DEPARTMENT OF MECHANICAL & AEROSPACE ENGINEERING

University of Strathclyde Engineering

Class Code/ Title:	ME420 – Individual Project
Technical paper title:	Investigating the relationship between Air Pollution & Respiratory Disease using Machine Learning & Satellite Data
Student Name/ Number:	Mohammed Zenudeen Shehbaz/201818756
Supervisor:	Dr Annalisa Riccardi
Date:	15/03/2022
Word count:	2740

Abstract

Monitoring concentrations of particulate matters (PM) is significant in understanding the safety of air for normal living conditions. Furthermore, there is clear evidence that an increase in ambient air pollution results in an increased number of cases for respiratory and cardiovascular diseases [1]. This study aims to analyse the correlation between predicted concentrations of $PM_{2.5/10}$ and the number of hospital admissions for respiratory diseases in England for April 2020/21 using data from remote sensing satellites and a machine learning prediction model. Precursors and chemical composition of $PM_{2.5/10}$ from Sentinel-5P were used alongside auxiliary features in the analysis. A gradient-boosting model known as Light Gradient-Boosting Machine (LightGBM) proved to be most suitable in predicting $PM_{2.5/10}$ concentrations. The predictions were then used to analyse the correlation with the number of hospital admissions for respiratory diseases.

The concentrations were well estimated, based upon metrics evaluating model performance. However, only a weak positive correlation was observed, which may have been due to limitations including spatial-temporal availability of clinical data, and the simplifying assumptions made.

Contents

Nomenclature

Abbreviations	Meaning
AOD	Aerosol Optical Depth
CCG	Clinical Commissioning Group
CND	Column Number Density
CTM	Chemical Transport Model
DB	Deep Blue
DT	Dark Target
LightGBM/LGBM	Light Gradient Boosting Machine
MAE	Mean Absolute Error
ML	Machine Learning
MODIS	Moderate Resolution Imaging Spectroradiometer
MSE	Mean Squared Error
NCEP	National Centres for Environmental Prediction
NDVI	Normalised Difference Vegetation Index
PM	Particulate Matter
RMSE	Root Mean Squared Error
WHO	World Health Organisation

Symbols	Meaning
Kg/m^2	Kilograms per metre squared
Km	Kilometres
m	Metres
µm	Micrometres/Microns
$µg/m^3$	Micrograms per meter cubed
Mol/m^2	Mols per metre squared
m/s	Metres per second
Pa	Pascal

Parameter	Units
Cloud Height	m
CO	Carbon Monoxide (Mol/m^2)
Elevation	m
MSE	$(µg/m^3)^2$
NDVI	Dimensionless
NO_2	Nitrogen Dioxide (Mol/m^2)
O_3	Ozone (Mol/m^2)
$PM_{2.5/10}$/RMSE/MAE	$µg/m^3$
Population Density	Dimensionless
Precipitable Water	Kg/m^2
Relative Humidity	%
Tropopause Pressure	Pa
U component of wind	m/s
V component of wind	m/s

1.0 Introduction

Air pollution presents one of the most detrimental impacts on society and the environment. Ever-increasing traffic, industrialisation, and urbanisation, all contribute to the increase of air pollution. According to the World Health Organisation (WHO) [2], air pollution is the cause of approximately 7 million premature deaths annually. In perspective, air pollution causes more deaths each year than COVID-19 ever has [3]. Clearly, this issue must be dealt with as soon as possible.

Air pollution is the contamination of an area, by cause of chemical or biological means, which has altered the natural state of the atmosphere [2]. The most destructive of pollutants are PM, which originate from the burning of fossil fuels, forest fires and exhausts from automobiles. These pollutants are responsible for a large number of respiratory and cardiovascular disease cases. In this study $PM_{2.5}$ and PM_{10}, which exhibit a diameter of 2.5 and 10 microns or less [4] respectively, will be analysed.

Air-quality monitoring stations provide reliable measurements of PM, however, only provide measurements local to its surrounding. There are approximately 300 stations in the UK, primarily located in larger cities [5]. However, it is clear from Figure 1 shown below, that there are far fewer located towards the North. The cost of construction and development inhibit air-quality stations from being common globally, especially in poorer regions.

Satellites offer a continuous distribution of data including meteorological, vegetation and PM chemical composition, across much of the planet [6]. Utilising satellite measurements to estimate concentrations of PM can prove to be a viable alternative to the limited spatial coverage of air-quality monitoring stations.

Machine learning (ML) is a subset of Artificial Intelligence, capable of identifying patterns from past data to make informed predictions for unseen data [7]. This analysis is classed as a regression problem where the output is in the form of continuous numerical values.

The main aims of this study are:

- To make effective use of Data Science techniques to assimilate and pre-process satellite and health board data.
- To develop an ML model which can accurately predict the concentration of $PM_{2.5/10}$ in the UK.
- To analyse the correlation between the predicted PM concentrations and the number of hospital admissions for respiratory disease in England.

2.0 Literature Review

This section highlights previous methods of predicting PM from remote sensing satellites. Among these are methods which depend on PM chemical composition, spatial interpolation, and Aerosol Optical Depth.

2.1 Aerosol optical Depth

Due to its high correlation with PM, utilising Aerosol Optical Depth (AOD) as a primary feature is common in remote sensing satellite studies. (Chen et al.,2018) [8], created a method which used satellite measured AOD and a random forest model to estimate daily concentrations of PM_{10} in China. In this approach, Temperature, Humidity, Wind Speed, etc. were used as auxiliary features. AOD was downloaded from the Moderate Resolution Imaging Spectroradiometer (MODIS) at 10km spatial resolution. AOD products, Deep Blue (DB) and Dark Target (DT), were merged to improve the spatial coverage, as there is significant missing data with AOD products. To attempt to fill the gaps in data, regression models were formed relating the products to one another. The random forest model performed best based on the model's evaluation metrics. Despite promising results, this method presents limitations in missing AOD data, which may be a larger problem if a similar analysis was carried out on a larger scale.

2.2 Spatial Interpolation

To overcome the limitations of missing AOD, (Xiao et al.,2017) [9] presented a statistical method utilising a Chemical Transport Model (CTM). This study focused on Mexico City at a spatial resolution of 1km between 2013 - 2014 and was driven by overcoming missing AOD. The ML model took data from each pixel of the image and provided reliable measurements for estimated $PM_{2.5}$. However, this approach is highly reliant on the presence of air-quality monitoring stations and would not be as effective in areas where stations are scarce.

2.3 PM Chemical Composition/Precursors

Recent studies consisting of chemical compositions/precursors of PM prove to be a viable method of predicting accurate concentrations of PM. Sentinel-5P satellite's measurements are utilised as the main features in a study by (Ayako Kawano., 2021) [10]. A prediction model for $PM_{2.5}$ was constructed, and the concentration was related to adverse birth defects in Nepal and Bangladesh. Due to the scarcity of monitoring stations in this region, data points from India were used to increase the data used for training the model. From Sentinel-5P, Nitrogen Dioxide, Ozone and Carbon Monoxide were used, alongside meteorological, urbanisation and vegetation data. To remove potential multicollinearity, low contributing features were removed from the model. The accuracy of the concentrations was relatively high, but it was noted that the performance varied based on the time of year. However, the quality of the results and the availability of data makes this methodology more suitable for this analysis.

Figure 1: Spatial distribution of UK monitoring stations [5]

3.0 Experimental Data

3.1 Ground measured PM

The accuracy of the predicted $PM_{2.5/10}$ concentration is determined by the closeness to ground measured concentrations in the same location. Ground measurements of $PM_{2.5/10}$ were collected for the UK from the open-source air-quality monitoring repository, OpenAQ [11]. OpenAQ provides real-time and historical air quality data across the globe.

3.2 NHS Health Board Data

The number of emergency hospital admissions for Respiratory Disease, Asthma and Chronic Obstructive Pulmonary Disease (COPD) for several clinical commissioning groups (CCGs) in England between April 2020 and April 2021 was obtained from the NHS Digital platform [12]. Table 1 below represents the first rows of what is a larger data set of NHS based data.

3.3 Satellite data sources

Satellite measurements were obtained from the Google Earth Engine catalogue and accessed with the Python module. The data includes air quality, meteorological, vegetation and urbanisation data. The maximum, minimum and mean were all considered [13].

3.4 Sentinel-5P

The Sentinel-5 Precursor (Sentinel-5P) launched on October 13, 2017, is a satellite which was dedicated to monitor the condition of the atmosphere [14]. The aim of this satellite was to acquire atmospheric measurements with a high spatial-temporal resolution, which was carried out by the Tropospheric Monitoring Instrument (TROPOMI) on board the satellite. This study acquired measurements for, Nitrogen Dioxide, Carbon Monoxide, and Ozone.

3.5 Meteorological/Vegetation data

The meteorological features were the horizontal velocity of wind, vertical velocity of wind, relative humidity, specific humidity, and total precipitable water. These originate from the National Centres for Environmental Prediction (NCEP) [15]. Normalised Difference Vegetation Index (NDVI) was an input feature. NDVI is a measurement of the difference between near-infrared light (which vegetation reflects) and red light (which vegetation absorbs) [16]. Surface elevation, cloud free coverage and population density were also included. Table 2 below summarises the satellite measurements for the final model.

Table 1: NHS Emergency Admissions England 2020/21

CCG Location	Respiratory Admissions	Asthma Admissions	COPD Admissions
Northumberland	749	156	130
North Cumbria & Morecambe Bay	721	107	117
County Durham	797	139	214
Newcastle, Tyneside & Sunderland	935	170	229
Tees Valley	926	152	205

Table 2: Input features of final model

Satellite/Date/Resolution	Bands/Units	Description
Sentinel-5P OFFL CO: 2018/06/28 – 2022/02/24 Resolution: 1113 meters	CO CND Cloud Height	- Vertically integrated CO column density. - Scattering layer height.
Sentinel-5P OFFL NO2: 2018/06/28 – 2022/02/17 Resolution: 1113 meters	NO_2 CND NO_2 Slant CND NO_2 Tropospheric CND NO_2 Stratospheric CND Absorbing Aerosol Index Tropopause Pressure	- Total vertical column of NO_2 - NO_2 slant column density - Tropospheric vertical column of NO_2 - Stratospheric vertical column of NO_2 - Aerosol index - Tropopause pressure
Sentinel-5P OFFL O3: 2018/09/08 – 2022/02/24 Resolution: 1113 meters	O_3 CND	- Total atmospheric column of O_3 between surface and the top of atmosphere
NOAA CDR AVHRR NDVI: 1981/06/24 – 2022/02/23 Resolution: 5566 meters	NDVI QA	- Normalized difference vegetation index - Quality control bit flags
NCEP: 2015/07/01 – 2022/02/26 Resolution: 27830 meters	Relative Humidity U velocity of wind V velocity of wind Precipitable Water	- Relative humidity 2m above ground - U component of wind 10m above ground - V component of wind 10m above ground - Precipitable water for entire atmosphere
HYCOM: 1992/10/02 – 2022/02/17 Resolution: 8905 meters	Surface elevation	- Sea surface elevation anomaly relative to modelled elevation mean

DMSP OLS: 1996/03/16 – 2014/01/01 Resolution: 928 meters	Average visibility Cloud free Coverage	- The average of the visible band digital number values with no further filtering. - Cloud-free coverages tally the total number of observations that went into each 30-arc second grid cell
GPWv411: 2000/01/01 – 2020/01/01 Resolution: 928 meters	Population Density	- Estimated number of people per km^2

4.0 Methodology

The programming language chosen was Python, due to the useful modules and libraries available, and the language's versatility. The methodology flow chart is shown in Figure 2.

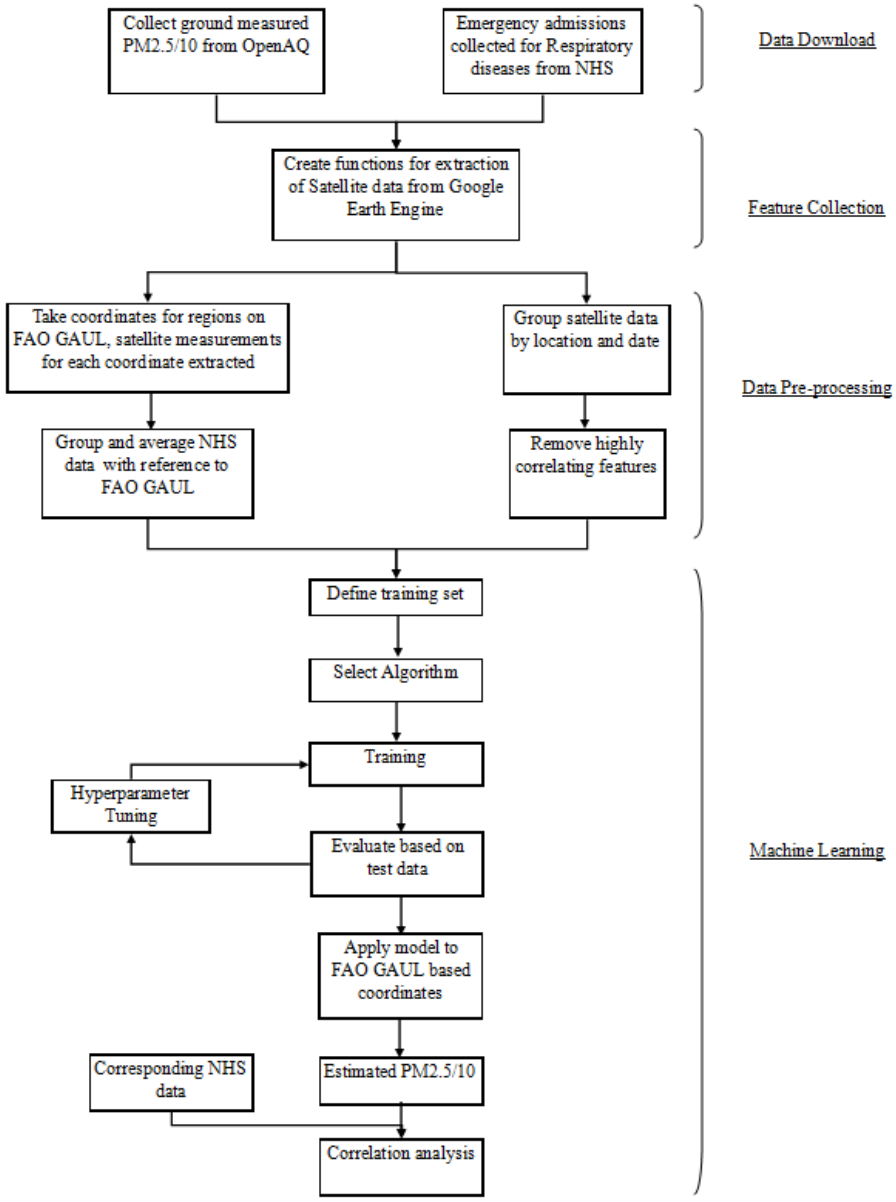

Figure 2: Methodology flow chart

4.1 Data Download

Data from OpenAQ was added and filtered to remove locations in other countries which have the same name as those located in the UK. The data consisted of locations, dates, coordinates, and the PM concentration which would later be used to train the model. The data was collected from 27/09/2018 - 31/03/2020.

The number of emergency hospital admissions for Respiratory disease, COPD and Asthma in England for April 2020/21 were downloaded from NHS Digital.

4.2 Feature Collection

The Google Earth Engine, and other data analysis modules were imported to the programme. Functions for each satellite were created in Python to extract measurements based on dates and coordinates. The outputs of all the functions were merged based on corresponding locations and dates.

4.3 Data Pre-processing

Before data was input to the model, a correlation matrix was created to remove features which observed a correlation greater than 80%, to avoid potential multicollinearity. These included: H_2O, Temperature and Specific Humidity.

NHS data was grouped into areas which corresponded to the regions marked on the FAO GAUL map in Figure 3 below. The number of admissions for each area was averaged, which represented an overall value. Approximately 15 coordinates for each region were taken. Satellite measurements for each coordinate were obtained. This defined the data used for the predictions.

4.4 Machine Learning

Input data was separated into training and testing sets with an 80:20 split, respectively. The machine learning algorithms which were compared in this analysis were: Random Forest, Gradient-Boosting Regressor, XGBoost and LightGBM. A five-fold cross-validation was implemented for hyperparameter tuning to prevent overfitting. The models were compared, and the best was selected based upon the evaluation metrics Root-Mean-Squared, Mean-Squared, Mean-Absolute error (RMSE, MSE, MAE) and R^2 which are commonly used to analyse the model performance in regression analysis.

The best performing model was used to predict concentrations in England between April 2020/21. These concentrations were correlated with the number of hospital admissions for corresponding regions. The metrics which were used to analyse this were the Pearson and Spearman correlation metrics.

Figure 3: FAO GAUL Level 2

5.0 Results

Tables 3 and 4 below provide a summary of the performance of each of the models on the test set. The LightGBM model is shown to be the most effective in the case of both $PM_{2.5}$ and PM_{10}. RMSE of 1.22 - 1.37, and R^2 of 0.96 - 0.97, represent a fairly low error in the predictions. Low RMSE and R^2 close to a value of one are desirable, and were calculated with equations in Figure 4:

5.1 Evaluation Metrics

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Where,

$\hat{y}$ – *predicted value of y*

$\bar{y}$ – *mean value of y*

Figure 4: RMSE/R^2 score equations [17]

Table 3: PM2.5 Evaluation Metrics

Model (PM2.5)	MAE (µg/m^3)	MSE ((µg/m^3)^2)	RMSE (µg/m^3)	R^2
Random Forest	0.76	1.48	1.22	0.96
Gradient Boosting	0.85	1.27	1.13	0.97
XG Boost	0.72	1.75	1.32	0.95
Light GBM	0.65	1.48	1.22	0.96

Table 4: PM10 Evaluation Metrics

Model (PM10)	MAE (µg/m^3)	MSE ((µg/m^3)^2)	RMSE (µg/m^3)	R^2
Random Forest	1.11	2.69	1.64	0.95
Gradient Boosting	1.19	2.51	1.59	0.96
XG Boost	0.89	2.05	1.43	0.96
Light GBM	0.86	1.88	1.37	0.97

5.2 Model Performance

In Figure 5 and 6 below, the predicted concentration for $PM_{2.5}$ and PM_{10} were plotted against ground measured concentrations of $PM_{2.5}$ and PM_{10}, respectively, for the LightGBM model. From the scatter plot, it is evident that a mostly linear variation is present, which is desirable. A higher percentage of low concentrations is observed.

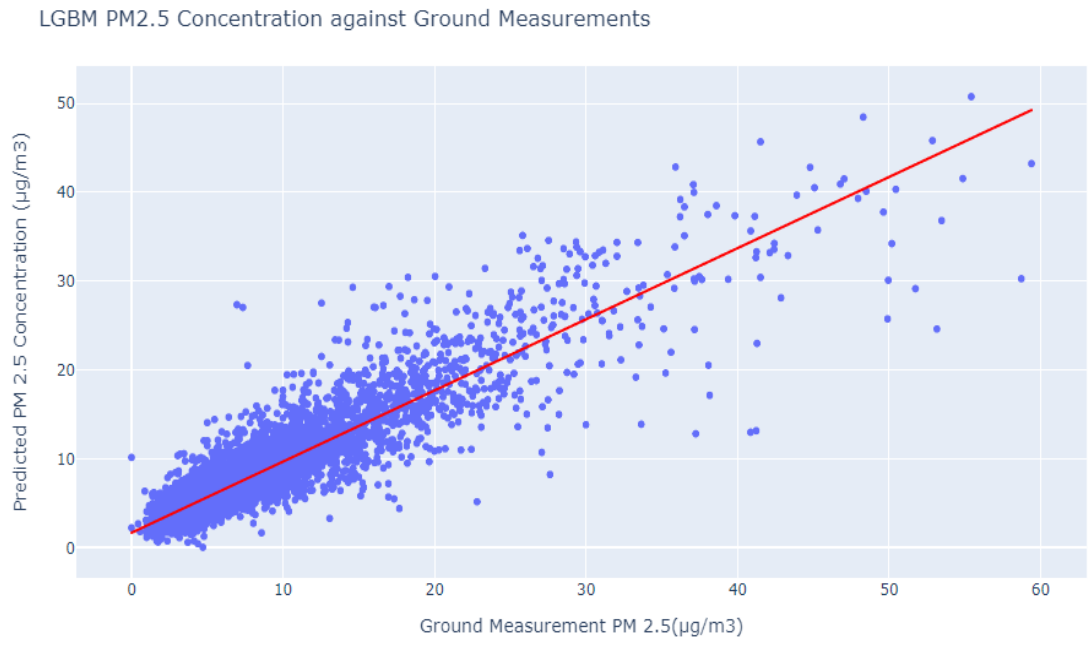

Figure 5: LGBM PM2.5 model performance

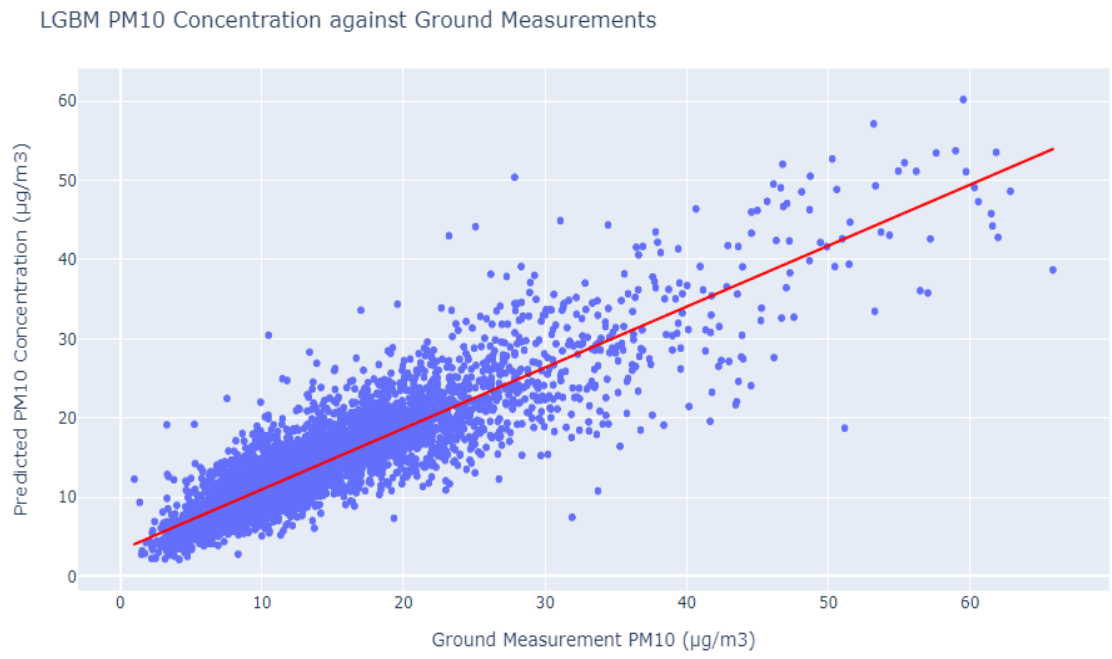

Figure 6: LGBM PM10 model performance

5.3 Model Temporal Variation

The temporal variation of the training set for the LightGBM model is seen in Figure 7 and 8 below, for $PM_{2.5}$ and PM_{10} respectively. The figures display the model's performance over a period of approximately 18 months. It can be evaluated that the model has performed adequately in this aspect, apart from the small number of outliers.

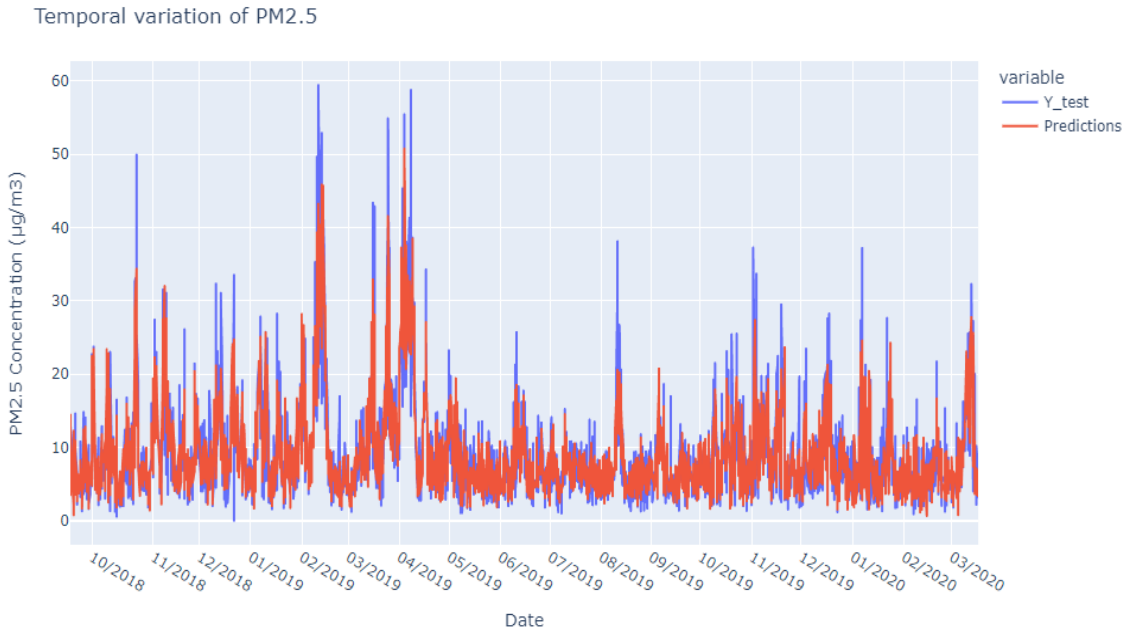

Figure 7: LGBM PM2.5 temporal variation

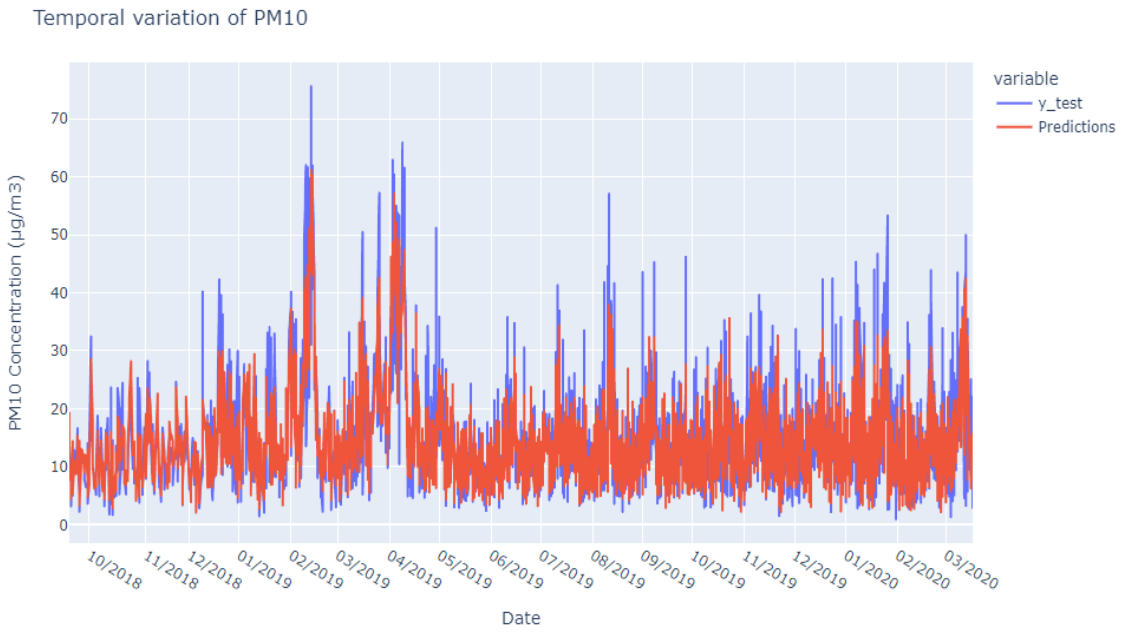

Figure 8: LGBM PM10 temporal variation

5.4 Feature Importance

The highest contributing factors in this model were the velocity components of wind, as shown in Figure 9. However, there is a relatively equal spread in terms of importance, and that the majority of the features contributed significantly to the model's performance.

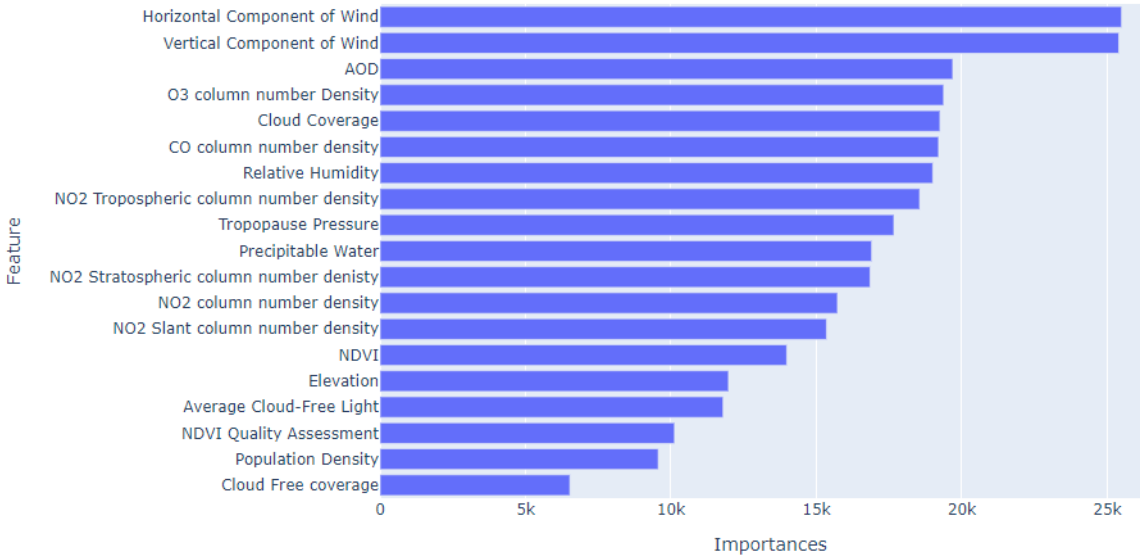

Figure 9: LGBM Feature importance

5.5 Correlation Metrics

Tables 5 and 6 below summarise the Pearson and Spearman correlation metrics between the concentrations of $PM_{2.5/10}$ from LightGBM and the number of hospital admissions for different respiratory diseases. The Pearson correlation is a dimensionless statistical parameter which analyses the linear variation of two features [18]. The Spearman correlation is a dimensionless measurement of the magnitude and direction between two variables [19]. A relatively higher correlation is seen between respiratory disease and COPD with Pearson correlation ranging between 0.238 – 0.283 and Spearman correlation range between 0.195 – 0.277. A lower correlation is seen with Asthma with values of 0.043 – 0.139 for Pearson correlation and 0.043 – 0.181 for Spearman.

Table 5: PM2.5 correlation metrics

Disease	Pearson Correlation	Spearman Correlation
Respiratory	0.238	0.195
Asthma	0.139	0.181
COPD	0.277	0.230

Table 6: PM10 correlation metrics

Disease	Pearson Correlation	Spearman Correlation
Respiratory	0.283	0.206
Asthma	0.043	0.043
COPD	0.268	0.277

5.6 Spatial Correlation

Figures 10 and 11 below represent the correlation analysis with respect to specific locations for respiratory disease with $PM_{2.5}$ and PM_{10} respectively. Despite the few locations which have a stronger correlation, it is evident that there is a lack of correlation between most of the locations. In Figure 10, there is a larger number of hospital admissions with respect to the concentration of $PM_{2.5}$. Conversely, in Figure 11 there is a relatively higher concentration of PM_{10} with respect to corresponding number of hospital admissions. PM_{10} appears to correlate overall better with respiratory admissions compared to $PM_{2.5}$.

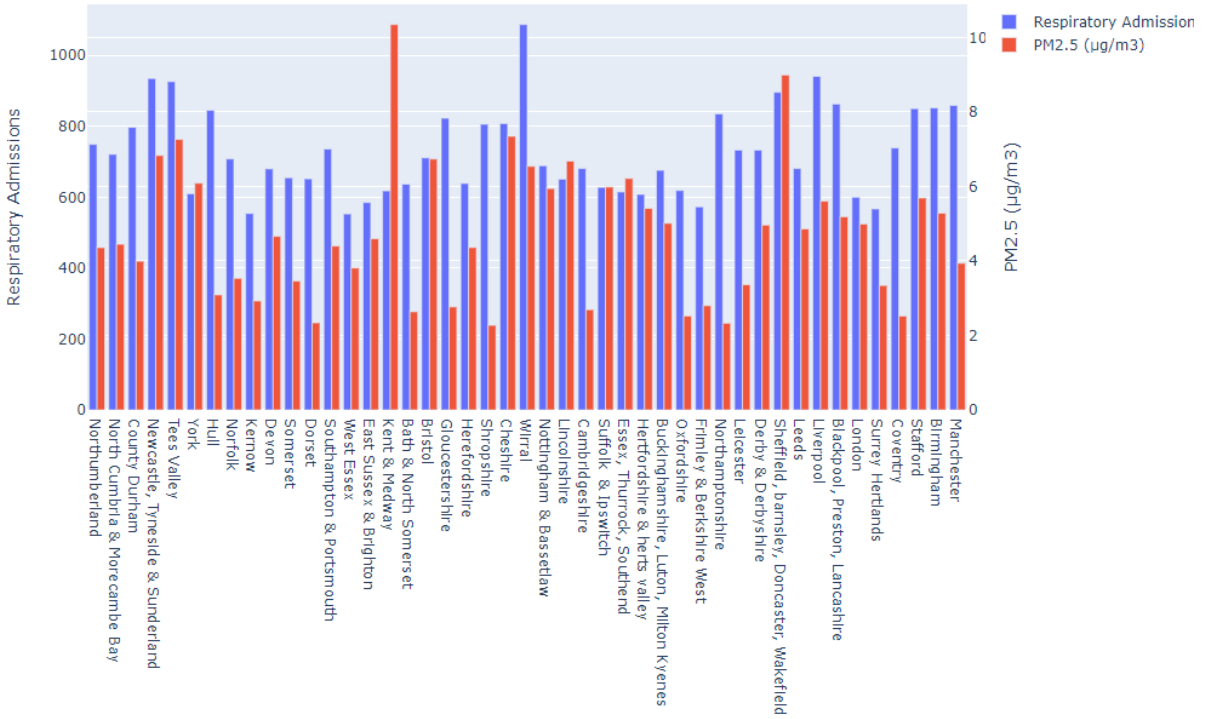

Figure 10: PM2.5 spatial correlation

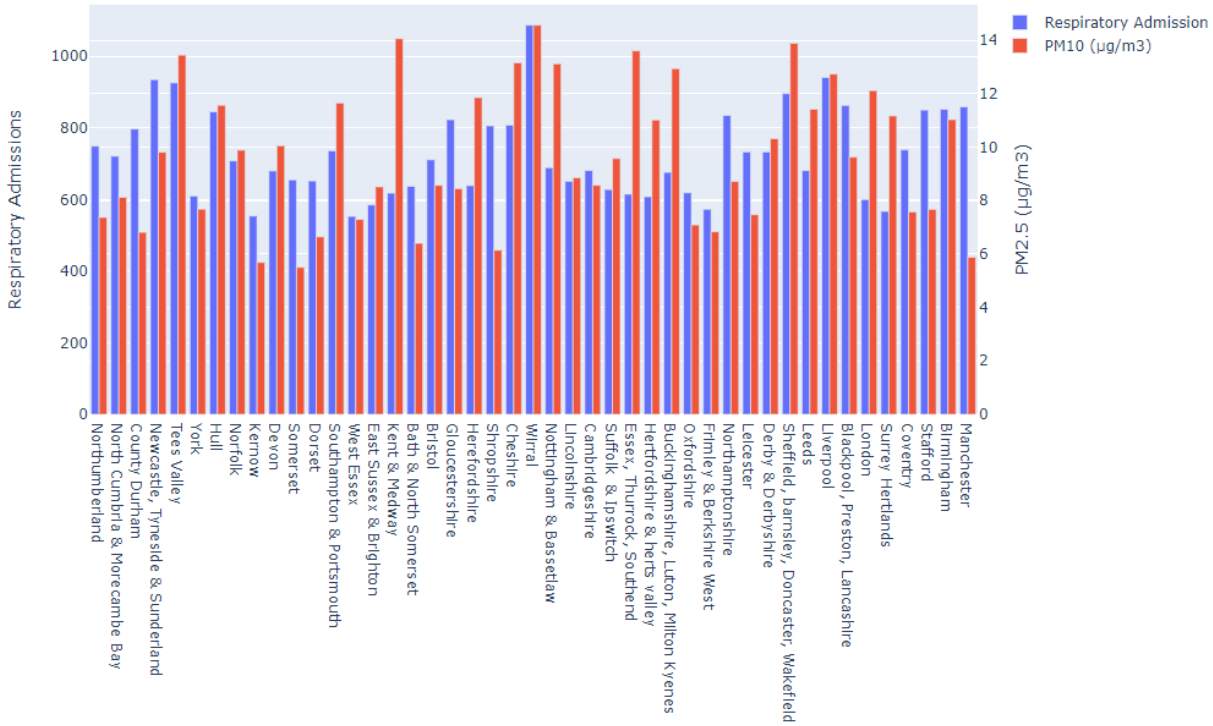

Figure 11: PM10 spatial correlation

6.0 Discussion

The standout prediction model in this analysis was LightGBM, which provided the most accurate representation of ground-based measurements for PM concentration. This model was also found to predict particulate matter concentrations which were the most highly correlated with the number of hospital admissions for each of the diseases compared to the other models. Despite the well-estimated predictions, the correlation between the concentrations and number of hospital admissions was not great. There is, at most, a weak positive correlation observed between the two. A more notable correlation was observed for respiratory disease and COPD based on the correlation metrics with a maximum value for Pearson and Spearman of 0.277. However, the correlation with Asthma was very weak, where 0.043 was a minimum value. In this aspect of the study, there are clearer limitations which have contributed to a weak correlation.

The time of interest for this study was partly during the COVID-19 pandemic, where during the year 2020 there was a significant change in the level of air pollution in the UK [20]. The models were trained mainly in the time prior to COVID-19, however, were tested during the pandemic. This anomaly may have been a key contributing factor in the weak correlation results.

As previously mentioned, the data collected for the number of hospital admissions is sourced from NHS Digital. The available data is an annual value for each CCG, which may not have been the best for use in this analysis. One value to represent a year of data does not provide as much information as a monthly report. For example, had there been a monthly value, it may have provided a further insight as to which periods in the year the correlation was higher.

Additionally, this analysis was limited due to the number of CCGs available on the NHS Digital page. There are approximately 100 CCGs, where some cover a far larger population than others, which may not be the most suitable in this circumstance. The data is averaged for each 100,000 people, which is also a source of possible error. A way to improve this would of course be to have a wider spread of CCGs across England which would allow for a more varied and more reliable representation of the true values.

To simplify the methodology, a select number of coordinates were collected from each region for the test data. However, this simplification may have led to the low correlation results. A method which can collect coordinates for each pixel in a certain region may have proved to return a greater correlation than that of the procedure in this study, due to the larger spread and representation of a true value.

7.0 Conclusion and Future work

In this study a prediction model for the level of concentration of $PM_{2.5}$ and PM_{10} was developed, and the correlation between these values and the number of hospital admissions for different diseases was analysed. This was done using Sentinel-5P and meteorological features as inputs in machine learning models. The final model

(LightGBM) used was able to provide an accurate estimation of PM in the UK as compared to the ground measurements. The results of the RMSE and R^2 were promising and can be a viable way of estimating particulate matters. However, the weak Pearson and Spearman correlations for different respiratory diseases proves that there are still limitations to this analysis. These limitations include the lack of wider spatial and temporal NHS data, the possible confusion of the model due to COVID-19 pandemic which saw a significant change in the amount of air pollution and assumptions made in the methodology.

To further develop this analysis, a study could include a process of gathering coordinates for every pixel in a region for a more accurate representation of the true PM concentration, as opposed to a select number of chosen coordinates. Additionally, the time of interest may be adjusted to a time prior to COVID-19 or sometime after to avoid potential anomalies in results.

8.0 References

[1]((Dominici, F., Peng, R.D., Bell, M.L., Pham, L., McDermott, A., Zeger, S.L. and Samet, J.M. (2006). Fine Particulate Air Pollution and Hospital Admission for Cardiovascular and Respiratory Diseases. JAMA, 295(10)

[2] World Health Organization (2019). Air pollution. [online] Who.int. Available at: https://www.who.int/health-topics/air-pollution#tab=tab_

[3] World Health Organization (2022). WHO COVID-19 dashboard. [online] World Health Organization. Available at: https://covid19.who.int/.

[4] GOV.UK. (2021). Concentrations of particulate matter (PM10 and PM2.5). [online] Available at: https://www.gov.uk/government/statistics/air-quality-statistics/concentrations-of-particulate-matter-pm10-and-pm25

[5] Department for Environment, F. and R.A. (Defra) webmaster@defra gsi gov uk (n.d.). Search for monitoring sites- Defra, UK. [online] uk-air.defra.gov.uk. Available at: https://uk-air.defra.gov.uk/networks/find-sites [Accessed 14 Mar. 2022].

[6] May, S. (2011). What Is a Satellite? [online] NASA. Available at: https://www.nasa.gov/audience/forstudents/5-8/features/nasa-knows/what-is-a-satellite-58.html

[7] Brown, S. (2021). *Machine learning, explained*. [online] MIT Sloan. Available at:

https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained.

[8] Chen, G., Li, S., Knibbs, L.D., Hamm, N.A.S., Cao, W., Li, T., Guo, J., Ren, H., Abramson, M.J. and Guo, Y. (2018). A machine learning method to estimate PM2.5 concentrations across China with remote sensing, meteorological and land use information. Science of The Total Environment, [online] 636, pp.52–60. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0048969718314281 [Accessed 14 Nov. 2021].

[9] Xiao, Q., Wang, Y., Chang, H.H., Meng, X., Geng, G., Lyapustin, A. and Liu, Y. (2017). Full-coverage high-resolution daily PM2.5 estimation using MAIAC AOD in the Yangtze River Delta of China. Remote Sensing of Environment, [online] 199, pp.437–446. Available at: https://www.sciencedirect.com/science/article/pii/S0034425717303371?casa_token=J7QPH4vBd8YAAAAA:7KqMh2u7FtcBieIyWWvfz_-fWwyQpaz3mNW5ifHRgpq0aEYx-wwjDOR8NQhP961AaNhA8kdYebs.

[10] Ogawa Kawano, A. (2021). [Poster] Association between prenatal exposure to ambient particulate matter (PM2.5) and adverse birth outcomes in Nepal and Bangladesh using the prediction model for daily PM2.5 concentrations matched with demographic and health survey data: spatio-temporal analysis. ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS).

[11] OpenAQ. (n.d.). OpenAQ. [online] Available at: https://openaq.org/#/.

[12] fingertips.phe.org.uk. (n.d.). Inhale - INteractive Health Atlas of Lung conditions in England - Data - OHID. [online] Available at: https://fingertips.phe.org.uk/profile/inhale/data

[13] Google Developers. (n.d.). Earth Engine Data Catalog. [online] Available at: https://developers.google.com/earth-engine/datasets

[14] sentinel.esa.int. (n.d.). Sentinel-5P - Missions - Sentinel Online - Sentinel. [online] Available at: https://sentinel.esa.int/web/sentinel/missions/sentinel-5p.

[15] US Department of Commerce, N. (n.d.). National Centers for Environmental Prediction. [online] www.weather.gov. Available at: https://www.weather.gov/ncep

[16] GISGeography (2018). *What Is NDVI (Normalized Difference Vegetation Index)? - GIS Geography*. [online] GIS Geography. Available at: https://gisgeography.com/ndvi-normalized-difference-vegetation-index/.

[17] DataTechNotes (n.d.). Regression Accuracy Check in Python (MAE, MSE, RMSE, R-Squared). [online] Available at: https://www.datatechnotes.com/2019/10/accuracy-check-in-python-mae-mse-rmse-r.html

[18] Ramzai, J. (2020). Clearly explained: Pearson V/S Spearman Correlation Coefficient. Medium. [online] 25 Jun. Available at: https://towardsdatascience.com/clearly-explained-pearson-v-s-spearman-correlation-coefficient-ada2f473b8.

[19] Laerd Statistics (2018). Spearman's Rank-Order Correlation - A guide to when to use it, what it does and what the assumptions are. [online] Laerd.com. Available at: https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php.

[20] AIR QUALITY EXPERT GROUP Estimation of changes in air pollution emissions, concentrations and exposure during the COVID-19 outbreak in the UK. Rapid evidence review -June 2020. (n.d.). [online] Available at: https://uk-air.defra.gov.uk/assets/documents/reports/cat09/2007010844_Estimation_of_Changes_in_Air_Pollution_During_COVID-19_outbreak_in_the_UK.pdf.