Comparison of Air Quality between two cities and Predicting the Air Quality concentration using Time Series Predictive Analysis

Siddharthan Saravanan*

04/08/2021

Abstract

*C0020581

Introduction

Over the last few years, air pollution in urban areas has steadily increased. The air is becoming increasingly harmful at an alarming rate as a result of growing industrialization and rising levels of harmful chemicals in the atmosphere. Since the outbreak of the Coronavirus pandemic, it has become more important to limit air pollution in order to mitigate the virus's effects. According to the WHO, 91% of the world population is breathing polluted air, which is one of the leading factors in the development of non-infectious disease. The UK Government through the Department for Environment, Food and Rural Affairs (Defra) manages the Ambient Air Quality by controlling the source of emission and by monitoring and managing the air quality.

Urban observatories have been established in several locations around the United Kingdom as a result of rapid technology advancements in IoT and sensor development. These urban observatories collect vital data about their community on a variety of topics, from air quality to traffic tracking, where environmental and human-caused data is gathered, stored, and evaluated for a variety of purposes. The number of topics and focus themes vary according on the locations and observatories. Manchester Urban Observatory [1] and Newcastle Urban Observatory [2] are two of the most active observatories in the United Kingdom that are being considered for this project. Despite the fact that both observatories have a few shared elements, this study considers air quality data collected by sensors in both locations.

Sensors with an Air Quality theme collects the data on characteristics that can be used to determine the location's air quality. Multiple sensors have been deployed across the city, and data is collected continuously 24 hours a day, seven days a week. The data observatories records the gases emissions and substances that are toxic for human beings. The most harmful emissions that are recorded through the sensors are tropospheric ozone (O_3), sulpher dioxide (SO_2), nitrogen dioxide (NO_2), Benzene, suspended particulate matter ($PM_{2.5}$ and PM_{10}), Lead, Carbon monoxide, PAH and 1,3-Butadiene. These are the Target Essential Variables (EV) established as the UK Government's National Air Quality Strategy (NAQS).

Project Objectives

The goal of this research is to compare, visually represent and predict the Air Quality of Manchester and Newcastle upon Tyne using data from their respective Urban Observatories. The EV's as established by Defra, only the Suspended Particulate Matter PM_{10} and Nitrogen Dioxide NO_2 are considered as part of this research.

More specifically the objective of the project includes:

- Data preparation and visual analysis of air quality data collected over a two-month period in Manchester and Newcastle upon Tyne.
- Comparing the variations and findings in air quality over time.
- Extend the data from two months to two years to better understand the impact of COVID19 on air quality — before lockdown, during lockdown-1, after lockdown-1, during lockdown-2, after lockdown-2.
- Analyze the data as a time series to better identify seasonality trends and extend the analysis to forecasting air quality data.

Literature Review

Methodology

Data Analysis

Data Understanding

The Air Quality data of both Manchester and Newcastle city are available for public use in their respective Urban Observatories. The Newcastle Urban Observatory monitors and records the data at the minutes granularity, whereas the Manchester Ubran Observatory monitors and records the data at 5-minute granularity. Further the Newcastle Urban Observatory has the feature of accessing the dataset through API's whereas the data for Manchester city should be manually downloaded from the Manchester Urban Observatory.

The Manchester city data have been downloaded manually from the Manchester Urban Observatory portal and postcodes for each sensor in the city is manually updated as part of the data preparation. The final data after data preparation step consists of 5 variables.

Table 1: Dataset description of Manchester

Variable Name	Class	Information
Timestamp	Character	Time series of the data recorded
NO2	Numeric	Value of the sensor recorded
PM10	Numeric	Value of the sensor recorded
StationName	Character	Station name which the sensor is located
PostCodes	Character	Location of the sensor

The Newcastle city data have been downloaded from the Newcastle Urban Observatory portal through API. The final data after data preparation step consists of 5 variables.

Table 2: Dataset description of Newcastle

Variable Name	Class	Information
Sensor.Name	Character	The name of the sensor which records data
Variable	Character	The Essential Variables from the data
Timestamp	Character	Time series of the data recorded
Value	Numeric	Value of the EV's recorded
Location..WKT.	Character	Location of the sensor

Data Preparation

The R has been used predominantly for pre-processing the data through the Project Template Package. During the initial stages of the research, February month's data of the year 2020 is considered and the time period is expanded after completing the data pipeline for the analysis. The Manchester city data has been downloaded manually and the data is stored as separate files based on each station location. The postcodes of each station was manually updated which helps in the future analysis process. The dataset has been binded together as a single data file using Python. The combined dataset have been stored inside the data folder of the R's Project Template.

The Newcastle data has been downloaded through the API from Newcastle Urban Observatory portal and the dataset has been stored in the data folder of the Project Template. Both the cities dataset has been

wrangled by removing the N/A values. The data has been pre-processed by assigning valid data types for the variables, conversion of postcodes into Latitude and Longitude for geospatial visualization of the EV's concentrations. The variable Timestamp of both the dataset was in text format and hence they are converted into date format and new variables "date" and "hour" has be obtained from the Timestamp variable.

Visual Representation

The datasets obtained from the step 3.1.2 has been stored in a way such that one file holds the NO_2 concentration among Newcastle and Manchester city and one file holds the PM_{10} concentration among Newcastle and Manchester city. Tableau has been extensively used for visual representation by creating dashboards. The visualization has been generated by three different granularity level which helps to better understand and visually view the outcome in each each cities.

Geospatial Representation

The dataset contains the location of each sensor that has been placed in each city. For Newcastle city, the dataset contains the Latitude and Longitude values of each sensor location in a single "Location..WKT." column. During the Data mining process the column is split in-order to get the Latitude and Longitude separately. While for the Manchester city, instead of the Latitude-Longitude information only the postcode informations where available in the Manchester Urban Observatory website. Hence, the postcodes are entered manually for each station and these postcodes are then converted into Latitudes and Longitudes through the Data Mining process.

Through Tableau the

Statistical Analysis

The statistical analysis has been performed over the two cities datasets to statistically understand the Air Quality difference between the cities over the EV's. Hypothesis testing has been introduced for performing the statistical analysis which compares the quality between the cities.

Welch T-Test The Welch T-Test was performed between the two EV's - NO_2 and PM_{10}. The dataset obtained from step 3.1.2 has been used from which the dataset has been filtered and stored separately for each cities. The values of NO_2 and PM_{10} of each cities are stored separately in an array from which Welch T-Test was performed.

Predictive Analysis

The FB-Prophet model has been used for forecasting the Time-Series data of both Manchester city and Newcastle City's Air Quality dataset. A function has been created for fitting the model with the desired training data which would be good for predicting the Air Quality. The model has been trained from July 1ˆst 2019 to November 30ˆth 2019 by setting up the daily seasonality to True. The model has been built in such a way that it would forecast for the next 30 days. The MSE, RMSE, MAE, MAPE and MDAPE are the performance metrics that are computed from the forecasting FB-Prophet model. The seasonality trend plots have also been generated which represents the trend lying behind the dataset.

Results

Visual Representation:

The dashboards are generated through Tableau under different granularities of the data for both NO_2 and PM_{10} EV's. The graphs are represented in two different colors for each cities. The plots are generated based on hourly granularity of each week, daily granularity and weekly granularity.

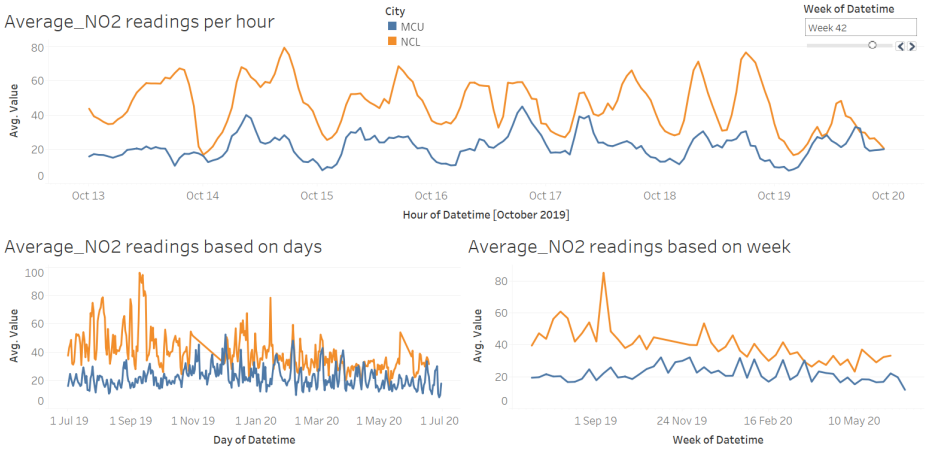

Figure 1: Dashboard Visualisation of NO_2

The **Figure 1** represents the dashboard generated for NO_2 concentration across the timeline.

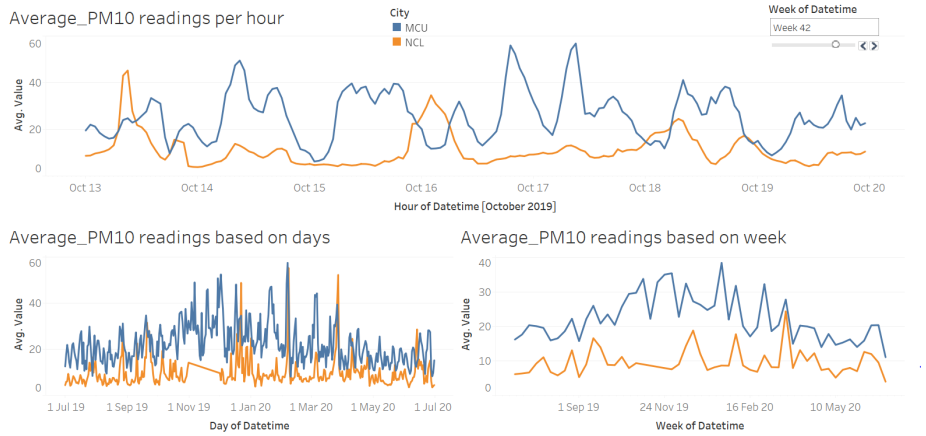

Figure 2: Dashboard Visualisation of PM_{10}

The **Figure 2** represents the dashboard generated for PM_{10} concentration across the timeline.

Geospatial Representation:

The **Figure 3** represents the geospatial visualization based on the location of the sensor. The color representation is based on the average concentration value of the Air Quality.

The **Figure 4** represents the geospatial visualization based on the location of the sensor. The color representation is based on the average concentration value of the Air Quality.

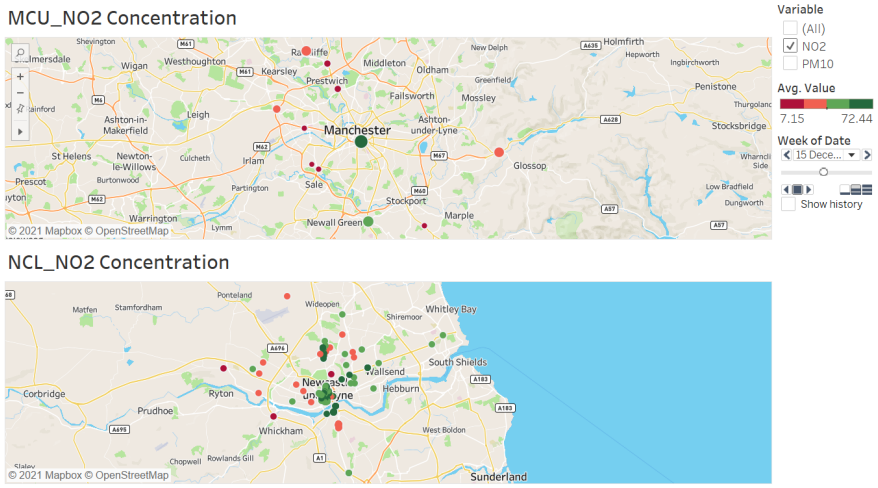

Figure 3: Dashboard Geospatial Visualisation of NO_2

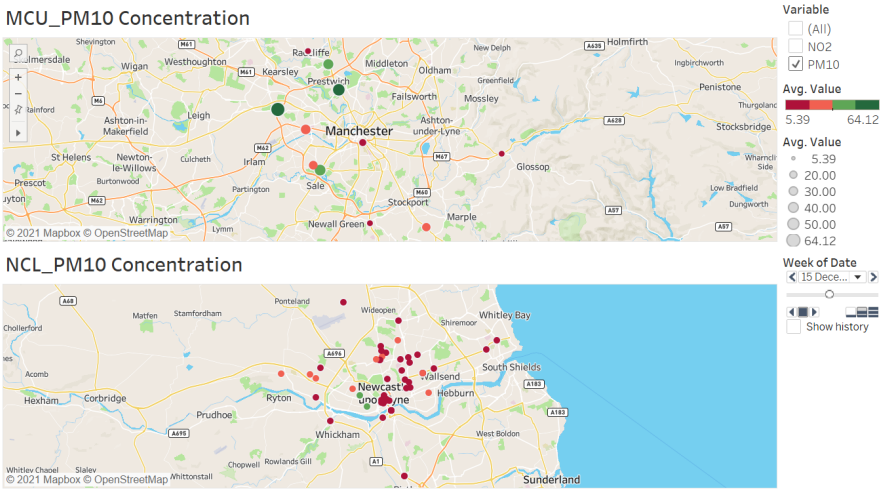

Figure 4: Dashboard Geospatial Visualisation of PM_{10}