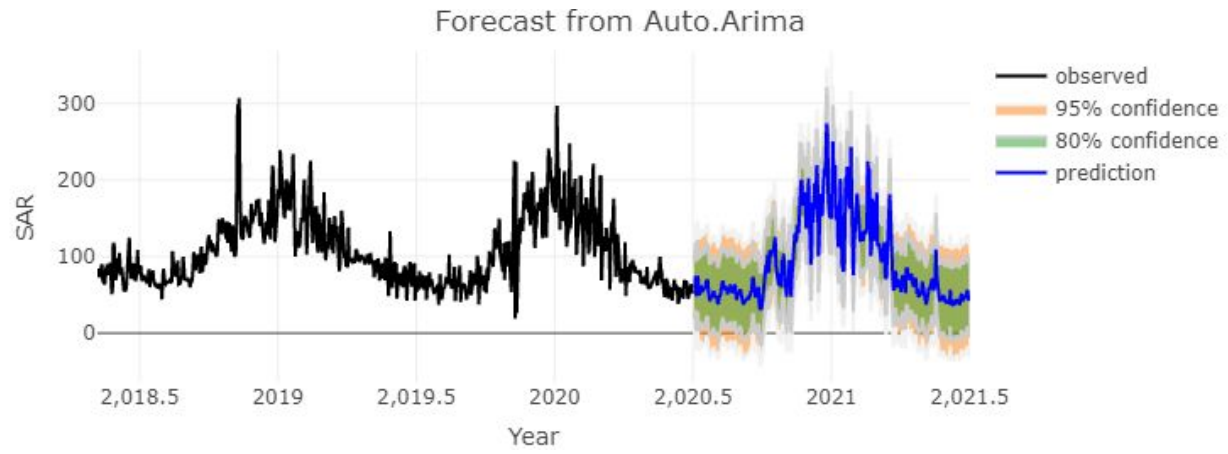
PM 2.5 pollution across continents



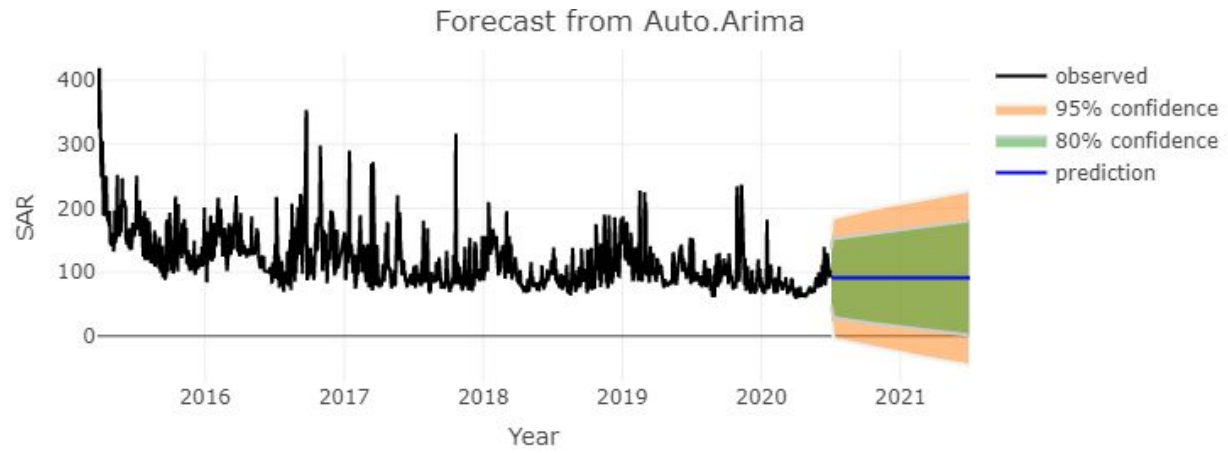
Delhi Forecast:



Mumbai Forecast:



Chennai Forecast:



Models:

Delhi: SARIMA(1,1,2)(0,1,0)

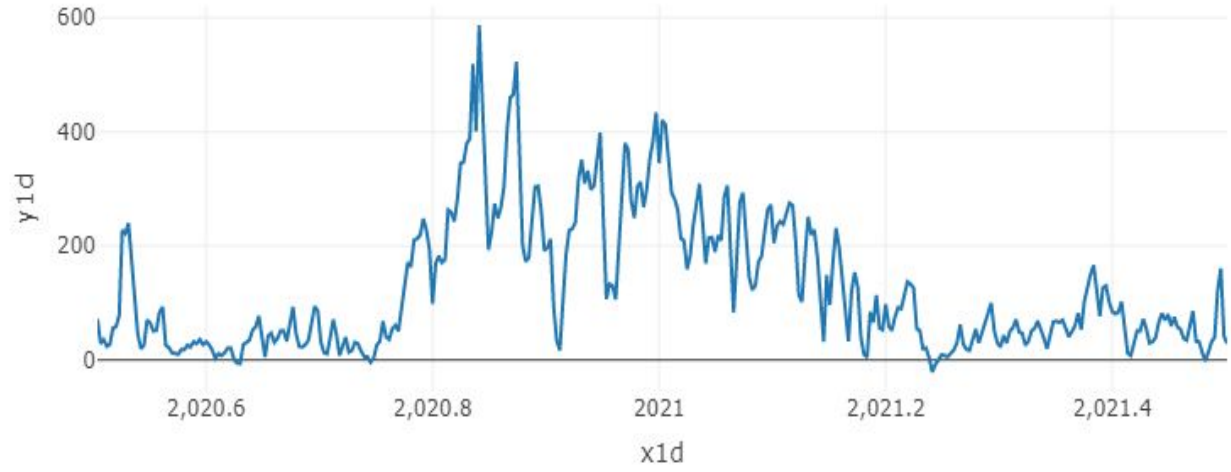
Mumbai: SARIMA(1,0,0)(0,1,0)

Chennai: SARIMA(4,1,3)

Inference and Result:

We have chosen SARIMA models for forecasting as the SARIMAX models with the exogenous variables gave us large error bands and many x variables were insignificant (p value<2). One of the ways to tackle this would be to use a smoothing technique to remove variability in the exogenous variables. We haven't done that as it will complicate the model and also it will result in loss of data. The Chennai AQI data we had had a lot of missing values. Therefore we were only able to forecast for Delhi and Mumbai. Further forecasted values of X also have uncertainties. **Vector Autoregression** is a possible solution to try.

Delhi Forecast:



Mumbai Forecast:

Residual Analysis:

The Residual analysis showed conditional heteroscedasticity (Low p value for Res^2). We have not provided a variance estimate as it will complicate the understanding. But this can be found using the “fgarch” library in R.

Packages Used:

library(htmltab) - This library was used to deal with webpages.
library(flexdashboard) - The flexdashboard library was used to create the final dashboard.
library(tidyverse) - This library was used for some data manipulation function.
library(readxl) - This library was used to read data from excel.
library(ggplot2) - This was used to plot better ACF, PACF plot.
library(tsbox) - This is a library that helps in creating and manipulating ts objects.
library(stR) - This has some useful string functions for data cleaning.
library(quantmod) - This library has the ARIMA, ACF, PACF modeling functions.
library(tseries) - This library also helps in Time Series manipulation with the ts and zoo.
library(forecast) - This has forecast functions for auto.arima.
library(dygraphs) - This library has some interactive plots.
library(leaflet) - This library helps to plot geographical markers on maps.
library(plotly) - This library has some interactive plots.

Conclusion:

In this project we have identified the SARIMA function using Hyndman & Khandakar algorithm in auto.arima to be the best model for our problem. Due to the limitations of computing power and relatively low quality of data, we were only able to forecast for Delhi and Mumbai. With distributed computing, this can be overcome. The model can be automated and scaled with better quality data to make an interactive dashboard for policy makers to make decisions and to track climate change and fluctuations. Technologies like HDFS, Map Reduce with R can be on a cloud service provider used for scaling.

The source code and the entire project is available on
https://lwellingtondaniel7.github.io/SAR-Project/#overview .

References:

1. Arora, Himanshu & Solanki, Arun. (2020). Prediction of Air Quality Index in Metro Cities using Time Series Forecasting Models. Xi'an Jianzhu Keji Daxue Xuebao/Journal of Xi'an University of Architecture & Technology. 12. 3052-3067. 10.37896/JXAT12.05/1721.
2. Wu, Lifeng & Gao, Xiaohui & Xiao, Yanli & Liu, Sifeng & Yang, Yingjie. (2017). Using grey Holt–Winters model to predict the air quality index for cities in China. Natural Hazards. 88. 1-10. 10.1007/s11069-017-2901-8.
3. Dun, Meng & Xu, Zhicun & Chen, Yan & Wu, Lifeng. (2020). Short-Term Air Quality Prediction Based on Fractional Grey Linear Regression and Support Vector Machine. Mathematical Problems in Engineering. 2020. 1-13. 10.1155/2020/8914501.
4. Ikram, Maria & Yan, Zhijun & Liu, Yan & Qu, Weihua. (2015). Seasonal effects of temperature fluctuations on air quality and respiratory disease: a study in Beijing. Natural Hazards. 79. 10.1007/s11069-015-1879-3.
5. Ikram, Maria & Yan, Zhijun & Liu, Yan & Qu, Weihua. (2015). Seasonal effects of temperature fluctuations on air quality and respiratory disease: a study in Beijing. Natural Hazards. 79. 10.1007/s11069-015-1879-3.
6. Zheng, Yu & Yi, Xiuwen & Li, Ming & Li, Ruiyuan & Shan, Zhangqing & Chang, Eric & Li, Tianrui. (2015). Forecasting Fine-Grained Air Quality Based on Big Data. 2267-2276. 10.1145/2783258.2788573.
7. Du, Zhehua & Lin, Xin. (2020). Air Quality Prediction Based on Neural Network Model of Long Short-term Memory. IOP Conference Series: Earth and Environmental Science. 508. 012013. 10.1088/1755-1315/508/1/012013.