IBM DIGITAL CREDENTIAL

FORECASTING AIR POLLUTION PARTICULATE MATTER (PM2.5) IN THE CITY OF LONDON

MIREL MORA

CERTIFICATE APPLIED DATA SCIENCE CAPSTONE

APRIL 27, 2021

Abstract

Air pollution in London is a mixture of emission created locally, and those from background concentrations, particles of pollution can remain suspended for weeks and so can be transported long distances. This has a negative effect on a number of different aspects of human health. In London, 9,400 premature deaths are attributed to poor air quality and a cost of between £1.4 and £3.7 billion a year to the health service[1]. Air pollution also contributed to other environmental issues, such as global warming, acid rain and ozone depletion. However, human well is the primary importance to study about air pollution. Then, Particulate matter (these are usually split into 2 sizes: PM2.5 and PM10) and NO_2. These are usually seen as dangerous form of air pollution due to their high concentrations and the negative health impacts they create. In this report predictive model for forecasting particulate matter concentration in London

[1] London Councils - https://www.londoncouncils.gov.uk/

Contents

1 — Introduction

1.1 Air pollution in London

It is often assumed that air pollution in London is a recent phenomenon, however, legislation attempting to control air pollution was enacted as early as 1306. Coal smoke and its associated problems remained a matter of concern in London up until the late 20th century with the famous smogs of the 1950s and 60s.

In recent years, the pollutants in the capital's air have altered considerably. This is primarily because of the decline in the use of coal in industry and domestic heating, which has led to large reductions in the emissions of sulphur dioxide and particles of soot over the past 40 years. At the same time the increased number of motor vehicles is producing considerable amounts of nitrogen dioxide and small particles.

1.2 Problem statement

There is mounting evidence of health effects from everyday exposure to air pollution. The modern-day small particles are our main problem for air pollution health effects, whilst other pollutants such as nitrogen dioxide and ozone are also a major cause for concern.

There are different effects depending on the length and intensity of exposure. For example, short term exposure (a few hours) to high levels of NO_2 can irritate the airways and cause severe coughing and exacerbate existing respiratory illnesses, which is uncomfortable at best, and dangerous at worst for vulnerable people (sick and older or younger people for example).

Long term exposure can contribute to someone developing a number of illnesses, such as asthma, pulmonary disease and lung cancer. It has also been shown to stunt the growth of children's lungs. This is particularly worrying, as around one-third of London's schools have been found to be close to busy roads that suffer illegal levels of NO2 pollution.

1.2.1 Important pollutant

The pollutants most widely referred to in the literature are:

- Particulate matter (these are usually split into 2 sizes: PM2.5 and PM10)
- Nitrogen dioxide (NO_2)
- Sulphur dioxide (SO_2)
- Ozone (O_3)
- And occasionally, Carbon Monoxide (CO)

PM and NO2 are commonly seen as the most dangerous forms of air pollution due to their high concentrations and the negative health impacts they create.

Particulate Matter 2.5 and 10 (PM)

PM is made up of a wide range of materials and arises from a variety of sources. Concentrations of PM comprise primary particles emitted directly into the atmosphere from combustion sources and secondary particles formed by chemical reactions in the air. It consists of a complex mixture of solid and liquid particles of human-made (such as diesel soot) and natural substances suspended in the air (such as sea spray and Saharan dust). In the UK the biggest human-made sources are stationary fuel combustion (power generators) and transport.

NO_2

Nitrogen dioxide is human made, with the major sources of emissions of NO_2 being combustion processes (heating, power generation, and engines in vehicles and ships)1. Nitrogen is released during the combustion of fuel and then combines with oxygen atoms to create nitric oxide (NO). This further combines with oxygen to create nitrogen dioxide (NO2). Nitric oxide is not considered to be hazardous to health at typical ambient concentrations, but nitrogen dioxide is. Nitrogen dioxide and nitric oxide are referred to together as oxides of nitrogen (NO_x). NO_x gases can also react to form smog and contribute to acid rain. NO_x is also central to the formation of fine particles or particulate matter (PM) and ground level ozone (O_3), both of which are associated with adverse health effects.

1.3 Air pollution forecast

Air pollution forecasting is a worthwhile investment on multiple levels - individual, community, national and global. Accurate forecasting helps people plan ahead, decreasing the effects on health and the costs associated.

If people are aware of variations in the quality of the air they breathe, the effect of pollutants on health as well as concentrations likely to cause adverse effects and actions to curtail pollution. Furthermore, there is a greater likelihood of motivating changes in both individual behaviour and public policy [1], as people want air quality information.

[1] Kelly F. (2012). Monitoring air pollution: Use of early warning systems for public health./

2 — Data description

2.1 Data set description

London is the capital and largest city of United Kingdom that is situated between the latitude of 52.3555° North and 1.1743° West. The data has been collected from London Air via the Openair project tools.

Openair is an R package developed for the purpose of analysing air quality data — or more generally atmospheric composition data. The package is extensively used in academia, the public and private sectors. The project was initially funded by the UK Natural Environment Research Council (NERC), with additional funds from Defra. The most up to date information on openair can be found in the package itself and at the book website here.

Openair tool provide a extensive access to UK air quality data. The networks includes:

- For importing data from the UK national network called Automatic Urban and Rural Network. This is the main UK network. (importAURN)]
- For accessing data from Air Quality Scotland network. (importSAQN)
- For accessing data from the Air Quality Wales network. (importWAQN)
- For accessing data from the Air Quality England network of sites. (importAQE)
- For accessing data from the Northern Ireland network of sites. (importNI)
- A simplified version of a function to give basic access to hourly European data based on Stuart Grange's package see.
- For accessing data from the sites operated by King's College London, primarily including The London Air Quality Network. (importKCL)

In this case, the networks used were UK national, Air Quality England and The London Air Quality.

2.2 Data cleaning

Data downloaded from three different network from Openair are combined into two data frames. The process is divide in the follow steps:

1. Extract data from Openair with R (Github - Jupyter)
2. Create a general data set with three networks (UK national, England and The London Air Quality).
3. Choosing stations (Stations in Greater London).

4. Filling missing values using Spatial interpolation.

5. Filling missing values using Panda method.

2.2.1 Extract data from Openair with R

Openair Tool works with programming language R. Openair has documents where explains how it works the package. It has been collected data from all stations in the United Kingdom for the years 2019 and 2020. It is generated two data set of 1495 stations around the United Kingdom. Data sets are divide into two groups that are particular and chronological data.

a) Particular data in the United Kingdom, England and London

b) Chronological data in the United Kingdom, England and London

2.2.2 Create a general data set with three networks

Generated one data frame with Particular data in the United Kingdom, England and London and another data frame with Chronological data in the United Kingdom, England and London using Pandas Library. Chronological data refers to that information, which quickly alters in a brief length of time. Particular parameters are values that change in area and time which cause changes within the concentration of air pollution 2.1.

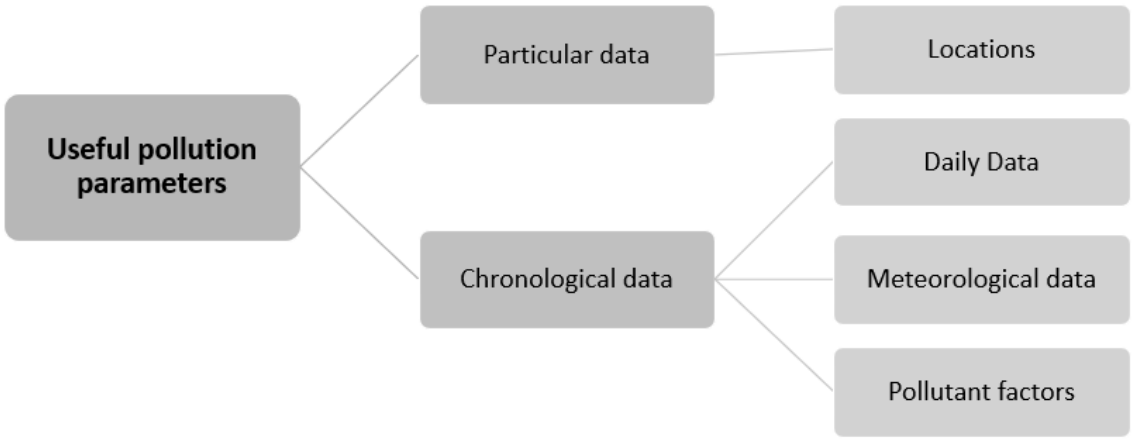
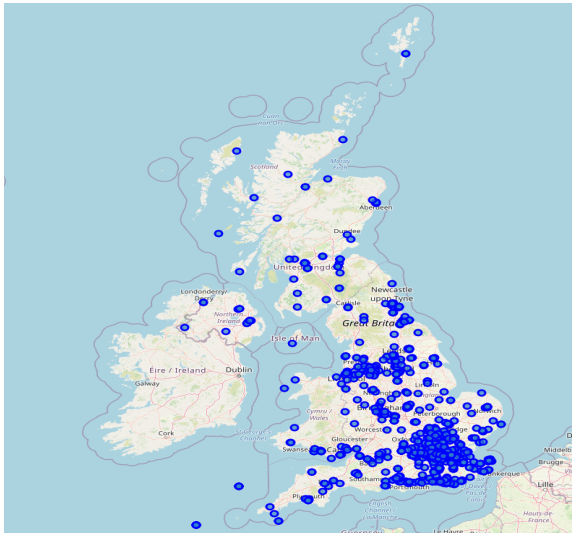


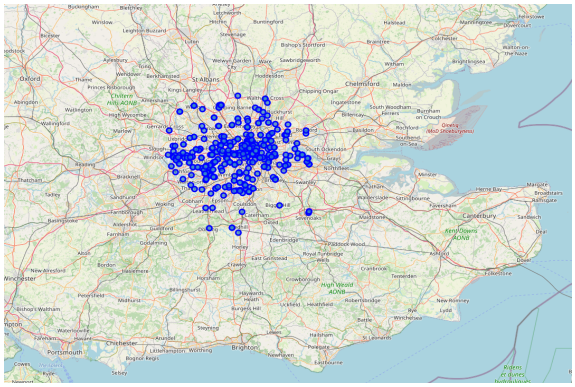
Figure 2.1: Data frames

2.2.3 Choosing stations (Stations in Greater London)

Two general data frame consider around 1495 stations. However, the project only is considering stations in London. So, the latitude between 51.20 to 51.20 and longitude from -0.55 and 0.20 is limited in the study and the rest of the stations are discarded.



(a) UK stations.



(b) Greater London stations .

Figure 2.2: Maps

Chronological data frame has different values like as site, code of station, date (date and hour), Oxides of Nitrogen (NO_x), Nitrogen Dioxide(NO_2), Nitric Oxide(NO), Ozone(O_3), Particulate Matter(PM_{10}) of 10 microns or less, Particulate Matter($PM_{2.5}$) of 2.5 microns or less, Volatile Component(v_{10}) of 10 microns or less, Volatile Component($v_{2.5}$) of 2.5 microns or less, Non-Volatile Component(nv_{10}) of 10 microns or less, Non-Volatile Component($nv_{2.5}$) of 2.5 microns or less, wind speed (ws), wind direction (wd), air temperature, Sulfur dioxide (SO_2), Carbon Monoxide (CO).

The stations should need to have more than 25% of total Particulate Matter($PM_{2.5}$) values and also a strong relationship with the Particulate Matter($PM_{2.5}$). So, it deleted stations that do not fulfil with features. Finally, the new data frame has 34 stations with the following values: code of station, date (date and hour), Oxides of Nitrogen (NO_x), Nitrogen Dioxide(NO_2), Nitric Oxide(NO), Particulate Matter(PM_{10}), wind speed (ws), wind direction (wd), air temperature.

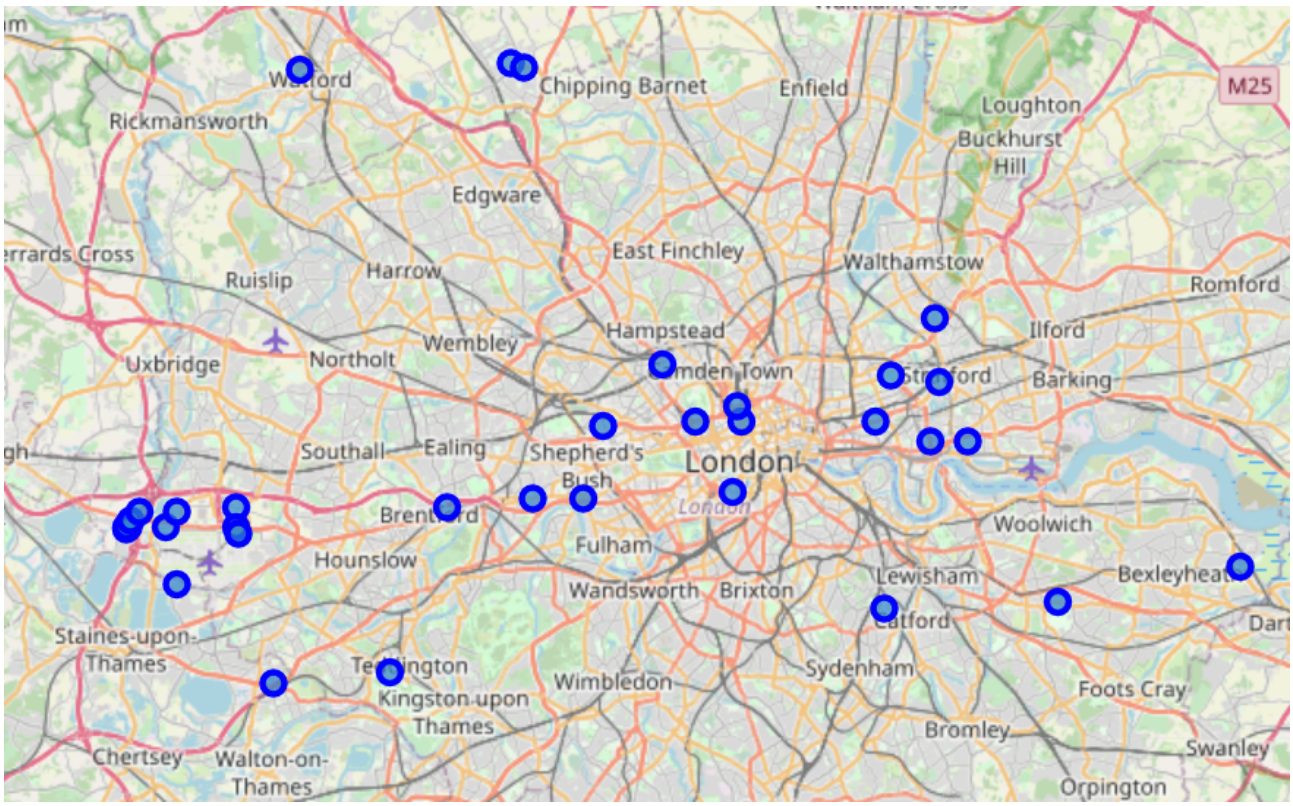


Figure 2.3: London Stations

2.2.4 Filling missing values using Spatial interpolation

The table 4.1 shows the percentage of missing values for each value in the Chronological dataframe.

Code	Date	NOx	NO_2	NO	O_3	PM_{10}	$PM_{2.5}$	Ws	Wd	Air Temperature
0%	0%	28%	26%	26%	77%	22%	12%	68%	68%	68%

Table 2.1: Percentage of missing before spatial interpolation

Applying interpolation spatial, the percentage of the missing values are found. This has two steps are the following calculate distance matrix and spatial interpolation of the nine nearest neighbours.

Distance matrix

It is used the library "scikit-learn" to calculate the distance matrix. Metric intended for two dimensional vector spaces and it is used the haversine distance because the particular data frame has latitude and longitude variables. The haversine formula determines the great-circle distance between two points on a sphere given their longitudes and latitudes.

Let the central angle Θ between any two points on a sphere be:

$$\Theta = \frac{d}{r} \tag{1}$$

where:

- d is the distance between the two points along a great circle of the sphere (see spherical distance),
- r is the radius of the sphere.

To find the distance, the haversine function hav(θ), applied above to both the central angle (θ) and the differences in latitude and longitude is.

$$\begin{aligned} d &= 2r \arcsin\left(\sqrt{\operatorname{hav}(\varphi_2 - \varphi_1) + \cos(\varphi_1)\cos(\varphi_2)\operatorname{hav}(\lambda_2 - \lambda_1)}\right) \\ &= 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \end{aligned} \tag{2}$$

where

- φ_1, φ_2 are the latitude of point 1 and latitude of point 2 (in radians),
- λ_1, λ_2 are the longitude of point 1 and longitude of point 2 (in radians).

	CA1	BEX	CLL2	LON6	HRL	HP1	MY1	KC1	TED2	HORS	CD009	HF5	LHRBR	T55	LHR2	T54	HB014	HB009	HIL4	HS5	HS4	NEW2	NEW3	SSE009	SSE007	SLH6	SLH5	SLH9	SBC01	TH004	TH2P	TH002	WL1
CA1	0.00	26.40	4.20	19.85	19.44	14.20	2.80	3.69	17.73	6.27	3.69	6.68	19.76	22.66	19.74	23.08	14.59	14.18	21.98	11.15	8.08	12.00	13.54	24.28	24.14	24.24	24.01	23.51	21.77	11.99	9.52	9.83	11.96
BEX	26.40	0.00	22.41	8.04	43.47	15.51	24.33	28.25	37.05	22.17	22.78	28.53	43.47	46.55	43.35	46.04	38.25	37.67	46.08	34.38	30.76	15.20	12.98	48.22	48.12	48.19	48.08	47.69	42.14	14.45	16.93	17.23	16.95
CLL2	4.20	22.41	0.00	15.67	22.17	10.14	1.99	6.06	18.66	3.10	0.66	7.60	22.37	25.39	22.31	25.46	18.41	17.95	24.78	13.27	9.68	8.73	9.76	27.05	26.93	27.02	26.83	26.37	23.24	8.17	5.80	6.72	9.48
LON6	19.85	8.04	15.67	0.00	35.73	7.51	17.44	21.11	29.03	14.80	16.16	20.96	35.68	38.76	35.55	38.14	33.11	32.57	38.32	26.70	23.16	10.72	7.93	40.42	40.32	40.39	40.29	39.92	34.12	8.85	11.03	12.15	13.29
HRL	19.44	43.47	22.17	35.73	0.00	28.34	20.22	16.20	9.70	21.46	22.09	15.02	0.86	3.23	1.07	4.19	22.53	22.72	2.62	9.11	12.77	30.90	31.72	4.87	4.75	4.84	4.66	4.23	7.79	30.14	27.91	28.84	31.31
HP1	14.20	15.51	10.14	7.51	28.34	0.00	11.47	14.55	21.54	8.24	10.77	13.84	28.26	31.33	28.13	30.66	28.53	28.06	30.92	19.40	15.96	10.08	8.08	32.98	32.89	32.96	32.87	32.51	26.63	7.54	8.11	10.11	12.74
MY1	2.80	24.33	1.99	17.44	20.22	11.47	0.00	4.08	17.09	3.47	1.89	5.88	20.43	23.44	20.38	23.57	17.39	16.97	22.82	11.38	7.85	10.68	11.74	25.10	24.97	25.06	24.87	24.40	21.54	10.15	7.78	8.63	11.27
KC1	3.69	28.25	6.06	21.11	16.20	14.55	4.08	0.00	14.04	6.36	5.91	3.25	16.44	19.42	16.40	19.65	16.09	15.79	18.79	7.56	4.39	14.74	15.80	21.07	20.94	21.04	20.83	20.36	18.10	14.21	11.86	12.65	15.14
TED2	17.73	37.05	18.66	29.03	9.70	21.54	17.09	14.04	0.00	16.70	18.87	11.24	9.12	11.60	8.90	10.03	26.76	26.72	11.58	7.54	9.66	26.87	26.86	13.01	12.96	12.99	13.06	12.91	5.09	25.41	23.64	25.14	28.09
HORS	6.27	22.17	3.10	14.80	21.46	8.24	3.47	6.36	16.70	0.00	3.71	6.43	21.54	24.61	21.45	24.39	20.87	20.45	24.08	12.35	8.68	10.17	10.38	26.29	26.18	26.26	26.11	25.69	21.52	8.85	6.94	8.52	11.58
CD009	3.69	22.78	0.66	16.16	22.09	10.77	1.89	5.91	18.87	3.71	0.00	7.72	22.31	25.31	22.26	25.45	17.77	17.31	24.69	13.26	9.73	8.83	10.03	26.97	26.84	26.93	26.73	26.27	23.38	8.45	6.02	6.75	9.38
HF5	6.68	28.53	7.60	20.96	15.02	13.84	5.88	3.25	11.24	6.43	7.72	0.00	15.11	18.18	15.02	18.03	19.02	18.76	17.65	5.92	2.25	16.20	16.75	19.85	19.74	19.82	19.67	19.25	15.66	15.19	13.07	14.28	17.08
LHRBR	19.76	43.47	22.37	35.68	0.86	28.26	20.43	16.44	9.12	21.54	22.31	15.11	0.00	3.09	0.24	3.53	23.29	23.46	2.69	9.19	12.87	31.10	31.84	4.76	4.66	4.73	4.61	4.26	6.96	30.27	28.07	29.06	31.58
T55	22.66	46.55	25.39	38.76	3.23	31.33	23.44	19.42	11.60	24.61	25.31	18.18	3.09	0.00	3.21	2.50	24.96	25.20	0.84	12.26	15.94	34.12	34.90	1.67	1.57	1.64	1.54	1.31	8.29	33.33	31.12	32.06	34.54
LHR2	19.74	43.35	22.31	35.55	1.07	28.13	20.38	16.40	8.90	21.45	22.26	15.02	0.24	3.21	0.00	3.47	23.42	23.58	2.85	9.11	12.78	31.04	31.76	4.87	4.77	4.84	4.74	4.41	6.73	30.19	28.00	29.01	31.54
T54	23.08	46.04	25.46	38.14	4.19	30.66	23.57	19.65	10.03	24.39	25.45	18.03	3.53	2.50	3.47	0.00	26.69	26.89	3.12	12.20	15.83	34.17	34.76	3.21	3.22	3.21	3.42	3.54	6.07	33.20	31.09	32.18	34.79
HB014	14.59	38.25	18.41	33.11	22.53	28.53	17.39	16.09	26.76	20.87	17.77	19.02	23.29	24.96	23.42	26.69	0.00	0.62	24.12	19.31	18.80	23.10	25.58	26.06	25.91	26.02	25.65	25.10	28.70	24.35	22.08	21.20	21.40
HB009	14.18	37.67	17.95	32.57	22.72	28.06	16.97	15.79	26.72	20.45	17.31	18.76	23.46	25.20	23.58	26.89	0.62	0.00	24.36	19.24	18.61	22.51	25.01	26.32	26.17	26.28	25.91	25.36	28.76	23.79	21.54	20.63	20.80
HIL4	21.98	46.08	24.78	38.32	2.62	30.92	22.82	18.79	11.58	24.08	24.69	17.65	2.69	0.84	2.85	3.12	24.12	24.36	0.00	11.73	15.40	33.51	34.34	2.30	2.17	2.26	2.04	1.61	8.59	32.76	30.53	31.44	33.88
HS5	11.15	34.38	13.27	26.70	9.11	19.40	11.38	7.56	7.54	12.35	13.26	5.92	9.19	12.26	9.11	12.20	19.31	19.24	11.73	0.00	3.68	21.98	22.65	13.94	13.82	13.91	13.75	13.34	10.75	21.08	18.91	19.98	22.63
HS4	8.08	30.76	9.68	23.16	12.77	15.96	7.85	4.39	9.66	8.68	9.73	2.25	12.87	15.94	12.78	15.83	18.80	18.61	15.40	3.68	0.00	18.36	18.98	17.61	17.50	17.58	17.43	17.00	13.77	17.41	15.26	16.39	19.11
NEW2	12.00	15.20	8.73	10.72	30.90	10.08	10.68	14.74	26.87	10.17	8.83	16.20	31.10	34.12	31.04	34.17	23.10	22.51	33.51	21.98	18.36	0.00	2.79	35.78	35.66	35.75	35.55	35.09	31.66	2.55	3.24	2.18	2.76
NEW3	13.54	12.98	9.76	7.93	31.72	8.08	11.74	15.80	26.86	10.38	10.03	16.75	31.84	34.90	31.76	34.76	25.58	25.01	34.34	22.65	18.98	2.79	0.00	36.57	36.46	36.54	36.38	35.94	31.79	1.59	4.02	4.38	5.47
SSE009	24.28	48.22	27.05	40.42	4.87	32.98	25.10	21.07	13.01	26.29	26.97	19.85	4.76	1.67	4.87	3.21	26.06	26.32	2.30	13.94	17.61	35.78	36.57	0.00	0.15	0.04	0.43	0.96	9.28	35.00	32.78	33.72	36.17
SSE007	24.14	48.12	26.93	40.32	4.75	32.89	24.97	20.94	12.96	26.18	26.84	19.74	4.66	1.57	4.77	3.22	25.91	26.17	2.17	13.82	17.50	35.66	36.46	0.15	0.00	0.11	0.30	0.82	9.29	34.88	32.66	33.59	36.04
SLH6	24.24	48.19	27.02	40.39	4.84	32.96	25.06	21.04	12.99	26.26	26.93	19.82	4.73	1.64	4.84	3.21	26.02	26.28	2.26	13.91	17.58	35.75	36.54	0.04	0.11	0.00	0.40	0.93	9.28	34.97	32.75	33.68	36.14
SLH5	24.01	48.08	26.83	40.29	4.66	32.87	24.87	20.83	13.06	26.11	26.73	19.67	4.61	1.54	4.74	3.42	25.65	25.91	2.04	13.75	17.43	35.55	36.38	0.43	0.30	0.40	0.00	0.56	9.47	34.80	32.57	33.48	35.92
SLH9	23.51	47.69	26.37	39.92	4.23	32.51	24.40	20.36	12.91	25.69	26.27	19.25	4.26	1.31	4.41	3.54	25.10	25.36	1.61	13.34	17.00	35.09	35.94	0.96	0.82	0.93	0.56	0.00	9.51	34.36	32.12	33.01	35.43
SBC01	21.77	42.14	23.24	34.12	7.79	26.63	21.54	18.10	5.09	21.52	23.38	15.66	6.96	8.29	6.73	6.07	28.70	28.76	8.59	10.75	13.77	31.66	31.79	9.28	9.29	9.28	9.47	9.51	0.00	30.31	28.45	29.85	32.72
TH004	11.99	14.45	8.17	8.85	30.14	7.54	10.15	14.21	25.41	8.85	8.45	15.19	30.27	33.33	30.19	33.20	24.35	23.79	32.76	21.08	17.41	2.55	1.59	35.00	34.88	34.97	34.80	34.36	30.31	0.00	2.48	3.32	5.27
TH2P	9.52	16.93	5.80	11.03	27.91	8.11	7.78	11.86	23.64	6.94	6.02	13.07	28.07	31.12	28.00	31.09	22.08	21.54	30.53	18.91	15.26	3.24	4.02	32.78	32.66	32.75	32.57	32.12	28.45	2.48	0.00	2.09	5.13
TH002	9.83	17.23	6.72	12.15	28.84	10.11	8.63	12.65	25.14	8.52	6.75	14.28	29.06	32.06	29.01	32.18	21.20	20.63	31.44	19.98	16.39	2.18	4.38	33.72	33.59	33.68	33.48	33.01	29.85	3.32	2.09	0.00	3.13
WL1	11.96	16.95	9.48	13.29	31.31	12.74	11.27	15.14	28.09	11.58	9.38	17.08	31.58	34.54	31.54	34.79	21.40	20.80	33.88	22.63	19.11	2.76	5.47	36.17	36.04	36.14	35.92	35.43	32.72	5.27	5.13	3.13	0.00
HB004	20.27	45.97	24.46	40.01	19.14	34.38	22.92	20.18	26.32	26.19	23.91	22.24	20.00	20.60	20.19	22.84	9.13	9.73	19.81	19.95	21.10	30.81	33.02	21.21	21.07	21.17	20.78	20.28	26.60	31.63	29.20	28.74	29.52

Figure 2.4: Distance Matrix of 34 stations

Spatial Interpolation

The methodology to find the nearest neighbours interpolation considering multiple stations is. Let's denote $S_i^1 ....S_i^k$ the k-nearest stations to S_i such as $d(S_i, S_i^1) < d(S_i, S_i^2) < .... < d(S_i, S_i^k)$. Finally, it calculated the mean from the 9 nearest stations.

$$X_t(S_i) = \frac{1}{K}\sum_{k=1}^{K} X_t(S_i^k) \tag{3}$$

where $X_t(S_i)$ denotes the variable to be interpolated at a given point in time t and station S_i.

Year	Code	Date	NOx	NO_2	NO	PM_{10}	$PM_{2.5}$	Ws	Wd	Air Temperature
2020	0%	0%	0%	0%	0%	0%	0%	3.7%	3.7%	3.7%
2019	0%	0%	0%	0%	0%	0%	0%	4.8%	4.8%	4.8%

Table 2.2: Percentage of missing after spatial interpolation

2.2.5 Fill missing values using method in Panda

This is basically about the *'fillna'* function that is used in pandas, the method used is *'backfill'*. Finally, this method help to fill all the missing values in the data frame.

Year	Code	Date	NOx	NO_2	NO	PM_{10}	$PM_{2.5}$	Ws	Wd	Air Temperature
2020	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2019	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Table 2.3: Percentage of missing after apply fillna function

3 — Methodology

3.1 Model

Forecast the air pollution, we utilized machine learning strategy, such as linear regression. The proposed architecture is showing in the figure 3.1 of the prediction model for air pollution.

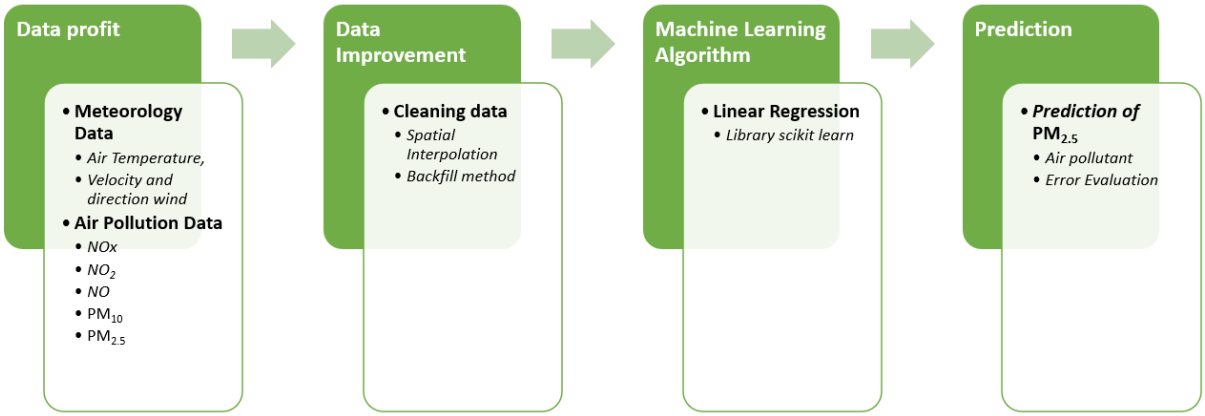


Figure 3.1: Architecture of prediction model for air pollution

After the data improvement, the filtered meteorological data and air pollution data are used to forecast the $PM_{2.5}$ level.

3.2 Correlation Coefficient

The word correlation is used in everyday life to denote some form of association.However, in statistical terms we use correlation to denote association between two quantitative variables. We also assume that the association is linear, that one variable increases or decreases a fixed amount for a unit increase or decrease in the other.

The degree of association is measured by a correlation coefficient, denoted by r. It is sometimes called Pearson's correlation coefficient after its originator and is a measure of linear association.

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{1}$$

where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

3.2.1 Looking at data: scatter diagrams

There are generated three different scatter plots to observed the relationship between values. The vertical scale represents $PM_{2.5}$ and the horizontal scale represents Oxides of Nitrogen (NO_x), Nitrogen Dioxide(NO_2), Nitric Oxide(NO), Particulate Matter(PM_{10}), wind speed (ws), wind direction (wd), air temperature.

Figure 3.2 shows the correlation in all station between 2019 and 2020. The strength correlation between all values is with PM_{10}. The other values can help in the regression but the model probably take PM_{10} as a unique correlation value.

It has compared the relation between all station. The station that showed the clear correlation between $PM_{2.5}$ and PM_{10} is the station in London Honor Oak Park (HP1), figure 3.3. On the other hand, the figure 3.5 shows a correlation but it is not that clear than station HP1.

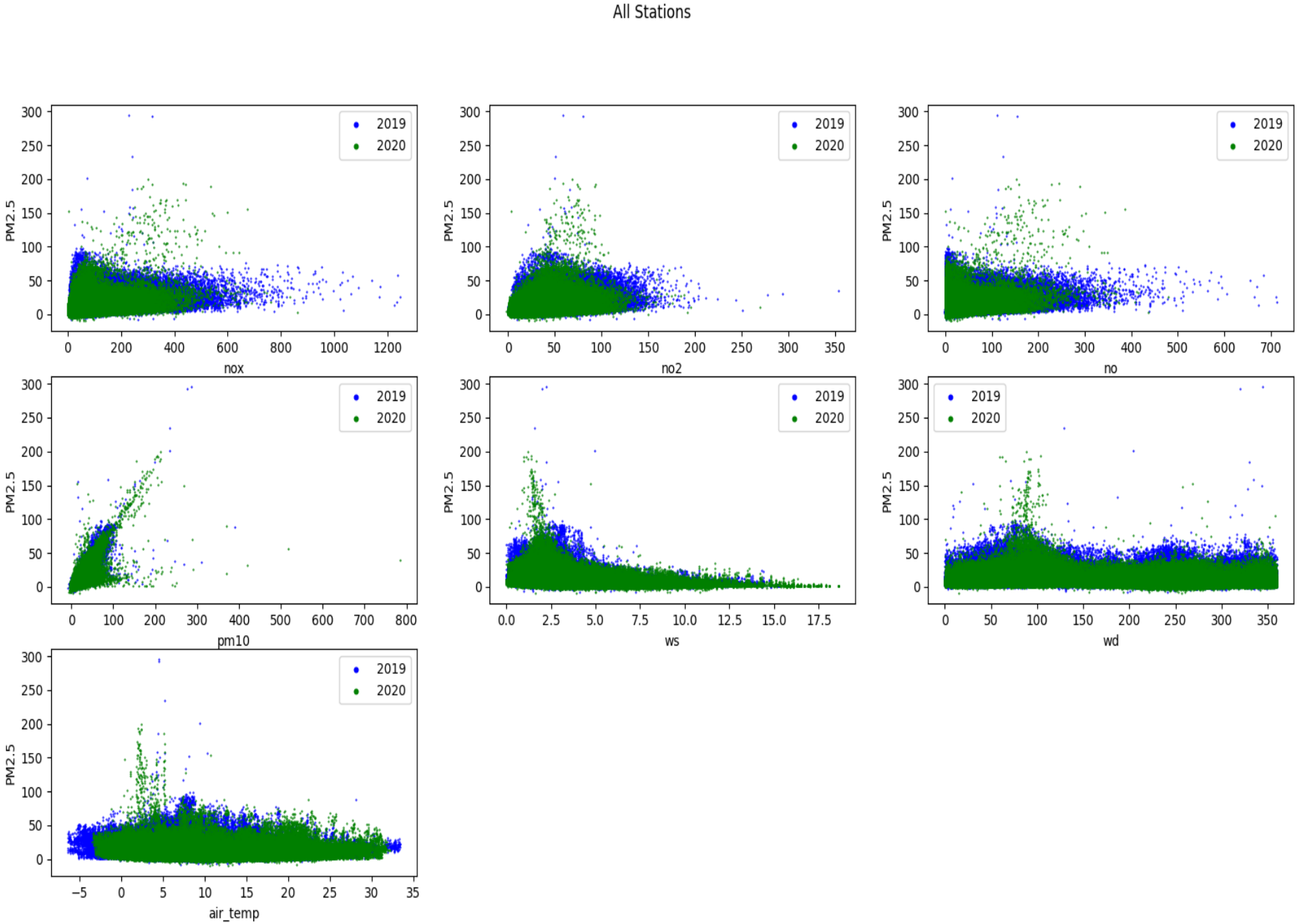


Figure 3.2: Correlation between $PM_{2.5}$ with all stations.

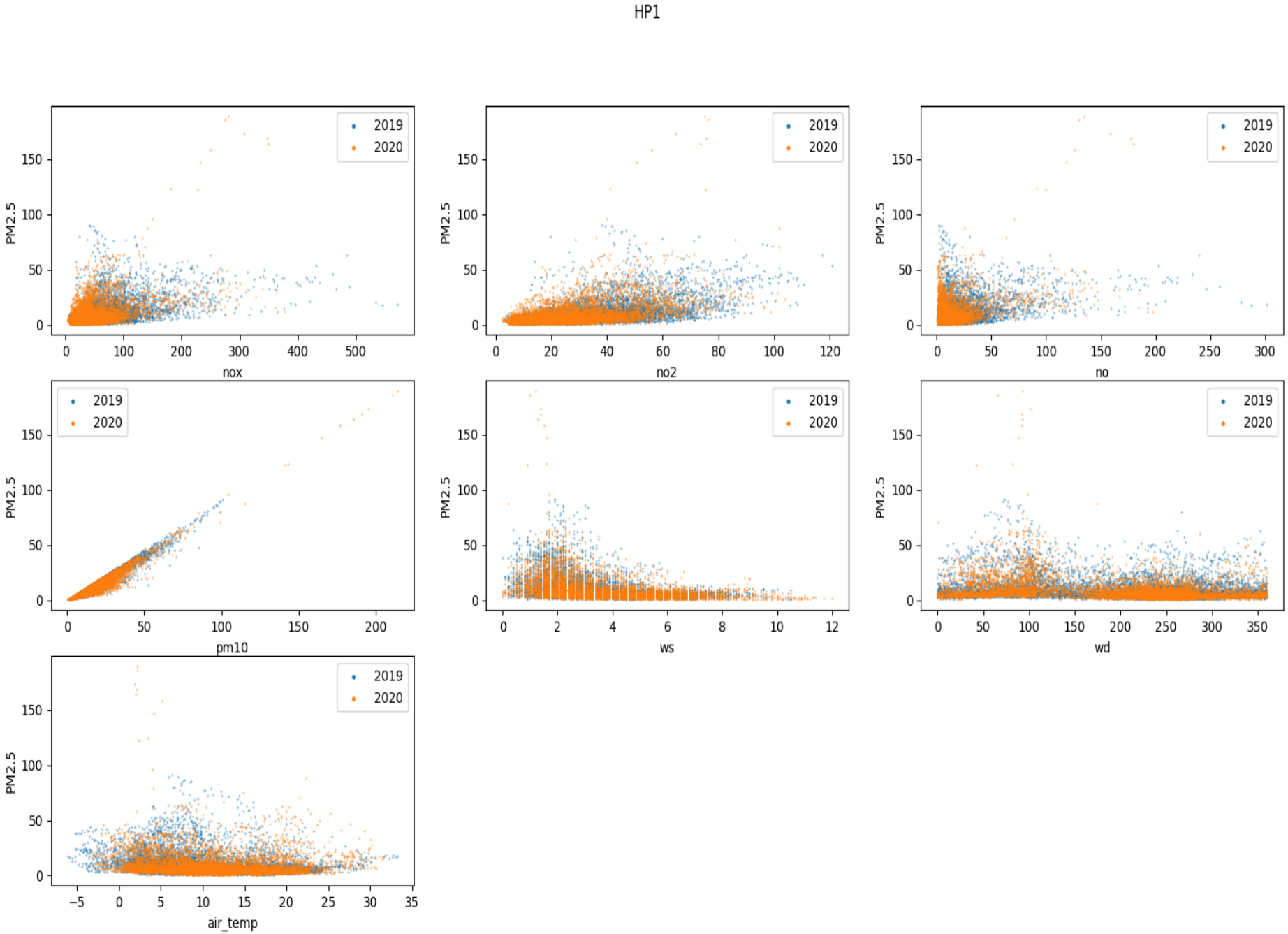


Figure 3.3: Correlation between $PM_{2.5}$ in London Honor Oak Park (HP1) Station.

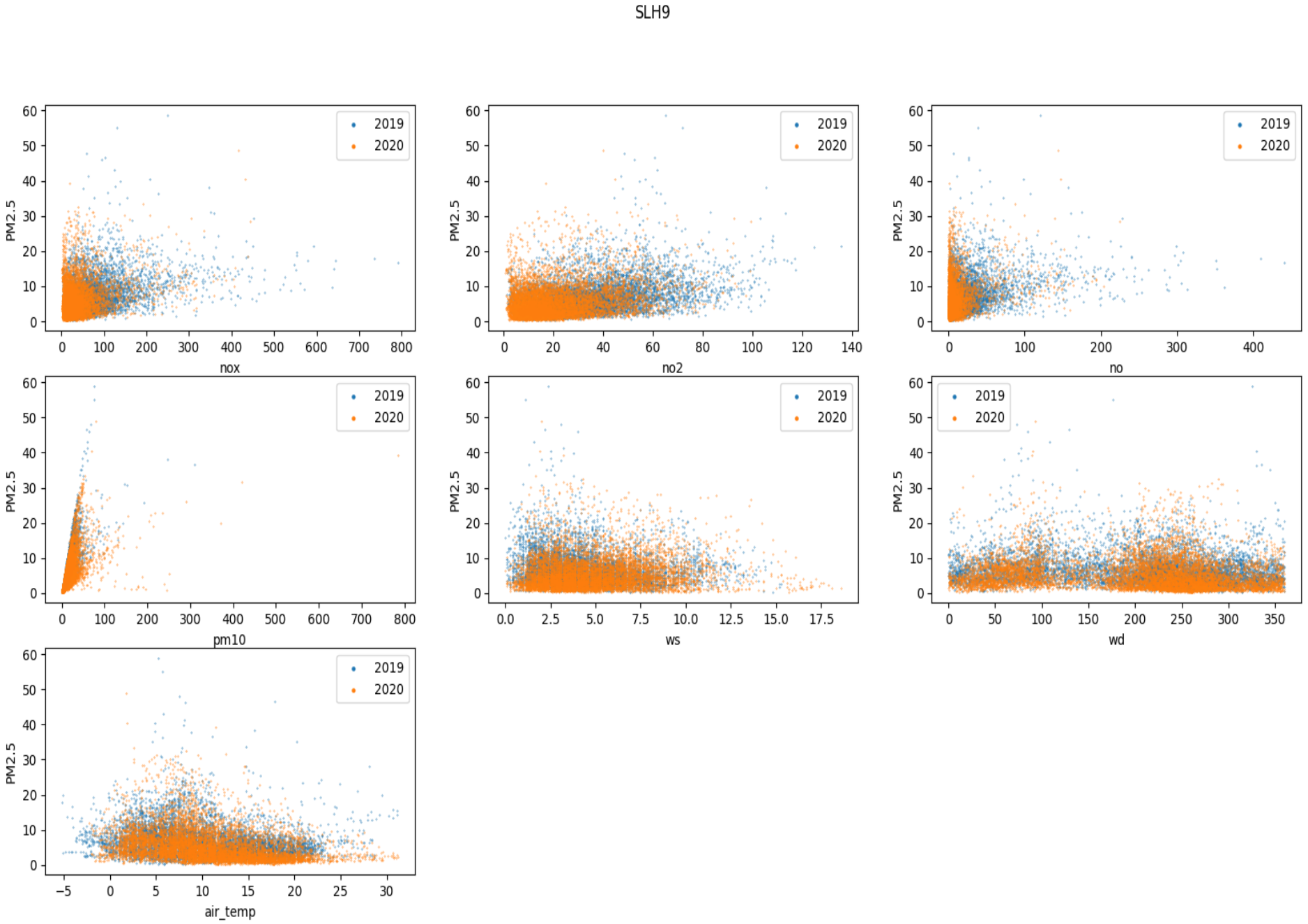


Figure 3.4: Correlation between $PM_{2.5}$ in Slough Lakeside Osiris (SLH9) Station.

3.3 Linear Regression

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables.

$$Y_{PM_{2.5}} = \beta_0 + \beta_1 X_{NO_x} + \beta_2 X_{NO_2} + \beta_3 X_{NO} + \beta_4 X_{PM_{10}} + \beta_5 X_{ws} + \beta_6 X_{wd} + \beta_7 X_{T_{air}} + \epsilon_i \quad (2)$$

where:

- $Y_{PM_{2.5}}$ dependent variable,
- X_i explanatory variables Oxides of Nitrogen (NO_x), Nitrogen Dioxide(NO_2), Nitric Oxide(NO), Particulate Matter(PM_{10}), wind speed (ws), wind direction (wd), air temperature
- β_i y-intercept (constant term)
- ϵ_i the model's error term (also known as the residuals)

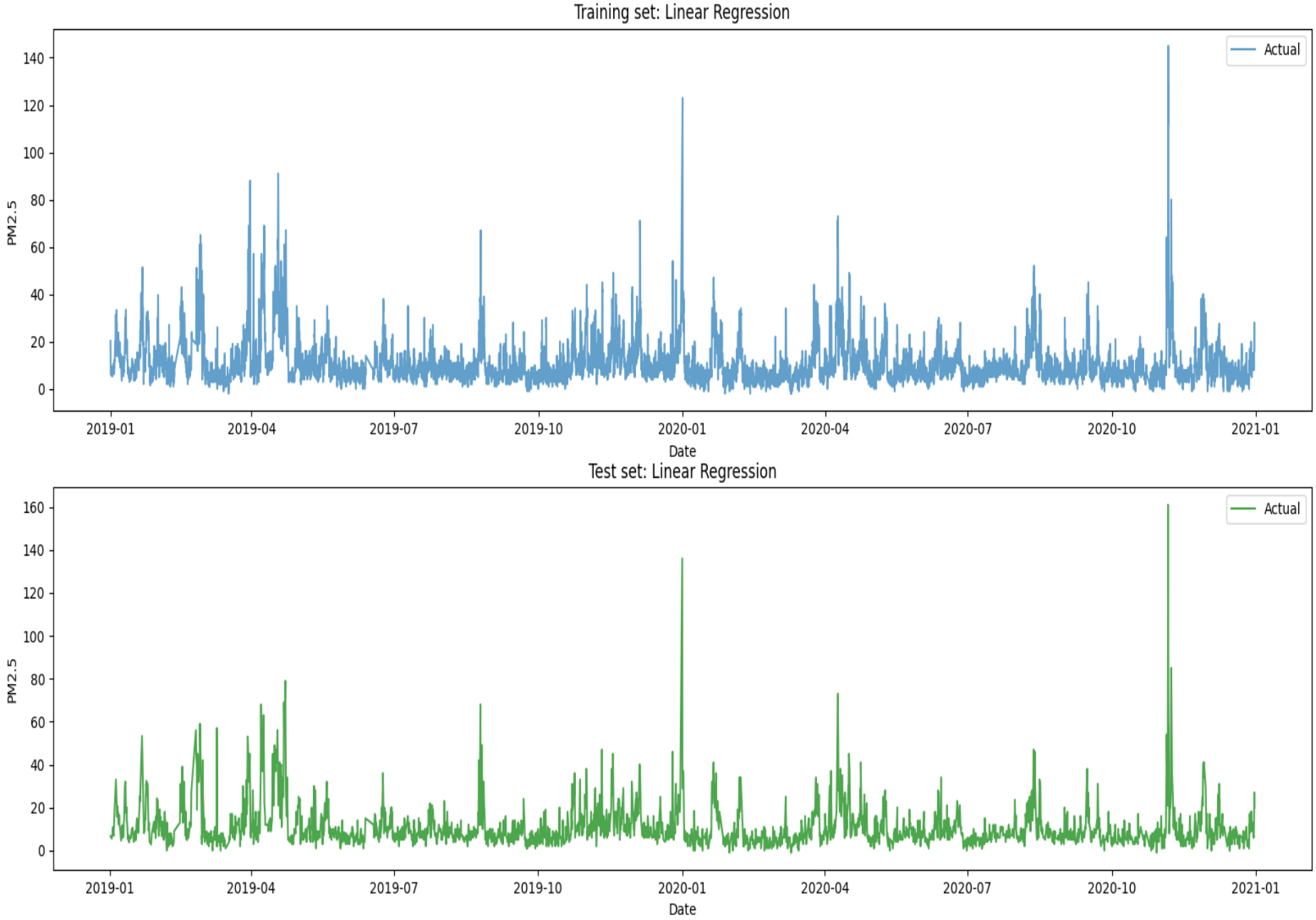


Figure 3.5: Prediction model for Lineal Regression.

4 — Results

4.1 Coefficient of Determination R^2

The UK air data set is stationary time-series data from January 2019 to December 2020. It works with different models for each stations. The figure 4.1 shows the mean squared error for each station with these results it can choose the best and worst model.

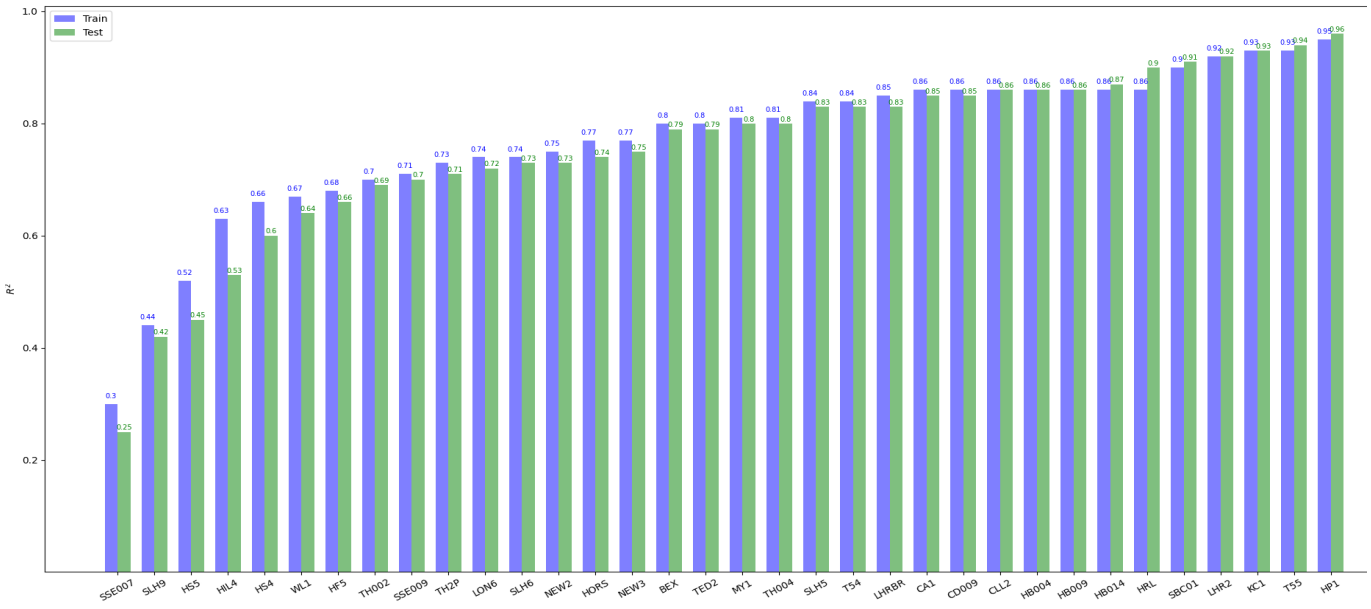


Figure 4.1: R^2 for each station.

4.2 Forecasting performance for station HP1

Figures 4.2 and 4.3 show the results of prediction using the best model (Station HP1) between studied stations around London, first graph is training set with a $R_2 = 0.95$ and test set with a $R_2 = 0.96$. Moreover, the blue and green line indicates the actual values in training and test set, respectively and brown line the prediction of Particular Matter $PM_{2.5}$ in both figures. Also, figures illustrate the error amount of $PM_{2.5}$ between actual values and predicted values.

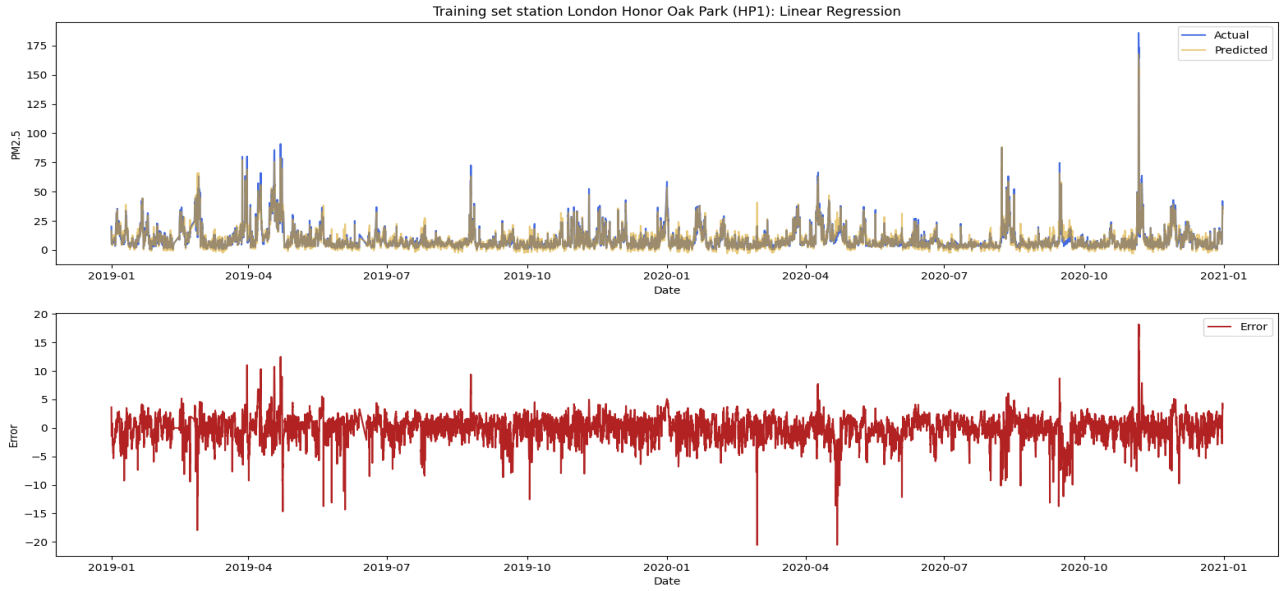


Figure 4.2: Prediction model for Lineal regression Station HP1.(Train set)

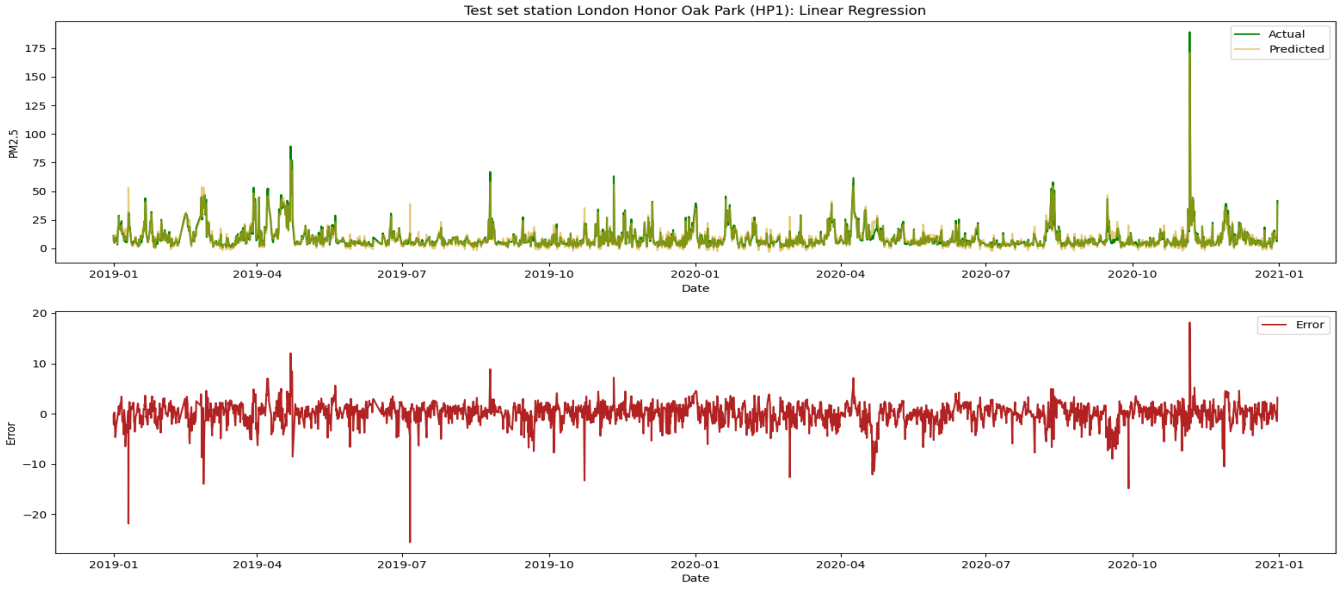


Figure 4.3: Prediction model for Lineal regression Station HP1.(Test set).

4.3 Forecasting performance for station SSE

Similarly, figures 4.4 and 4.5 show the result of prediction using the worst model (Station SEE007) between studied stations where first graph is training set with a $R_2 = 0.28$ and test set with a $R_2 = 0.32$

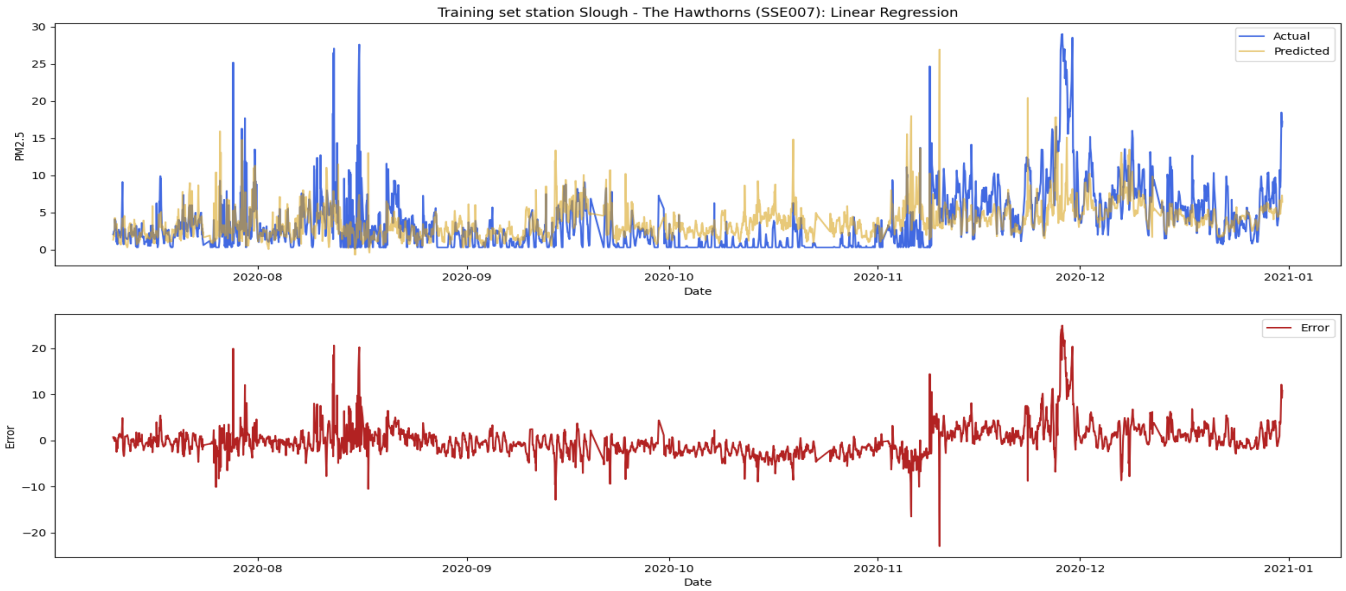


Figure 4.4: Prediction model for Lineal regression Station SSE.(Train set)

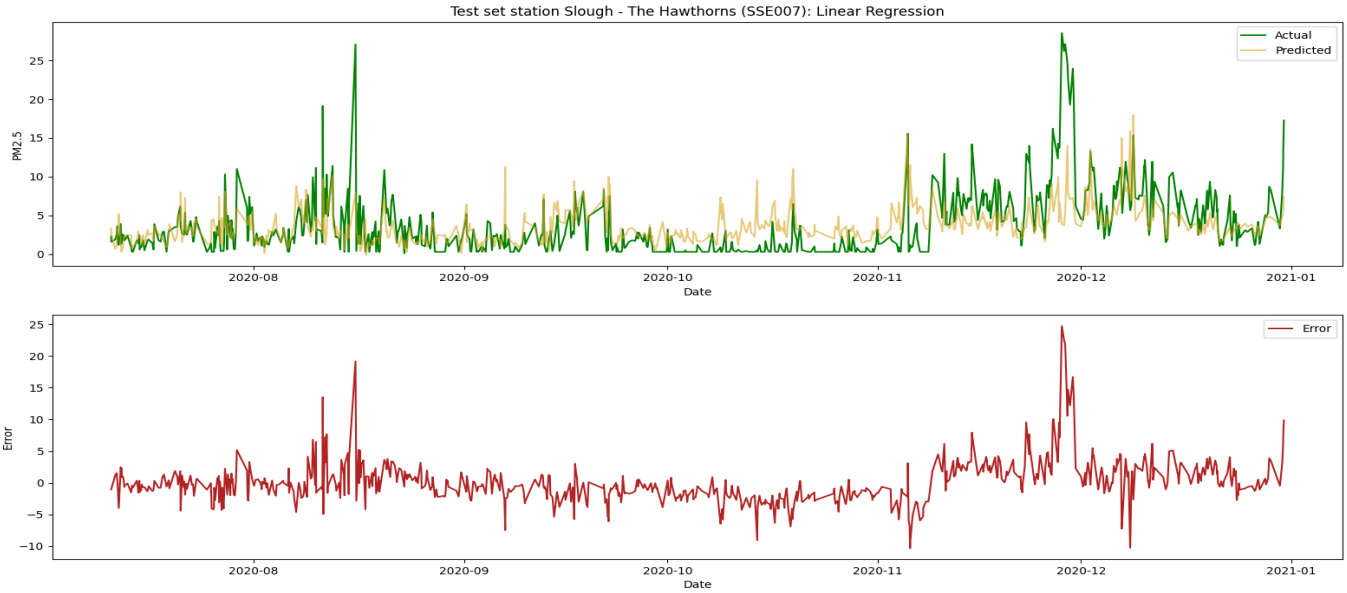


Figure 4.5: Prediction model for Lineal regression Station SSE.(Test set).

4.4 Regression coefficient and intercept

Figure 4.1 shows the regression coefficients and intercept for best station and worst stations.

Features	*Station HP1*	*Station SE007*
β_0 y-intercept	-0.102	4.10
β_1 Oxides of Nitrogen (NO_x)	-0.161	-0.012
β_2 Nitrogen Dioxide (NO_2)	0.144	0.016
β_3 Nitric Oxide (NO)	0.258	0.083
β_4 Particulate Matter (PM_{10})	0.797	0.194
β_5 wind speed (*ws*)	-0.254	0.065
β_6 wind direction (*wd*)	0.001	-0.003
β_7 air temperature (T)	-0.119	-0.210

Table 4.1: The best and worst stations results of the diagnostics train and test values for forecasting of the $PM_{2.5}$

Figure 4.6 shows the regression coefficient and intercept for all values.

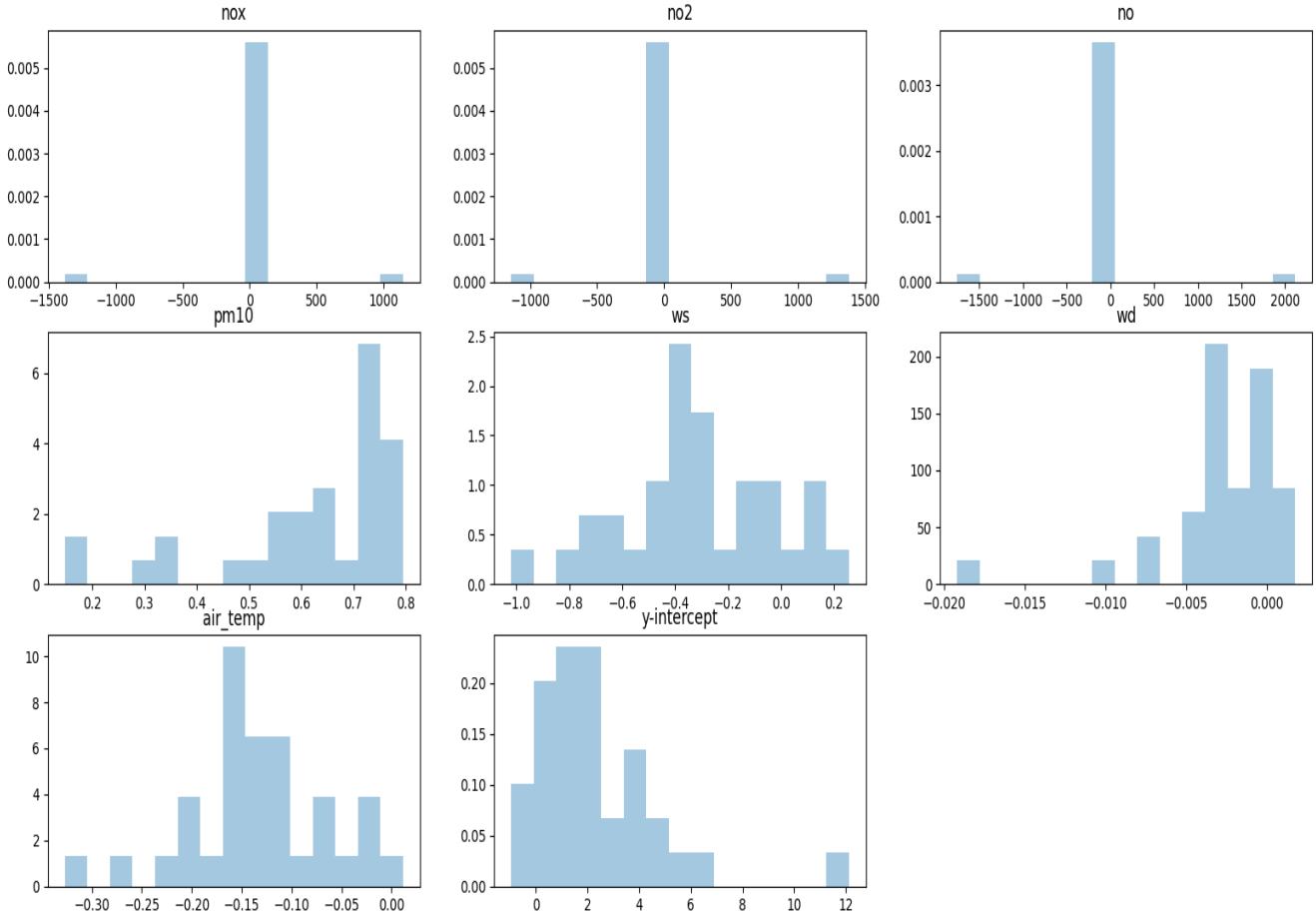


Figure 4.6: Prediction model for Lineal regression Station SSE.(Test set).

4.5 Model results

Figure 4.7 shows correlation between $PM_{2.5}$ and PM_{10} in the best and worst station.

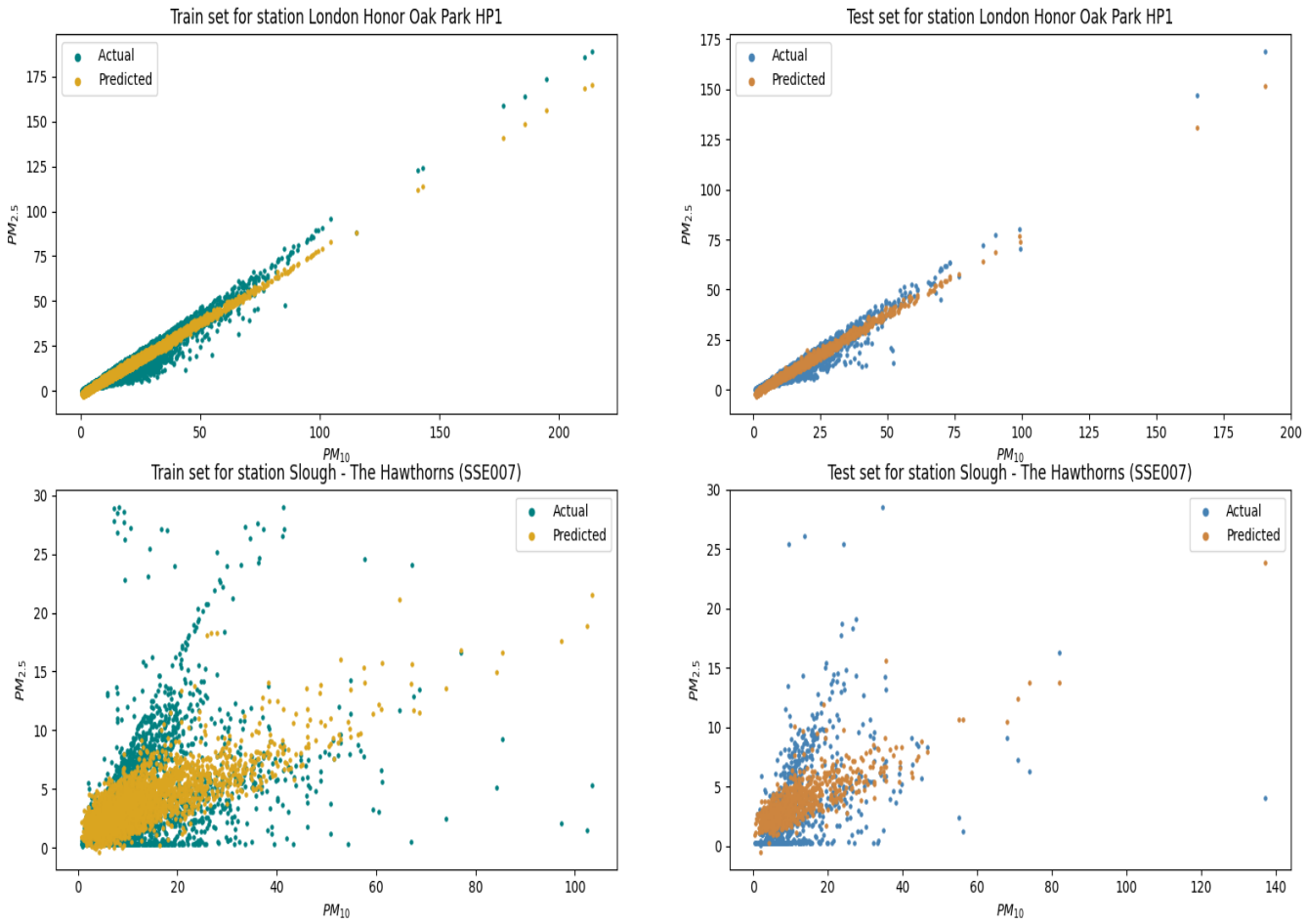


Figure 4.7: Prediction model for Lineal regression Station SSE.(Test set).

Table 4.2 shows the results for lineal regression algorithm performs compared with the worst and better station.

Features	*Station HP1*		*Station SE007*	
	Train	*Test*	*Train*	*Test*
Coefficient of Determination (R^2)	0.95	0.96	0.28	0.32
Mean Square Error (MSE)	4.95	4.70	11.98	10.19

Table 4.2: The best and worst stations results of the diagnostics train and test values for forecasting of the $PM_{2.5}$

5 — Discussion

5.1 Coefficient of Determination R^2

More than 50% of 32 stations have a coefficient of determination (R_2) bigger than 70%. So, it is possible to design algorithm for each station and can forecast $PM_{2.5}$.

5.2 Forecasting performance station HP1

The best coefficient of determination is for station HP1 with 95% and 95% to train and test respectively. This is in South London around Honor Oak Park, station has 98% of the data from 2019 to 2020. This help to design a better methodology to forecast $PM_{2.5}$. Figure 4.2 shows the efficiency of algorithm and error is attenuated.

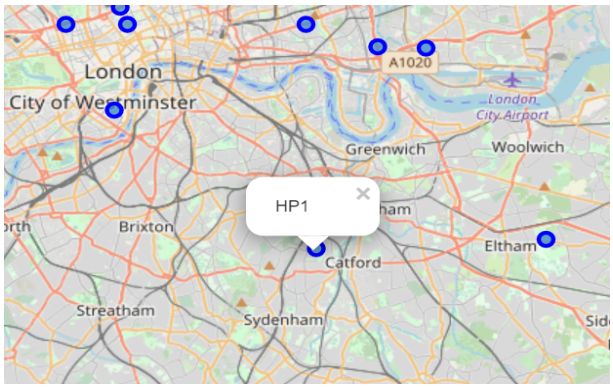


Figure 5.1: Location of Station in London Honor Oak Park (HP1).

5.3 Forecasting performance station SSE007

The worst coefficient of determination is for station HP1 with 30% and 25% to train and test respectively. This is near Heathrow Airport in Slough - The Hawthorns, station has 25% of the data from 2019 to 2020. The number of data is the reason that algorithm has a deficient performance to forecast $PM_{2.5}$. Figure 4.4 shows the poor efficiency of th model and error has values too high.

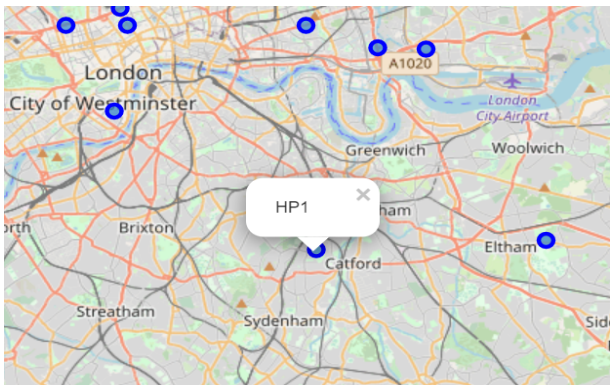


Figure 5.2: Location of Station in London Honor Oak Park (HP1).

The figure 5.3 shows the number of data for each station. This is important feature to design a algorithm with high performance.

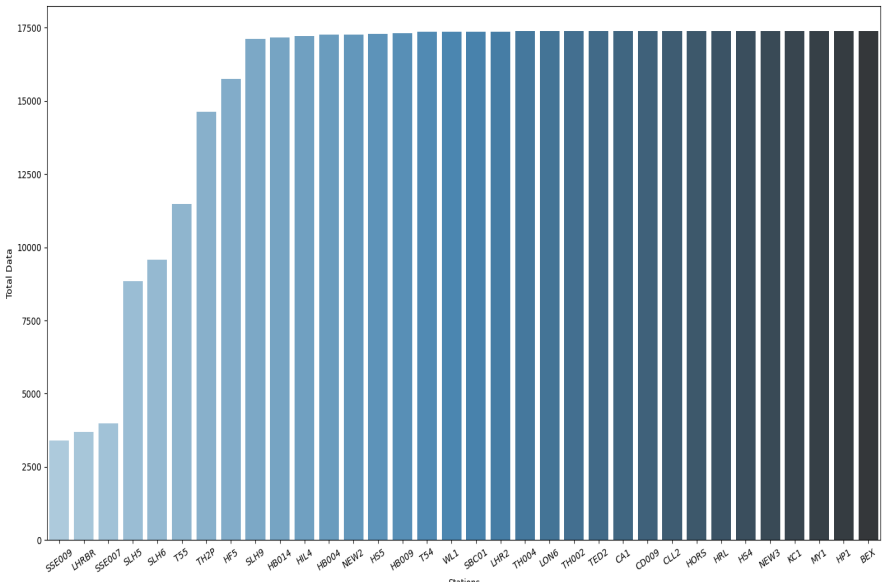


Figure 5.3: Location of Station in London Honor Oak Park (HP1).

5.4 Regression coefficient and intercept

The algorithm has 7 coefficient that it get more information of this values. However, figure 4.6 shows how the value PM_{10} is the most correlated with $PM_{2.5}$. It is possible that doing linear regression method with PM_{10} the results are similar but PM_{10} started data are same than $PM_{2.5}$ and the study has less data than the actually study.

5.5 Model results

Figure 4.7 shows how the linear regression can help to design an algorithm to forecast $PM2.5$ but it is better take the station that have more than 80% of the total data because station with less than 70% of data could design algorithm with high error.

6 — Conclusions and Future Work

In this work, the proposed machine learning models to analyse the air pollution on London Air Quality on the City of London is presented. There are 150 air pollution stations are recorded in London Air Quality data from the last few years. In the project only is used 32 stations from 2019 to 2020. The particular matter $PM_{2.5}$ forecast are done using linear regression. The results shows that this kind of model can help to forecast values to increased the data for future studies in this kind of matter or in other cases.

List of Figures

List of Tables