Assessing Urban Air Quality in Major English Cities Using Geospatial Analysis and Python

Akshay Suresh Varma

23252473

Department of Computer Science
Maynooth University
CO. Kildare.

MSc in Data Science and Analytics

NCG616 – Dissertation

12/08/2024

Table of Contents

1. Introduction:

Air pollution has become a major concern in urban environments, impacting people and the surrounding population. A good strategy for the mitigation of air pollution should be informed by the assessment of the extent of air pollution and the distribution of the pollutants in space. Technologies related to geography and space together with the use of tools based on Geospatial technologies, and Python programming language allow for efficient data collection, analysis, and visualization of air quality indicators. The air quality of England cities such as London, Manchester, Birmingham, Bristol, etc. are evaluated in this study. The entire analysis is implemented in the Google Colab platform.

1.1 Background

The increased rates of urbanization and industrialization caused air quality in the cities to have highly deteriorated. Airborne pollutants like PM, NO2, and SO2 are prevalent in urban environments and are hazardous to health through respiratory and cardiovascular diseases (Zangari et al. 2020). Fixed station networks serve well but scientists and policymakers are increasingly interested in air quality at the major scale.

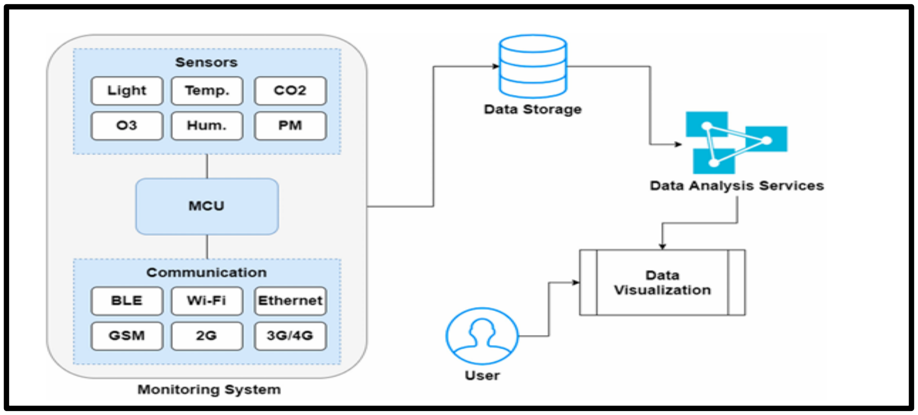


Figure 1: Air quality monitoring system

(Source: Saini et al. 2020)

The empirical research project plans to employ geographical information systems and other analytical tools to develop complex models of air quality in cities with the hope of establishing the location of polluting industries and areas vulnerable to pollution. Due to Geo pandas, a type of

Python library, it is easy to combine and work with data characterizing the air quality and geographical information (Saini *et al.* 2020). Matplotlib can only be used in the creation of static charts and graphs, which means that it can only be used in an analysis of a given data set whereby the data is made more comprehensible through charts. Folium, on the other hand, can be used to design maps and this makes the data more useful through the maps.

1.2 Aim and Objectives

Aim:

The project aims to develop a geospatial-based system that continuously monitors and visualizes the air quality mainly in urban areas. The system is used to map air pollutant distribution, identifying pollution hotspots and potential causes of pollutants.

Objectives:

- To gather and assimilate air quality data with spatial interpolation.
- To make a mapping of the air pollutants over the urban areas by using folium.
- To estimate the pollution levels in those areas of the city that are not covered by monitoring stations.
- To test the changes in air quality over time and relate the results to other factors such as traffic and population.
- To identify and bring attention to areas in which the quality of the air people breathe is hazardous to their health.

1.3 Research questions

Q1: Which areas in the urban environment are most impacted by air pollution?

Q2: How can air quality have changed over time in these urban areas?

Q3: Which spatial interpolation effectively highlights the zones with unsafe air quality standards or levels?

1.4 Research Rationale

The spatial distribution of air pollutants is important to analyze for better management and prevention measures in the interest of people's well-being and city planning. The identification of hot spots allows authorities to focus on health hazards and apply them to occurring pollution. This project makes use of Geospatial to improve the depth of the understanding of air quality in cities and the general way they can be transmitted. These space-joining techniques can supplement the coverage in the monitoring networks, and the degree of pollution. Further, understanding the pattern of air pollution with regards to traffic, population density etc. will help in identifying the causative factors of pollution and hence come up with solutions to reduce pollution.

1.5 Research Significance

The significance of this study is that it can increase the application of air quality monitoring and governance of cities. These two tools make the project more effective in defining detailed pollution irregularities through the integration of spatial data and air quality measurements. The works to be produced by this system help in generating maps and diagrams that are beneficial and indispensable for policymakers, urban planners, and public health workers, organizations so that they understand what to do for public health security (Jo et al. 2020). In addition, the structure of the project in terms of tracking trends and comparing them with possible sources of pollution is valuable for identifying goals for the long-term enhancement of the environment. The application of such Python tools as Geo pandas, Matplotlib, and Folium also proves the effectiveness of open-source solutions in handling environmental issues. Hence, the findings of this study are readily applicable to other areas and conditions.

2. Literature Review:

The literature review is crucial in knowing the existing knowledge, and preliminary findings in the field of air quality monitoring, and visualization. It examines various approaches, tools, and models that have been used to model and monitor the quality of the atmosphere, especially in the context of cities. Based on the comparison of the current and the continued developed methods for air quality assessment, the benefits and drawbacks of the standard methods, including fixed monitoring stations, and the perspective of using remote sensing methods are discussed. Moreover, it expands on the GIS connection to an air quality data source and focuses on the significance of spatial approaches in assessing pollution hotspots and the spatial pattern of air pollutants.

2.1 Air Quality Monitoring System

Air quality monitoring technique is a crucial aspect of health considering the quality of air that people breathe, and conventional techniques have always been of immense benefit in this area of endeavor. Traditional air quality monitoring networks are usually point sources which are established in both the urban and rural areas to ensure they are evenly distributed to cover the entire country, and they mainly measure pollutants such as PM2, PM10, NO2, and SO2.

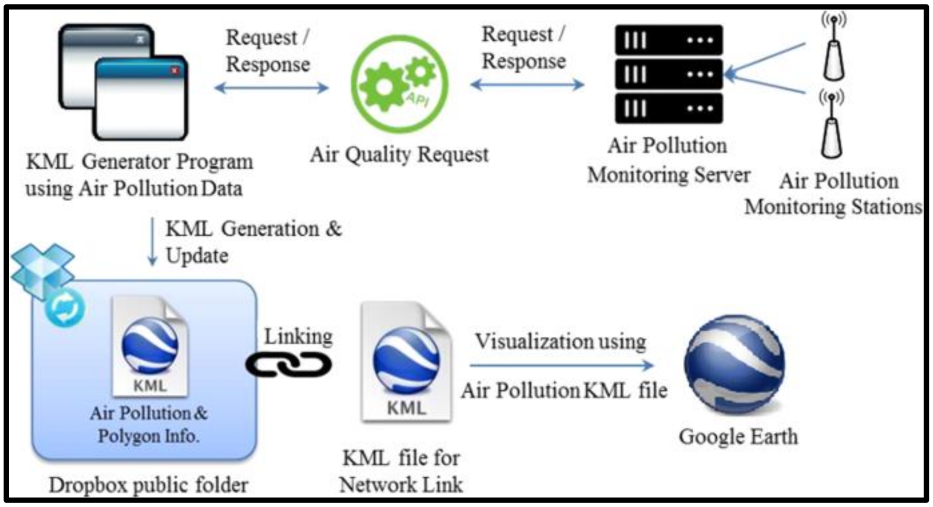

Request / Response
Request / Response
KML Generator Program using Air Pollution Data
Air Quality Request
Air Pollution Monitoring Server
Air Pollution Monitoring Stations
KML Generation & Update
KML
Air Pollution & Polygon Info.
Dropbox public folder
Linking
KML
KML file for Network Link
Visualization using Air Pollution KML file
Google Earth

Figure 2: Real-time monitoring of urban air quality monitoring

(Source: Chen, 2019)

These stations give real-time, accurate and dependable information and hence accurately measure the changes in air quality in any given period (Ren and Cao, 2020). Probably, the main asset of these networks is the stability and uniformity of measurements, primarily useful for compliance with various requirements and the continuation of long-term monitoring of environmental changes. But as much as these traditional methods are correct in their findings, they have a lot of drawbacks. This is because there is only limited spatial coverage, which prevents proper analysis of the differences between various zones or districts of the city.

Satellites and aerials/drones are more applicable to large areas and can give much better spatial coverage than the point-based sensors in the list above for monitoring air quality. Satellites collect information almost worldwide; this makes it easy to capture trends that would not be captured by stations on the ground. While a system refers to a specific means for monitoring an area, drones provide the ability to monitor certain parts of a region that may be hard to access or that have little to no monitoring systems in place. However, these methods are also associated with some limitations (Schürholz *et al.* 2020). Satellite sensors in most cases come with lower spatial resolution hence it is difficult to capture small-scale pollution events or to measure the quality of air in a localized area. However, range and practicality of function are major downfalls when it comes to drones, which although, provide much higher-resolution imagery. The two methods also employ complex processing of data to arrive at the best results to increase the efficiency of production. Thus, it is seen that even though remote sensing technologies support the existing monitoring networks in terms of spatial range, they do not offer the specificity and accuracy characteristic of fixed stations.

2.2 Geospatial Analysis in Air Quality Monitoring

Geospatial analysis plays a critical role in air quality monitoring by integrating geospatial with environmental data to assess and visualize pollution across different regions. This enables researchers to combine spatial data, such as topography, land use, and meteorological information, with air quality measurements to create detailed maps that highlight pollution hotspots and spatial patterns.

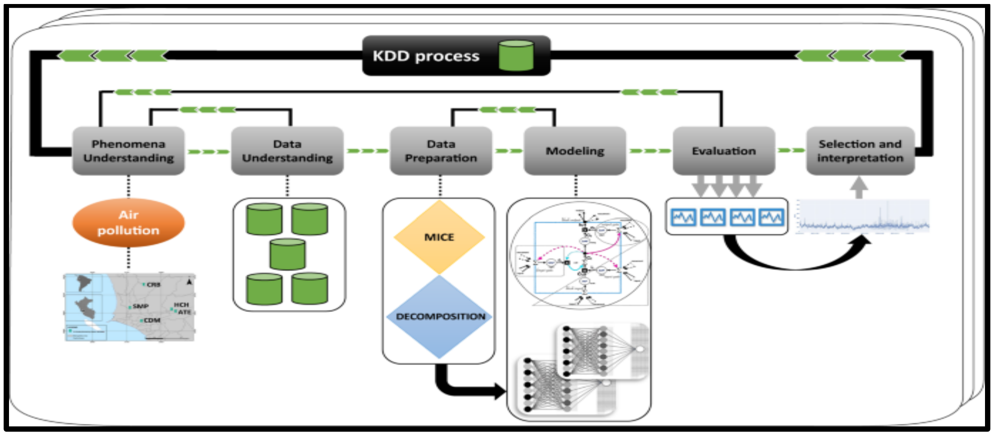


Figure 3: Air quality assessment using a geospatial system

(Source: Cordova et al. 2021)

But it does not consider both distance and spatial dependency of points with its data which makes Kriging more precise in its estimates. Kriging also comes with the added benefit of being able to give the corresponding standard errors of the estimates, which is important in environmental research (Ryu, 2022).

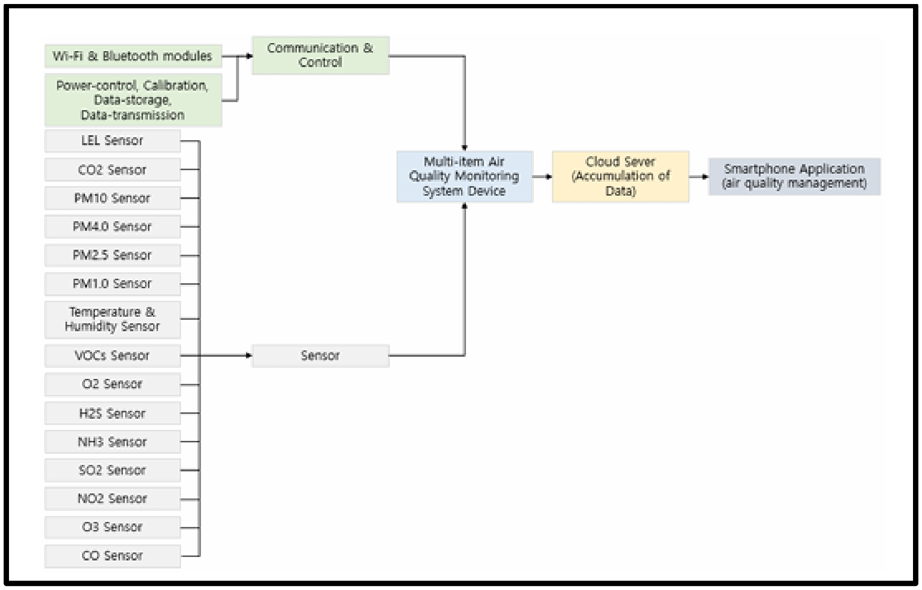

Wi-Fi & Bluetooth modules
Power-control, Calibration, Data-storage, Data-transmission
Communication & Control
LEL Sensor
CO2 Sensor
PM10 Sensor
PM4.0 Sensor
PM2.5 Sensor
PM1.0 Sensor
Temperature & Humidity Sensor
VOCs Sensor
O2 Sensor
H2S Sensor
NH3 Sensor
SO2 Sensor
NO2 Sensor
O3 Sensor
CO Sensor
Sensor
Multi-item Air Quality Monitoring System Device
Cloud Sever (Accumulation of Data)
Smartphone Application (air quality management)

Figure 4: Multi-item air quality-based monitoring system

(Source: Park et al. 2021)

These tools enable the stacking of data types that can facilitate the correlation of various air pollution levels to possible genitors/precipitates (Marques et al. 2020). With these relationships elaborated, therefore, Geospatial enables policy makers and urban planners to make informed decisions in combating air pollution to enhance people’s health. Spatial interpolation techniques are critical in air quality mapping because measurements in many places can be limited or nonexistent. Among methods that have been developed in estimating the pollution status of not monitored areas using data collected at monitored stations include inverse distance weighting (IDW), Kriging and SPLINE. Here, IDW supposes that the effect of a particular data point is inversely proportional to the distance and thus the method is characterized by great applicability for constructing smooth continuous surfaces of pollution levels.

2.3 Visualization Techniques for Air Quality Data

The provided dataset contains stratified and hour-based air quality data and factors that contribute to air quality. These columns are the time, and concentration of various gasses such as CO (carbon monoxide), NOx (nitrogen oxides), NO2, and benzene (C_6H_6**)**, among others, as well as sensor outputs, labeled PT08.S1 to PT08.S5, possibly corresponding to the mentioned gases or other conditions. Besides, it comprises other environmental parameters like temperature (T), relative humidity (RH), and absolute humidity (AH). In the context of the proposed study, Air Quality monitoring and visualization this dataset can be used for the analysis of temporal changes in the concentration of pollutants, identification of peak values and their relation to environmental parameters such as temperature and humidity (Park *et al.* 2021). Discrete methods could be used on this data to estimate pollution at the non-representative time intervals and geographical locations and works with maps can be created to show distribution and locate concentration areas throughout time. Applying the methods based on the given dataset, it is possible to identify the connections between air quality changing with time and environmental factors. This may include the changes due to traffic intensity or meteorological conditions.

2.4 Applications of Geo Pandas, matplotlib, and Folium

The operations on the available spatial data are performed with the help of Geo Pandas in this project and this makes it easier to join and analyze the air quality across different cities. The pollution level in these dummy cities can then be displayed with Matplotlib to create static plots of the variations of different parameters in air pollution including PM2. 5 and NO2 concentrations. For a live production of the maps, Folium is used in generating working maps to demonstrate the dispersion of pollutants across the urban setups (Marques et al. 2020). Among these methods, the Ordinary Least Squares (OLS), is used in establishing the relationship between the air quality index and other parameters such as the traffic intensity and industrial emissions. It makes it possible to establish other factors that are significantly affecting pollution of the air and this assists in comparing fluctuations in the quality of air in various cities.

2.5 Factors Contributing to Air Pollution

According to research, the growth of cities increases the concentration of vehicles as well as construction activities resulting in increased emissions of NO2 and PM. These factors are often employed in geospatial studies using data on population distribution, traffic volumes, and land usage coupled with air quality (Schürholz *et al.* 2020). This analysis enables a researcher to establish the areas with concentrated pollution which may include roadsides, and densely populated areas among others, and to estimate the effects of urbanization on the quality of air. These studies are carried out using the geospatial method to map the direct relationships between the occurrences of urban growth and pollutant concentration that would help urban planners and policymakers immensely.

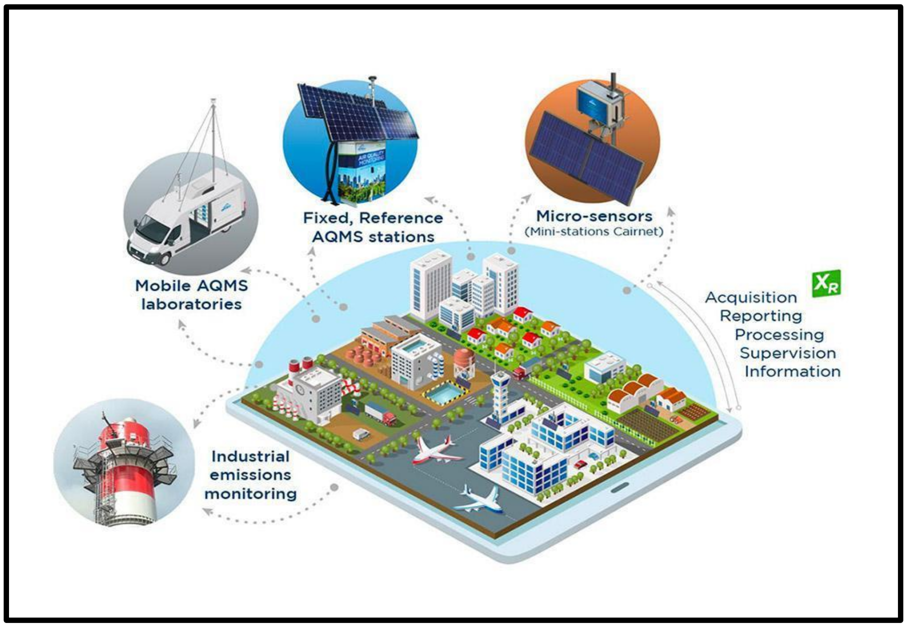


Figure 5: Factors contributing to air pollutants

(Source: Air Quality Monitoring, 2024)

Industrial activities are another major source of emission of air pollutants in the urban areas especially in cities that have established industries and processing firms. Mobile and fixed emission sources include industries like factories, power plants, and others releasing pollutants like SO2, VOCs, and heavy metals. Spatial modeling is used to filter possible effects of industrialization on pollution levels since industrial locations are compared with the position of residential areas and associated air quality results (Ren and Cao, 2020). These spatial approaches also contribute towards establishing major industrial offenders that cause pollution and its effects on the surrounding community.

The wind speed, temperature, and humidity play a critical role in proving the rate and the tendency of air pollution particles to spread and concentrate in certain areas. Wind can either blow pollutants far away from their sources or can blow them to some area and accumulate them there. Temperature inversion can also hold pollutants close to the earth's surface being felt more in urban areas (Narayana *et al.* 2022). Relative humidity also has an impact on the chemical composition of pollutants soluble in water and the formation of secondary pollutants: ozone, for example. In geospatial air quality models, these meteorological variables are assimilated and then used in the dispersion modeling to predict the movement and concentration of pollutants and hence, help in realizing regions with initial high pollution.

2.6 Health Impacts of Air Pollution

Researchers have established that air pollutants significantly affect human health through respiratory and cardiovascular diseases inclusive of their effects on airways. Extensive research published in scholarly journals indicates the negative effects of pollutants like particles (PM2. 5 and PM10), nitrogen dioxide (NO2), sulfur dioxide (SO2), and ozone (O3) on human health with consequences such as asthma, bronchitis, lung cancer, and heart diseases.

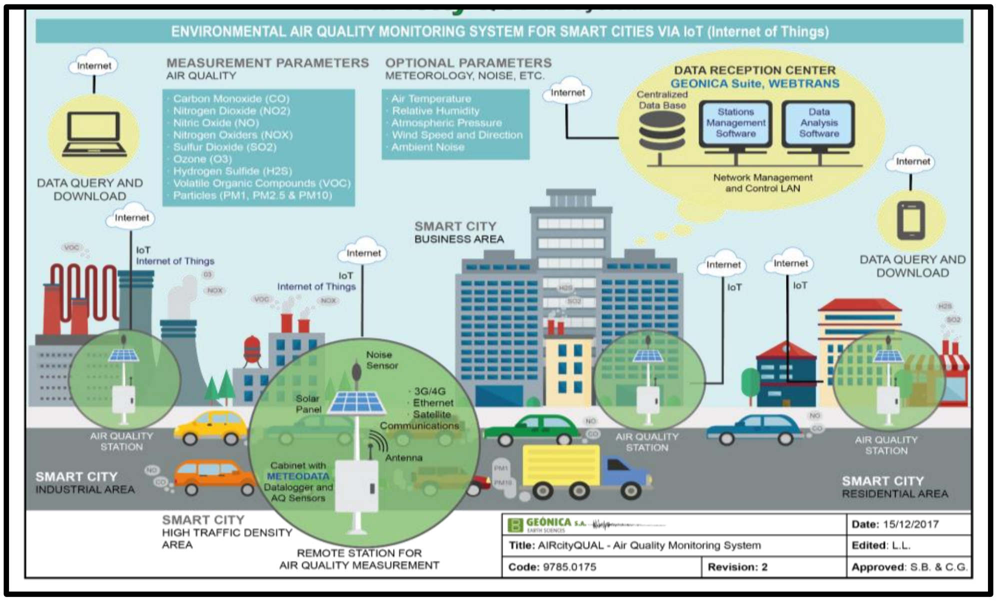


Figure 6: Air quality monitoring solutions

(Source: Geonica, 2020)

Literature that has adopted the spatial analysis method has been useful in the determination of the association between the rate of pollutants present in each environment and the effects to health. For instance, geospatial technologies have helped in spatial analysis of the emission of pollutants and other diseases such as asthma which is likely to appear in areas where NO2 levels are high. These analyses assist in determining not only which areas need the most cleaning out, but also which populations are most endangered, allowing for the formulation of interventions in public health and policies (Concas *et al*. 2021). Such studies stress the necessity to combine the geography of a region with health indicators to understand the effects of air pollution on inhabited environments and to work out effective preventative measures.

2.7 Vulnerable Populations

The analysis has presented a composition that air pollution is much more threatening to children, elders, and the poor. Due to the tender age of the children, they may suffer from respiratory-related illness that exists because of polluted air especially due to the developing respiratory system that is within the children. People of low-income levels are most likely to live in peri-industrialized areas or near commercial centers and highways which contribute to increased health risks from pollution. Spatial analysis has been important in finding out those groups that are most vulnerable to the effects of pollution by overlaying pollution data with demographic information (Concas et al. 2021). It extends beyond mere identification of environmental justice issues for fair policies to be adopted for the benefit of the needy persons.

With geospatial analysis, one can identify the areas where the vulnerable population is located about the sources of pollution and in this way, the social intervention’s priorities are correctly placed, among the most threatened groups of citizens. Urban air quality monitoring projects have increasingly employed geospatial tools and Python libraries to enhance data analysis and visualization. For instance, cities like London and New York have implemented extensive air quality monitoring networks that utilize Geospatial to map pollution levels across urban areas. These projects often employ Python libraries such as Geo Pandas for spatial data manipulation and Folium for interactive map creation, allowing for real-time visualization of air quality data (Tian *et al.* 2021).

3. Methodology:

3.1 Data, study area, and methods

3.1.1 Data

The data set used contains several variables of air quality determined on an hourly basis over forty-eight hours in different cities. These parameters are concentrations of carbon monoxide (CO), Non-Methane Hydrocarbons (NMHC), Benzene (C6H6), Nitrogen oxides (NOX), Nitrogen dioxide

(NO2), Ozone(O3), and other variables including temperature, relative humidity (RH), and absolute humidity (Liu *et al.* 2021). All these measurements are obtained from air quality monitoring stations in the various areas with the stations incorporating specific instruments for collecting real-time data.

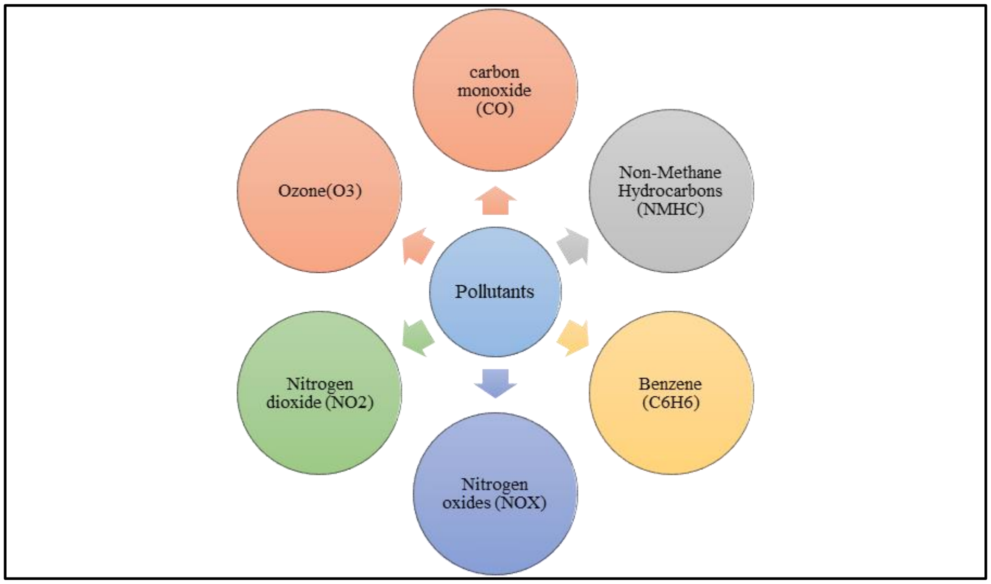


Figure 7: Common Pollutants

Based on enhancing this data, spatial coordinates, latitude and longitude of hypothetical monitoring stations in various cities like London, Manchester, Birmingham, Bristo, etc. can be accessed. Through the context of air quality monitoring and visualization, data can be gathered from numerous sources inclusive of public air quality measurement stations; additionally synthetic is going to be used to represent various cities. Based on spatial data manipulation, Geo pandas can be used with Matplotlib for static mapping of air quality in different urban regions. Folium can be used to produce maps on the dispersion of air pollutants with a view of pinpointing areas of high concentration and risky regions through the application of Interpolation (Han *et al.* 2021). Precisely, the data analysis can identify the relation between the level of air pollution and some factors like density of cars, or the quantity of people. Ordinary least squares (OLS) regression analysis can be used to identify trends of the above-mentioned variables about air quality and thereby predict the areas most affected and possible causes of pollution.

3.1.2 Study Area

A suitable area of observation encompasses cities as they are the centers of high air pollution detrimental to human health. Much of the study involves using air quality data to map pollutants' distribution. The respective cities are built utilizing geospatial technologies where various cities are represented with features of traffic, industrial, and population density. These are related to air quality indices to draw trends and areas of high pollution levels. Categorized spatial interpolation techniques are used to bring out regions referred to as "unsafe zones," that is, concentrations of the pollutants are above the recommended limit. This respective study makes it attainable to provide empirical data on the extent of spatial variability in air pollution in a city, which extends knowledge about the dispersion of such air pollutants as CO, NOx, and particulate matter (Fowler *et al.* 2020).

Tools such as Matplotlib for graphs and Folium for maps are used and through them, the study maps all the affected cities and areas of the most polluted regions. Based on understanding how the pollution levels might be explained by these factors, an OLS regression analysis approach. It is used to estimate the impact of pollution levels and potential explanatory variables like traffic intensity, population and industrial density. It contributes to getting insight into the specific factors causing differences in air quality and the foundation for measures to eliminate it. It also seeks to establish the changes in air quality over time to get a temporal dimension on how pollutants have changed in the urban areas probably due to changes in policies or the level of urbanization (Giovannini *et al.* 2020.). Integrating spatial and temporal data, the research should better explain air pollution patterns in the respective cities, which should help to improve the strategies of environmental and health management.

3.1.3 Methods

Implementing a geospatial Python project focused on air quality monitoring and visualization, the following methods can be employed:

Spatial Data Collection and Manipulation

The first procedure that needs to be followed is to collect information relating to the quality of air in different cities. These maps can incorporate spatial results from interpolation to indicate hotspots, regions with high pollution concentrations. Also, overlying other spatial variables like traffic flow or population density also assist in correlating pollution levels with possible causal factors.

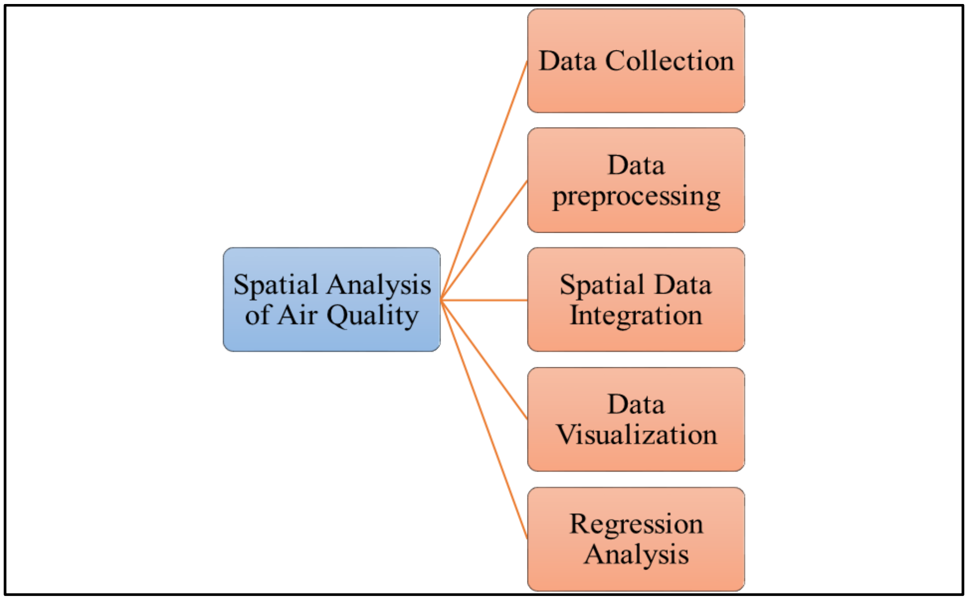


Figure 8: Spatial Data Model

The techniques such as Kriging are used then someone can be able to forecast the level of pollution in areas that were not tested and therefore create a GIS mapped surface model of the pollutant strength. It is very important for establishing relations in the space and for understanding the tendencies in emerging air quality issues in some parts of the city.

Visualization and Hotspot Identification

Matplotlib has been useful especially when it comes to the visualization of air quality data as well as Folium. When making static charts that display the trends of air pollutant concentration across Cities over time, one ought to use Matplotlib. When the activity requires a more significant level of students' engagement, it is advised to use Folium to create engaging maps that illustrate densities of pollutants (Stavroulas *et al.* 2020). This means that there should be better cognitive presentation of the result showing how insecure the zones are and for reasonable social presentation and formation of social policies.

Regression Analysis and Insight Generation

Through that regression analysis needs to be performed to identify the contributing factors to air pollution. This statistical method can help to express the link between air quality parameters including CO, NO2 and O3 and possible influencing factors which include traffic flow, industries, and population density (Biondo *et al.* 2022). Therefore, depending on the results of the regression coefficients, such factors influencing mostly the poor quality of air can be found.

3.3 Data analysis techniques

The structured approach to executing an air quality monitoring and visualization project using geospatial tools:

3.3.1 Data Pre-processing

Preparing the data for analysis is a very important step which is known as data pre-processing. This can be done before any form of analysis; begin by importing the dataset and cleaning and formatting the data if necessary. Since air quality data includes temporal and spatial localization, check that all the time stamps are consistent and formatted properly. Data can be cleaned using filtering where anomalies can be removed or use data handling techniques like the use of interpolated imputation (Ouni and Saleem, 2022). Also, the data density can be assessed for pollutants to the same level for comparability. If the work is conducted with spatial data, check the coordinates, and convert them to a single coordinate system using Geo pandas. This step entails the process of data screening to narrow down the geographical and temporal scopes.

3.3.2 Data Visualization

This process is critical to obtain insights and communicate the tendencies and changes of air quality. Firstly, it is necessary to plot simple time-series data with the help of Matplotlib to implement a further understanding of the variance of the pollutant level. Based on spatial visualization, it is best to use Folium to create interesting maps (Kumar and Pande, 2023). These maps should be used to plot concentrations of pollutants so that identification of ‘hot spots’ can easily be made. Based on improving such visualizations, it is recommended to use spatial interpolation methods (Kriging or IDW) to find out pollution rates in other regions not covered by monitoring. It helps define ‘unsafe zones’ more accurately for a given proximity to a source.

3.3.3 Spatial Analysis

Spatial analysis takes the data analysis to another level through the inclusion of the geographical factor. It is essential to start with Geo pandas to address the required actions on the spatial datasets.

The collective functionality can be assessed to obtain a spatial representation of pollutants combined with geographical areas, city districts or related areas. Spatial statistics can be employed to know if there is clustering in the levels of pollution and get the extent of spatial autocorrelation (Zhu *et al.* 2020). Also, the collective issue is to use spatial interpolation techniques to estimate pollution values in zones which are not monitored, pointing to potentially unsafe areas. The help of such analyses can construct one more map that reveals the current picture of polluted sites and can predict future tendencies that can be useful in urban planning and determining policies.

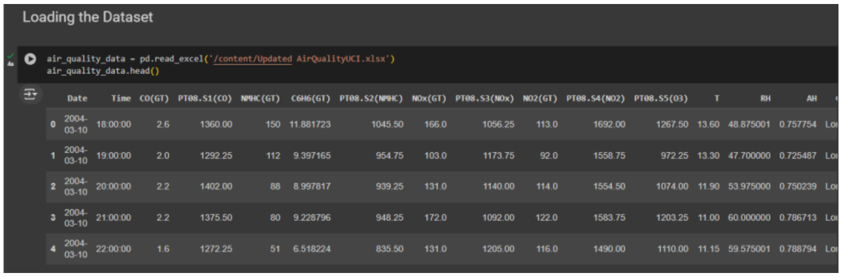
3.3.4 OLS Regression Analysis

Multiple linear regression can be employed to examine the relationships between the different factors that could influence air quality factors such as the volume of traffic, population density and industrial activities (Liang *et al.* 2020). The OLS regression can be performed using the Statsmodels library in Python which aids in the determination of the important predictors of air pollution. The regression analysis also provides the proportion of variance of pollutant levels that is contributed by these factors. Recognizing that, there are some assumptions used for OLS, for example, homoscedasticity and no multicollinearity, to verify your model.

4. Result and Analysis:

4.1 Implementation

The implementation process of several phases: data preprocessing step, spatial analysis and the regression analysis phase. Lacking latitudes and longitudes, these are created; further, a Geodata Frame is built for spatial analysis. Spatial weights are developed for spatial regression analysis and Ordinary Least Squares regression for examining pollutant measure dependability on its predictor variables. The Folium maps and the spatial distribution plots that accommodate the outcome are used. Every step guarantees the detailed consideration of spatial and temporal characteristics of air quality data.



Loading the Dataset

```
air_quality_data = pd.read_excel('/content/Updated AirQualityUCI.xlsx')
air_quality_data.head()
```

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.S5(O3)	T	RH	AH	
0	2004-03-10	18:00:00	2.6	1360.00	150	11.881723	1045.50	166.0	1056.25	113.0	1692.00	1267.50	13.60	48.875001	0.757754	Lo
1	2004-03-10	19:00:00	2.0	1292.25	112	9.397165	954.75	103.0	1173.75	92.0	1558.75	972.25	13.30	47.700000	0.725487	Lo
2	2004-03-10	20:00:00	2.2	1402.00	88	8.997817	939.25	131.0	1140.00	114.0	1554.50	1074.00	11.90	53.975000	0.750239	Lo
3	2004-03-10	21:00:00	2.2	1375.50	80	9.228796	948.25	172.0	1092.00	122.0	1583.75	1203.25	11.00	60.000000	0.786713	Lo
4	2004-03-10	22:00:00	1.6	1272.25	51	6.518224	835.50	131.0	1205.00	116.0	1490.00	1110.00	11.15	59.575001	0.788794	Lo

Figure 9: Loading the dataset

The figure shows the air quality dataset in a panda Data Frame but there are many syntax and formatting errors in the code. The range over a period consists of certain CO levels and different measures of gases. The problematic and abnormally placed punctuation characters as well as the wrong placement of parentheses interferes with the correct execution of code. To correct the above errors while loading the data the correct method to be used is ***pd. read_excel()***. The data in a clean format with correct mathematical headers from cells and overall correct syntax of the files is important when it comes to the analysis of the data set.

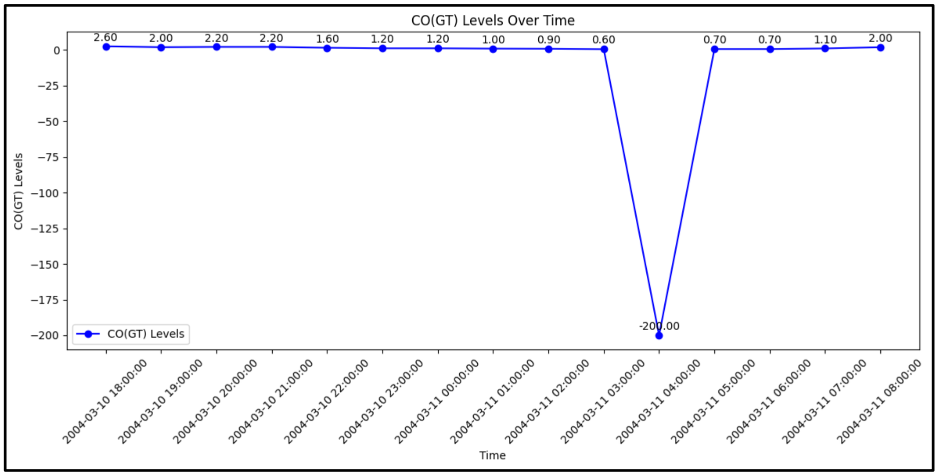


Figure 10: CO(GT) Levels Over Time

The above figure shows the date and time columns into one index, Datetime, that will allow chronological compatibility when analyzing air quality data. The plot is created with the purpose of presenting the first 15 data points only where the interest is in CO(GT) levels. The plot includes annotations to point out CO values at each time point. The x-axis formatting has been proposed for possible modifications to make these axes more easily understandable and herein the changes made would be to provide timestamps with an hourly rate.

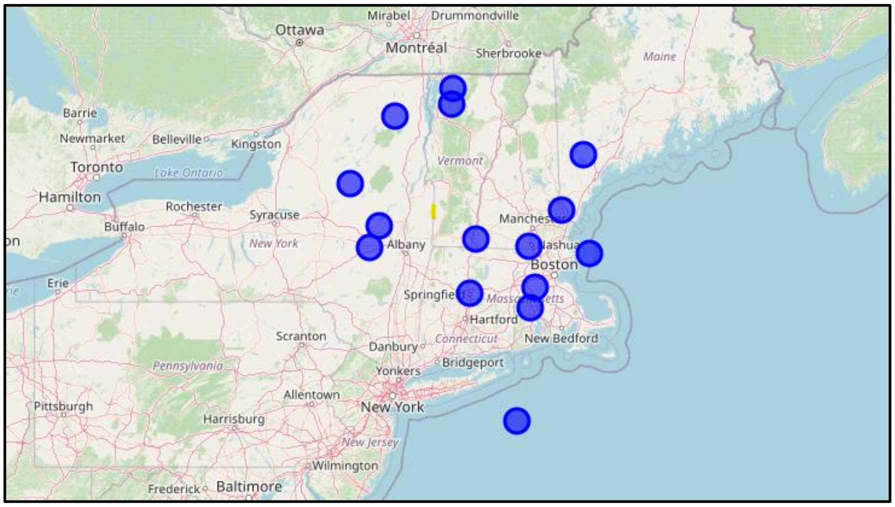


Figure 11: Interactive Map of Different Cities

The figure shows the air quality plot that checks the air quality of the cities by calculating latitude and longitudes. The map shows the level of air in different cities such as London, Birmingham, Manchester etc.

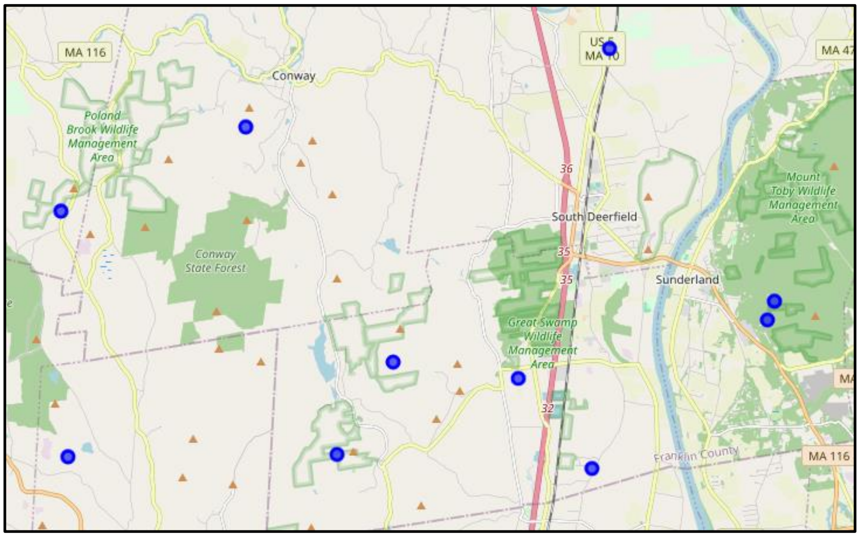
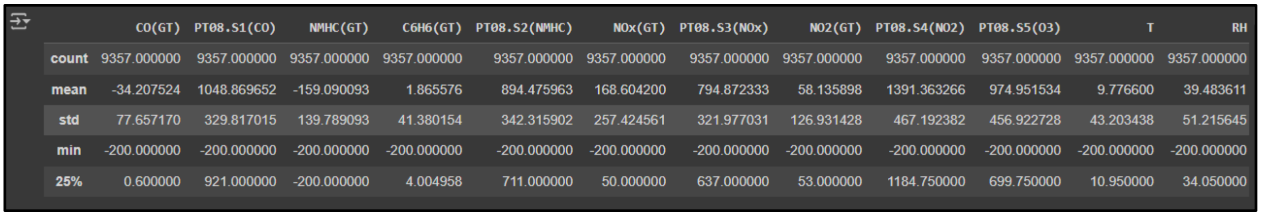

MA 116
Conway
US 5
MA 10
MA 47
Poland Brook Wildlife Management Area
36
Mount Toby Wildlife Management Area
South Deerfield
Conway State Forest
35
35
Sunderland
Great Swamp Wildlife Management Area
32
Franklin County
MA 116

Figure 12: Air Quality Map

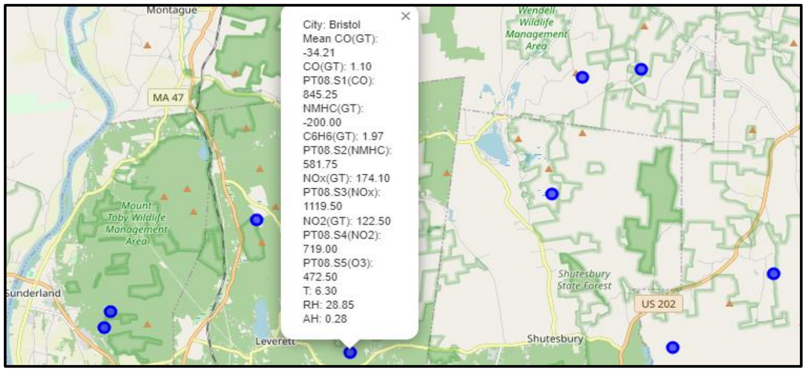
The map estimates the average of the latitude and longitude of all the air qualities to estimate the center of the geographical data. The average of the above-calculated co-ordinate is used for the positioning of the map at the start, should the map open with the data most dense region at its center.



	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.S5(O3)	T	RH
count	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000
mean	-34.207524	1048.869652	-159.090093	1.865576	894.475963	168.604200	794.872333	58.135898	1391.363266	974.951534	9.776600	39.483611
std	77.657170	329.817015	139.789093	41.380154	342.315902	257.424561	321.977031	126.931428	467.192382	456.922728	43.203438	51.215645
min	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000
25%	0.600000	921.000000	-200.000000	4.004958	711.000000	50.000000	637.000000	53.000000	1184.750000	699.750000	10.950000	34.050000

Figure 13: Descriptive Statistics

The summary statistics of the whole air quality dataset, and then refines it on certain specified variables of concerns, which mainly include pollution concentrations and weather conditions. The describe () function gives details such as the mean, standard deviation, and percentiles of the selected variables. The analysis to enlighten the mean, mode, median and range of the data distribution.



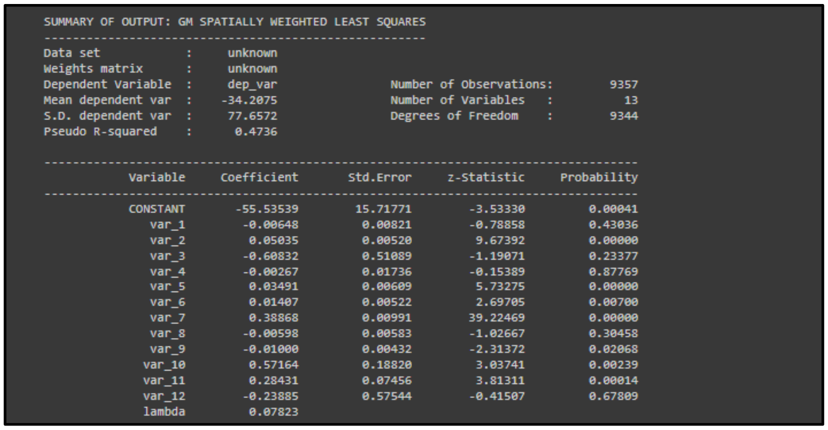
City: Bristol
Mean CO(GT): -34.21
CO(GT): 1.10
PT08.S1(CO): 845.25
NMHC(GT): -200.00
C6H6(GT): 1.97
PT08.S2(NMHC): 581.75
NOx(GT): 174.10
PT08.S3(NOx): 1119.50
NO2(GT): 122.50
PT08.S4(NO2): 719.00
PT08.S5(O3): 472.50
T: 6.30
RH: 28.85
AH: 0.28

Figure 14: Descriptive Statistics of cities

The code provides the descriptive statistics of the variables that include CO(GT), NMHC(GT), and NOx(GT), then creates a map of the data using Folium. Every point is a geographical location that when clicked, opens a pop-up with general information about the city and air quality indicators. The average mean CO(GT) scores in the set is equal to 2.33, individual places presenting CO(GT) ratios including 2.20 and 2.15.

4.2 Spatial Regression

Spatial regression models the spatial data by taking into consideration auto-correlation that is spatial dependence among various observations. It employs spatial weights matrices to express the connection between the variables while taking into consideration their spatial location. This method helps in increasing the precision of predictions especially when the outcome depends on spatial interaction and thus facilitates in the precise outcome of patterns and effects within geographical data.



```
SUMMARY OF OUTPUT: GM SPATIALLY WEIGHTED LEAST SQUARES
-----------------------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  :     dep_var              Number of Observations:        9357
Mean dependent var  :    -34.2075              Number of Variables   :          13
S.D. dependent var  :     77.6572              Degrees of Freedom    :        9344
Pseudo R-squared    :      0.4736

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT       -55.53539        15.71771        -3.53330         0.00041
               var_1        -0.00648         0.00821        -0.78858         0.43036
               var_2         0.05035         0.00520         9.67392         0.00000
               var_3        -0.60832         0.51089        -1.19071         0.23377
               var_4        -0.00267         0.01736        -0.15389         0.87769
               var_5         0.03491         0.00609         5.73275         0.00000
               var_6         0.01407         0.00522         2.69705         0.00700
               var_7         0.38868         0.00991        39.22469         0.00000
               var_8        -0.00598         0.00583        -1.02667         0.30458
               var_9        -0.01000         0.00432        -2.31372         0.02068
              var_10         0.57164         0.18820         3.03741         0.00239
              var_11         0.28431         0.07456         3.81311         0.00014
              var_12        -0.23885         0.57544        -0.41507         0.67809
              lambda         0.07823
```

Figure 15: Summary of Spatially

The figure shows the spatial error regression on the air quality data using the GM Error model to capture the result of the spatial data structures present in the data. Spatial weights are created based on geographical distance where a distance up to 0. 1, so that the changes of the outcome variables depending on the values of the explanatory variables in the regression are affected by the data points close to them. The outcomes of the regression model point out that there are potent predictors like var_2 (Coefficient = 0. 05035, p < 0.00001) and var_7 (Coefficient = 0. 38868, p < 0. 00001) having a positive relation with the dependent variable, that is CO(GT). It is details plotted on a Folium map with symbols embedded in red to identify the data points and the corresponding CO(GT) for ease of spatial analysis of air quality.

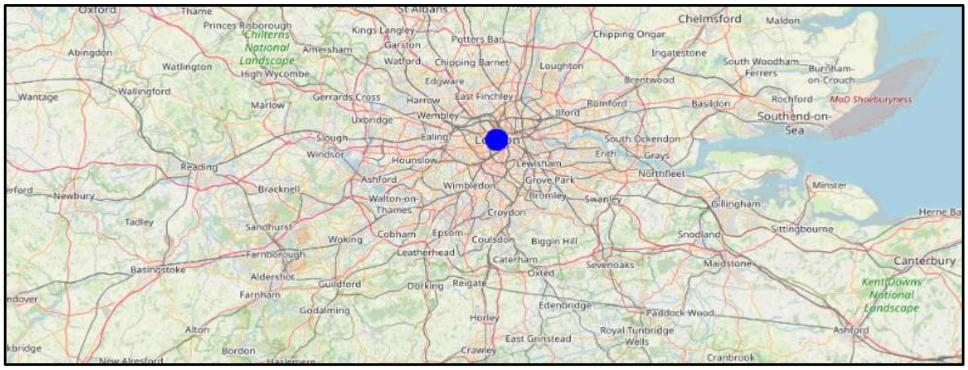


Figure 16: Spatially Air Quality Distribution

The Folium map created in the code displays points of interest using red circles to represent those air quality data points on the world map. The<|reserved_special_token_255|> map is an average map between latitudes and longitudes of the data set for orientation on the variation of CO(GT) in the spatial context. That is, each marker is clickable providing additional information regarding the CO(GT) value that was recorded at that point. Due to the use of the red markers, it is easier to see and locate the areas where actual CO(GT) levels have been recorded and this in turn makes it easier to analyze the air quality and look for areas of air pollution.

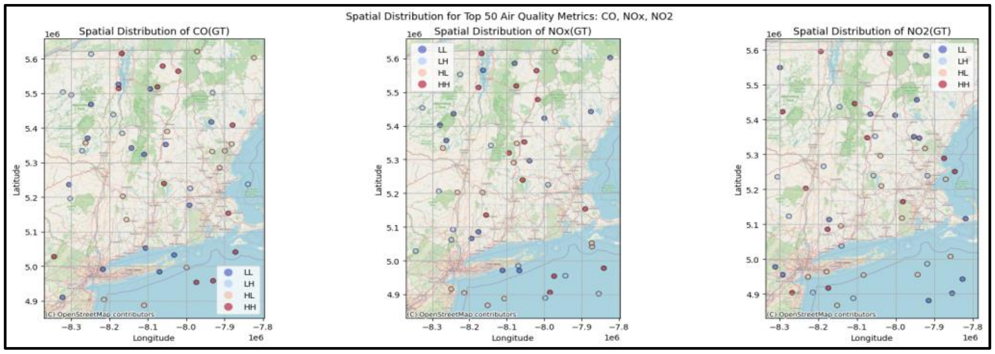

Spatial Distribution for Top 50 Air Quality Metrics: CO, NOx, NO2
Spatial Distribution of CO(GT)
Spatial Distribution of NOx(GT)
Spatial Distribution of NO2(GT)
1e6
Latitude
Longitude
LL
LH
HL
HH
5.6
5.5
5.4
5.3
5.2
5.1
5.0
4.9
−8.3
−8.2
−8.1
−8.0
−7.9
−7.8
(C) OpenStreetMap contributors

Figure 17: Spatial Distribution for Top 50 Air Quality Metrics: CO, NOx, NO2

The code involves developing a spatial analysis of air quality involving the highest fifty figures for fields such as CO(GT), NOx(GT), and NO2(GT) on a geographical map. It first converts the data to a Geodata Frame and reprojects it to a suitable coordinate reference system (EPSG:3857 for GPS mapping. The spatial distribution of each pollutant is represented on subplots of the figure with colors: ‘LL’, ‘LH’, ‘HL’, ‘HH’ amount to respective concentrations. The metrics applied to the analysis help to distinguish areas with high indexes, which means that there can be potential threats to the environment and, therefore, allow for more focused intervention and protection in these regions.

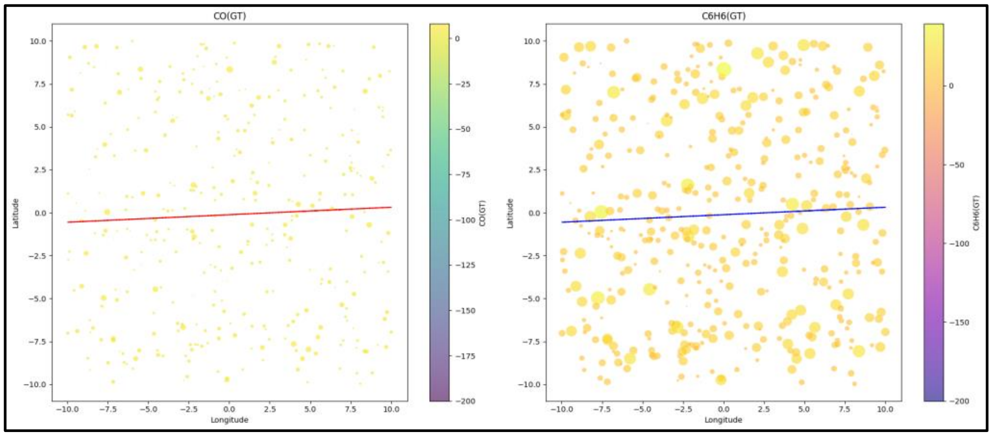


Figure 18: Geospatial Patterns of CO(GT) and C6H6(GT) in High Air Pollution Cities

The above figure shows the geospatial patterns CO(GT) and C6H6(GT) in High Air Pollution Cities. The process calculates the average CO(GT) of cities and determines the five cities with the highest average CO(GT). Scatter plots, which show the correlation between latitude, longitude, and air quality indexes (CO(GT) and C6H6(GT)). Pollutant data is analyzed using colored scatter plots and different regression lines where the trend in air quality is determined showing how the pollutant is distributed among the sampled cities.

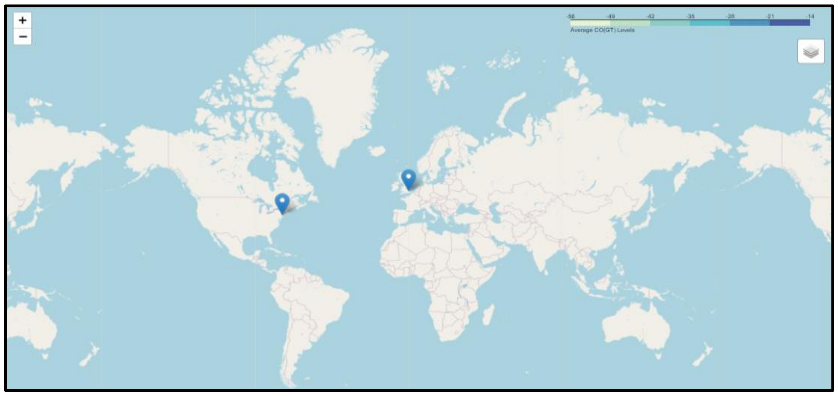
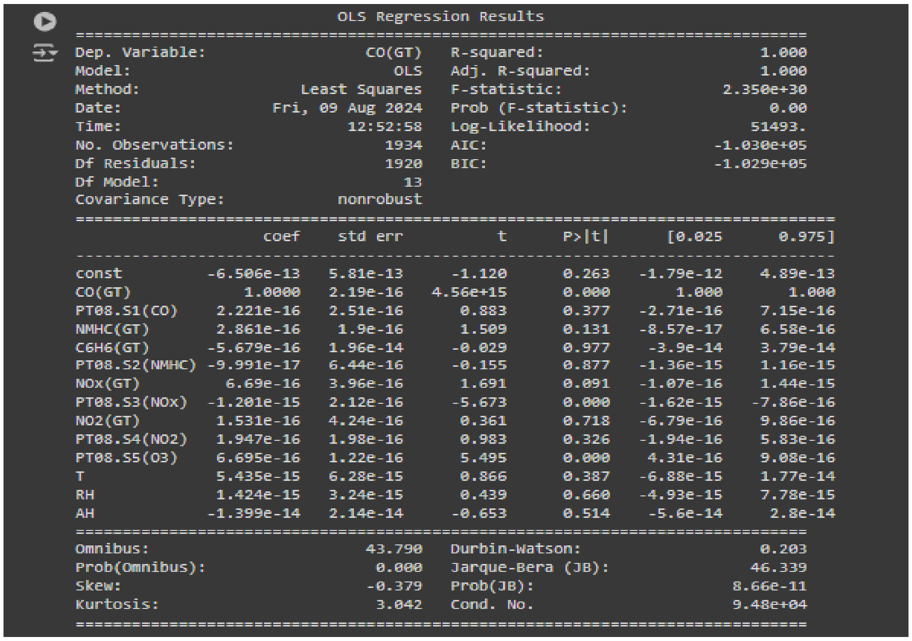


Figure 19: Choropleth Map of Average CO(GT)

The above figure shows the choropleth map of the average CO(GT) of GeoJSON for selected cities and plots air quality over a Folium map. The numeric air quality alarms are translated, and their arithmetical mean is calculated per city. Folium Choropleth layering of cities names the color based on the average CO(GT) levels to distinguish the geographical location of air pollution. This kind of visualization is useful in visualizing the distribution and density of the amounts of CO(GT) pollutants.

4.3 OLS Regression

Linear regression mainly Ordinary Least Squares (OLS) to analyze relations between a dependent variable and one or more predictors by summing residual squares. Output coefficients which describe the impact of each predictor against the outcome can be further tested statistically using hypothesis, and allow the presentation of goodness of fit measures such as R-squared and p-values.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              CO(GT)   R-squared:                          1.000
Model:                         OLS   Adj. R-squared:                     1.000
Method:              Least Squares   F-statistic:                    2.350e+30
Date:             Fri, 09 Aug 2024   Prob (F-statistic):                  0.00
Time:                     12:52:58   Log-Likelihood:                    51493.
No. Observations:             1934   AIC:                           -1.030e+05
Df Residuals:                 1920   BIC:                           -1.029e+05
Df Model:                       13
Covariance Type:         nonrobust
================================================================================
                    coef    std err          t      P>|t|      [0.025     0.975]
--------------------------------------------------------------------------------
const         -6.506e-13   5.81e-13     -1.120      0.263   -1.79e-12   4.89e-13
CO(GT)            1.0000   2.19e-16   4.56e+15      0.000       1.000      1.000
PT08.S1(CO)    2.221e-16   2.51e-16      0.883      0.377   -2.71e-16   7.15e-16
NMHC(GT)       2.861e-16    1.9e-16      1.509      0.131   -8.57e-17   6.58e-16
C6H6(GT)      -5.679e-16   1.96e-14     -0.029      0.977    -3.9e-14   3.79e-14
PT08.S2(NMHC) -9.991e-17   6.44e-16     -0.155      0.877   -1.36e-15   1.16e-15
NOx(GT)         6.69e-16   3.96e-16      1.691      0.091   -1.07e-16   1.44e-15
PT08.S3(NOx)  -1.201e-15   2.12e-16     -5.673      0.000   -1.62e-15  -7.86e-16
NO2(GT)        1.531e-16   4.24e-16      0.361      0.718   -6.79e-16   9.86e-16
PT08.S4(NO2)   1.947e-16   1.98e-16      0.983      0.326   -1.94e-16   5.83e-16
PT08.S5(O3)    6.695e-16   1.22e-16      5.495      0.000    4.31e-16   9.08e-16
T              5.435e-15   6.28e-15      0.866      0.387   -6.88e-15   1.77e-14
RH             1.424e-15   3.24e-15      0.439      0.660   -4.93e-15   7.78e-15
AH            -1.399e-14   2.14e-14     -0.653      0.514    -5.6e-14    2.8e-14
==============================================================================
Omnibus:                    43.790   Durbin-Watson:                      0.203
Prob(Omnibus):               0.000   Jarque-Bera (JB):                  46.339
Skew:                       -0.379   Prob(JB):                        8.66e-11
Kurtosis:                    3.042   Cond. No.                        9.48e+04
==============================================================================
```

Figure 20: OLS Regression

The perfect model fit evidenced by the R- Square of 1.000 shows that the independent variables in the model caused all the variation of the dependent variable, CO(GT). The high R-squared value, along with a high condition number of 9.48e+04 indicates possible multicollinearity issues of the independent variables, resulting in numerical problems. Some of the coefficients that have been posted for the above factors include PT08. S5(O3) and PT08. S3(NOx), are large but RH and AH are negligible quantities. Applying the Durbin-Watson statistic test to the given data, the Durbin-Watson statistic is 0.

4.4 Discussion

The given dataset contains the record of daily air quality indicated by different pollutants like CO, NMHC, NOx, NO2, and some weather parameters for a certain period. Most importantly, the subsequent things can be observed in terms of analyzing this data. Analyzing with Python’s data cleaning tool (Pandas), *Matplotlib,* and interactive mapping tool *Folium* can easily demonstrate and analyze air quality within various cities. The obtained data, which has some problems with formatting, shows a variability of CO(GT) values, and their decrease at night, especially in the period from one morning to another morning. Maps with random coordinates for dynamic figures enable the visualization of air quality parameters and the determination of areas with elevated risks and unsafe territories. Through measures of yield analysis that give quantitative information concerning the average and dispersion of pollution levels.

Spatial interpolation models like GM_Error contain information about the effects of neighboring points on the CO(GT) concentrations. The information gathered is plotted on the maps, and locations, which have higher values of pollution, are shown with red dots for CO levels. Hence, the estimated OLS regression analysis sums up a perfect fitness with the R-squared of 1.000, implying that all the independent variables incorporated in the model elaborate on the difference in CO(GT). However, based on a high multicollinearity value and a low Durbin-Watson statistic, there is a violation of certain assumptions of the model. Entirely, the analysis gives an effective overview of air quality changes, helps identify places where pollution levels are urgent and risky, and ensures the basis for targeted environmental management.

5. Conclusion and Recommendation:

5.1 Conclusion

This research incorporated the use of geospatial technologies in combination with Python libraries to track as well as observe the quantity of air pollution in large cities such as London, Manchester, Birmingham and Bristol. The linking of air pollution data with spatial data offered a good insight into pollution concentration and localized areas of severe air pollution. Geo Pandas tool for spatial data operation, Matplotlib for generation of static diagrams, and Folium for map interactions let efficient pollutant concentration as well as vehicle density and population relation visuals. All the sampling sites experienced pollution, but the extent varied significantly in space; thus, IDW estimated the pollution levels in other unspecified areas. This analysis not only helps to identify the causes of pollution in different areas but also to analyze the effectiveness of prevention measures and forecast future trends. The results can be presented qualitatively in the form of graphs using Matplotlib hence enabling the user to understand all the factors leading to poor quality of air in urban areas. Lessons learned from these projects highlight the importance of integrating technological tools with policy and community engagement to effectively address urban air quality challenges.

5.2 Recommendation

The recommendations from these implementations include more informed urban planning decisions, mainly targeted pollution control measures, and improved public awareness of air quality issues.

- The findings highlight the necessity of developing specific interventions for areas with high air pollution and prove the effectiveness of geospatial tools for better interpretation and controlling air quality.
- This study has clearly shown the capabilities of geospatial technologies and Python libraries in air quality and monitoring. By integrating the data on pollutant distribution, the plans showed the critical hotspots and offered clear suggestions as to the specifics of pollution in the given area.
- Comparing traffic data, the statistics of populations and air quality demonstrated considerable relations implying the impact of these factors on air pollution. Interactive maps and static charts helped in identifying other air quality problems that are prevalent in various cities.
- Accompanying these results is the need to include spatial analysis of the environment in the management of air quality issues adequately. Increasing the number of monitoring stations for air quality in these areas could give increased density to improve the obtained values.

References :

1. Air Quality Monitoring (2024), *ENVEA*, Available at: https://www.envea.global/solutions/ambient-monitoring/air-quality-monitoring/ [Accessed: 09 August 2024]

2. Biondo, E.J., 2022. *Real-time indoor air quality (IAQ) monitoring system for smart buildings* (Master's thesis, Instituto Politecnico de Braganca (Portugal)).

3. Chen, P. (2019). Visualization of real-time monitoring datagraphic of urban environmental quality. *EURASIP Journal on Image and Video Processing*, *2019*(1), 1-9. https://doi.org/10.1186/s13640-019-0443-6.

4. Concas, F., Mineraud, J., Lagerspetz, E., Varjonen, S., Liu, X., Puolamäki, K., Nurmi, P. and Tarkoma, S., 2021. Low-cost outdoor air quality monitoring and sensor calibration: A survey and critical analysis. *ACM Transactions on Sensor Networks (TOSN)*, *17*(2), pp.1-44.

5. Cordova, C.H. *et al*. (2021) *Air Quality Assessment and pollution forecasting using artificial neural networks in metropolitan Lima-peru, Nature News*. Available at: https://www.nature.com/articles/s41598-021-03650-9 (Accessed: 09 August 2024).

6. Demanega, I., Mujan, I., Singer, B.C., Anđelković, A.S., Babich, F. and Licina, D., 2021. Performance assessment of low-cost environmental monitors and single sensors under variable indoor air quality and thermal conditions. *Building and Environment*, *187,* p.107415.

7. Fowler, D., Brimblecombe, P., Burrows, J., Heal, M.R., Grennfelt, P., Stevenson, D.S., Jowett, A., Nemitz, E., Coyle, M., Liu, X. and Chang, Y., 2020. A chronology of global air quality. *Philosophical Transactions of the Royal Society A*, *378*(2183), p.20190314.

8. Geonica, 2020 *GEOcityQUAL system for air quality monitoring*, *Air Quality Monitoring Solutions*. Available at: https://www.geonica.com/en/air-quality-monitoring.php (Accessed: 09 August 2024).

9. Giovannini, L., Ferrero, E., Karl, T., Rotach, M.W., Staquet, C., Trini Castelli, S. and Zardi, D., 2020. Atmospheric pollutant dispersion over complex terrain: Challenges and needs for improving air quality measurements and modeling. *Atmosphere*, *11*(6), p.646.

10. Han, J., Liu, H., Zhu, H., Xiong, H. and Dou, D., 2021, May. Joint air quality and weather prediction based on multi-adversarial spatiotemporal networks. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 5, pp. 4081-4089).

11. Jo, J., Jo, B., Kim, J., Kim, S. and Han, W., 2020. Development of an IoT-based indoor air quality monitoring platform. *Journal of Sensors*, *2020*(1), p.8749764.

12. Kumar, D.N. and Priyanka, K., 2021. An extensive assessment of ambient air quality in city Delhi: Air quality monitoring, source apportionment and Analysis. *Journal of Indian Association for Environmental Management (JIAEM)*, *41*(1), pp.54-67.

13. Kumar, K. and Pande, B.P., 2023. Air pollution prediction with machine learning: a case study of Indian cities. *International Journal of Environmental Science and Technology*, *20*(5), pp.5333-5348.

14. Liang, Y.C., Maimury, Y., Chen, A.H.L. and Juarez, J.R.C., 2020. Machine learning-based prediction of air quality. *applied sciences*, *10*(24), p.9151.

15. Liu, Z., Gu, X., Dong, Q., Tu, S. and Li, S., 2021. 3D visualization of airport pavement quality based on BIM and WebGL integration. *Journal of Transportation Engineering, Part B: Pavements*, *147*(3), p.04021024.

16. Marques, G., Saini, J., Dutta, M., Singh, P.K. and Hong, W.C., 2020. Indoor air quality monitoring systems for enhanced living environments: A review toward sustainable smart cities. *Sustainability*, *12*(10), p.4024.

17. Narayana, M.V., Jalihal, D. and Nagendra, S.S., 2022. Establishing a sustainable low-cost air quality monitoring setup: A survey of the state-of-the-art. *Sensors*, *22*(1), p.394.

18. Ouni, R. and Saleem, K., 2022. Framework for sustainable wireless sensor network based environmental monitoring. *Sustainability*, *14*(14), p.8356.

19. Park, B., Kim, S., Park, S., Kim, M., Kim, T.Y. and Park, H., 2021. Development of multi-item air quality monitoring system based on real-time data. *Applied Sciences*, *11*(20), p.9747.

20. Ren, C. and Cao, S.J., 2020. Implementation and visualization of artificial intelligent ventilation control system using fast prediction models and limited monitoring data. *Sustainable Cities and Society*, *52*, p.101860.

21. Ryu, J.H., 2022. Prototyping a low-cost open-source autonomous unmanned surface vehicle for real-time water quality monitoring and visualization. *HardwareX*, *12*, p.e00369.

22. Saini, J., Dutta, M. and Marques, G., 2020. A comprehensive review on indoor air quality monitoring systems for enhanced public health. *Sustainable environment research*, *30*, pp.1-12.

23. Saini, J., Dutta, M. and Marques, G., 2020. Indoor air quality monitoring systems based on internet of things: A systematic review. *International journal of environmental research and public health*, *17*(14), p.4942.

24. Schürholz, D., Kubler, S. and Zaslavsky, A., 2020. Artificial intelligence-enabled context-aware air quality prediction for smart cities. *Journal of Cleaner Production*, *271*, p.121941.

25. Stavroulas, I., Grivas, G., Michalopoulos, P., Liakakou, E., Bougiatioti, A., Kalkavouras, P., Fameli, K.M., Hatzianastassiou, N., Mihalopoulos, N. and Gerasopoulos, E., 2020. Field evaluation of low-cost PM sensors (Purple Air PA-II) under variable urban air quality conditions, in Greece. *Atmosphere*, *11*(9), p.926.

26. Tian, D., Li, G., Cheng, S., Kong, L., Tang, X., Zhao, Q., Gao, Y., Shan, G. and Chi, X., 2021. Visual analysis system for fine-grained inline relationship of air quality data. *Journal of Computer-Aided Design & Computer Graphics*, *33*(9), pp.1326-1336.

27. Zangari, S., Hill, D.T., Charette, A.T. and Mirowsky, J.E., 2020. Air quality changes in New York City during the COVID-19 pandemic. *Science of the Total Environment*, *742*, p.140496.

28. Zhang, D. and Woo, S.S., 2020. Real time localized air quality monitoring and prediction through mobile and fixed IoT sensing network. *IEEE Access*, *8*, pp.89584-89594.

29. Zhang, H., Srinivasan, R. and Ganesan, V., 2021. Low cost, multi-pollutant sensing system using raspberry pi for indoor air quality monitoring. *Sustainability*, *13*(1), p.370.

30. Zhu, H.C., Yu, C.W. and Cao, S.J., 2020. Ventilation online monitoring and control system from the perspective of technology application. *Indoor and Built Environment*, *29*(4), pp.587-602.