Forecasting hourly NO_2 concentrations in Manchester city centre using XGBoost with kNN-based feature selection

A dissertation submitted to the University of Manchester for the degree of Data Science in the Faculty of Humanities

STUDENT ID NUMBER

11155827

YEAR OF SUBMISSION

2023

CANDIDATE'S SCHOOL

School of Social Sciences

LIST OF CONTENTS

Word count: 6513

The code can be found here.

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Nitrogen dioxide (NO_2) is one of the major atmospheric pollutants, causing negative effects on humans and the environment. This study aimed to train and test prediction models for NO_2 concentration in Manchester Piccadilly, using XGBoost with kNN-based feature selection. Meteorological data from the AURN and traffic volume data from TfGM were used for prediction. The period of each dataset is from 1 Jan 2016 to 31 Dec 2020. The five versions of kNN-based feature selection methods for the traffic dataset were used: 1) kNN1 by distance from Piccadilly monitoring site, 2) kNN2 by distance only for sites with the angular difference from wind direction (θ) < 90°, 3) kNN3 according to distance adjusted by θ, 4) kNN4: Adjusted kNN3 model with more emphasis on θ, and 5) kNN5: Adjusted kNN3 model that performs manual imputation. The models are again trained and tested by changing the forecasting period (T+1, T+6, T+24, and T+48) and the time-series period (the Entire period, pre-lockdown period, and 1st lockdown period). It was found that within a small range of 3km, wind direction plays only a limited role in predicting NO_2. If we are going to use only a small number of features (K=5), using wind direction for feature selection can be effective. Feature selection did not lead to higher model accuracy, only meaningful in selecting a smaller number of features while maintaining similar accuracy. Meanwhile, the lockdown period model showed higher accuracy than the pre-lockdown model. Also, the pre-lockdown model failed to make accurate predictions for the lockdown period. This may be due to the changes in the contribution rates by sources of NO_2 during the lockdown period. The greater impact of COVID-19 was on other industries than the transport sector, leading to relatively increased traffic contribution to NO_2. Further research needs to be conducted with a bigger study area, more features, regions, and algorithms.

DECLARATION

No portion of the work referred to in the dissertation has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

INTELLECTUAL PROPERTY STATEMENT

i. The author of this dissertation (including any appendices and/or schedules to this dissertation) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this dissertation, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has entered into. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trademarks, and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the dissertation, for example graphs and tables (“Reproductions”), which may be MSc Data Science 18 described in this dissertation, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this dissertation, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see

https://documents.manchester.ac.uk/display.aspx?DocID=24420), in any relevant Dissertation restriction declarations deposited in the University Library, The University Library's regulations (see https://www.library.manchester.ac.uk/about/regulations/) and in The University's Guidance for the Presentation of Dissertations.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my dissertation supervisor Prof. David Topping, for his continuous support and insights. His guidance and advice carried me through all the stages of my dissertation project. Also, I would like to thank my parents for their love and support.

1. INTRODUCTION

1.1. Background

Nitrogen dioxide (NO_2) is one of the major atmospheric pollutants, reported to be responsible for 64,000 premature deaths in 2020 in Europe (EEA, 2022). Asthma, chronic obstructive pulmonary disease, lung cancer, and decreased lung function growth in children are the well-known negative effects of NO_2 exposure (Hamra et al., 2015 cited in Cooper et al., 2022; ALA, 2023; CARB, 2023; EPA, 2023a). NO_2 can also cause environment damage through rain, such as changing nutrient levels and affecting species diversity (DEFRA, 2023).

Global efforts are being made to improve air quality and reduce air pollution. For example, Environmental Protection Agency (EPA) in US sets and revises the national ambient air quality standards (NAAQS) and take regulatory actions as an official government organisation (EPA, 2023b). The UK is also implementing national efforts, investing over £2.7 billion overall in air quality and transport alongside regulatory actions (DEFRA, 2017). In this context, forecasting air quality or pollutant concentrations has a more important meaning for more effective policy design and implementation.

There have been many studies to forecast air pollutants using machine learning (ML) (Méndez, Merayo and Núñez, 2023). ML models such as artificial neural network (ANN) and extreme learning machine (ELM) can outperform statistical and chemical transport approaches in terms of forecasting time and accuracy (Zaini et al., 2022). Especially, XGBoost in recent studies has been reported to outperform other machine learning algorithms such as Random Forest (RF), Support Vector Machine (SVM), Convolutional Neural Networks (CNN), and Deep Neural Networks (DNN) (Pan, 2018; Wong et al., 2021; Ayus, Natarajan and Gupta, 2023).

1.2. Aim and Objectives

On this basis, this study aims to train and test prediction models for NO_2 concentration in Manchester Piccadilly, using XGBoost with kNN-based feature selection. Meteorological data from the Automatic Urban and Rural Network (AURN) and traffic volume data from Transport for Greater Manchester (TfGM) will be used

for prediction. The period of each dataset is from 1 Jan 2016 to 31 Dec 2020.

Depending on the feature selection method or the prediction period, various versions of models will be made and compared with each other. Below are the objectives of this study.

1) To train and test various XGBoost models by changing kNN-based feature selection methods for the traffic dataset as follows:
    - kNN1: kNN feature selection by distance from Piccadilly monitoring site
    - kNN2: kNN feature selection by distance only for sites with the angular difference from wind direction (θ) < 90°
    - kNN3: kNN feature selection according to distance adjusted by θ
    - kNN4: Adjusted kNN3 model with more emphasis on θ
    - kNN5: Adjusted kNN3 model that performs manual imputation
2) To train and test models from 1) again, by changing the forecasting period in hours as follows: T+1, T+6, T+24, and T+48
3) To train and test models from 2) again, by changing the time series period as below:
    - Entire period: 1st Jan 2016 – 31st Dec 2020
    - Pre-lockdown period: 1st Jan 2016 – 22nd Mar 2020
    - 1st Lockdown period: 23rd Mar 2020 – 1st Jun 2020 (UK Parliament, 2021)

The difference between the models from kNN1 to kNN4 lies in how wind direction is incorporated into the feature selection process. One of the advantages of XGBoost is its capability to handle missing values (Bentéjac, Csörgő and Martínez-Muñoz, 2020), and comparing kNN3 and kNN5 is to check it in practice. The details of each feature selection method and XGBoost algorithm will be discussed in Chapter 3.

2. RELATED WORKS

2.1. Factors associated with NO_2 concentration

2.1.1. Traffic volume and other factors

Identifying the primary sources of NO_2 is crucial to understanding the relationship between traffic volume and NO_2 concentration. In general, the sources of NO_2 can be classified into two categories: 1) natural sources, and 2) anthropogenic (human-produced) sources (NOAA, 2010; MfE, 2021). Natural sources include forest fires, bacteria and volcanoes, but the primary source of NO_2 is human-produced sources (NOAA, 2010; MfE, 2021; EPA 2023). Human-produced sources can be divided into two categories, indoor sources and outdoor sources (WHO, 2010). Indoor sources include combustion appliances and tobacco smoking (Vilčekova, 2010; MfE, 2021; EPA, 2023a). Transport, combustion industry, and power plants are outdoor sources of NO_2 (DEFRA, 2022).

Figure 1 shows the annual mean concentration of NO_2 in the UK from 1990 to 2022 (DEFRA, 2022). We can see that NO_2 is in an overall decreasing trend, which can be seen as a result of technological progress or national efforts as mentioned above. Regarding this study, it is important to see the gap between “Roadside” and “Urban Background” emissions. This has implications for the diffusion mechanism of NO_2 from roadsides into urban areas and the influence of other sources on NO_2.

In Figure 2, we can directly check the contribution of each source to Nitrogen oxides (NOx) in the UK (NAEI, 2022 cited in DEFRA, 2022). NOx is a term that includes nitric oxide (NO) and NO_2, and NO reacts with oxygen or ozone to form NO_2 (NAEI, 2022 cited in DEFRA, 2022). It is notable that emission from ‘Passenger cars’ and 'heavy duty vehicles’ took up less than 20% of the total emission in 2020 (NAEI, 2022). This implies the limitation of forecasting NO_2 concentration only by traffic volume. We will discuss this later in detail with model results in chapter 5.

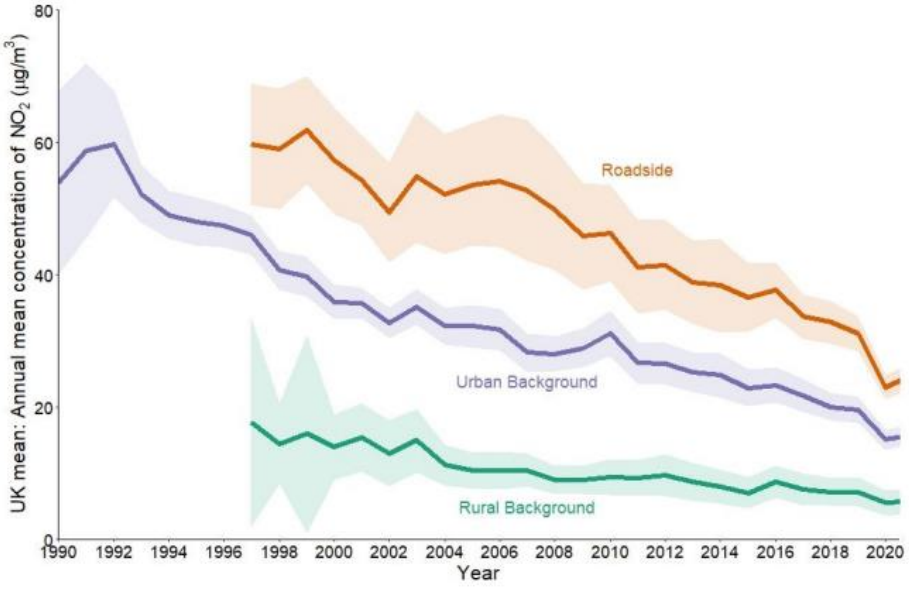

Figure 1. Annual mean concentration of NO_2 in the UK, 1990-2021 (DEFRA, 2022)

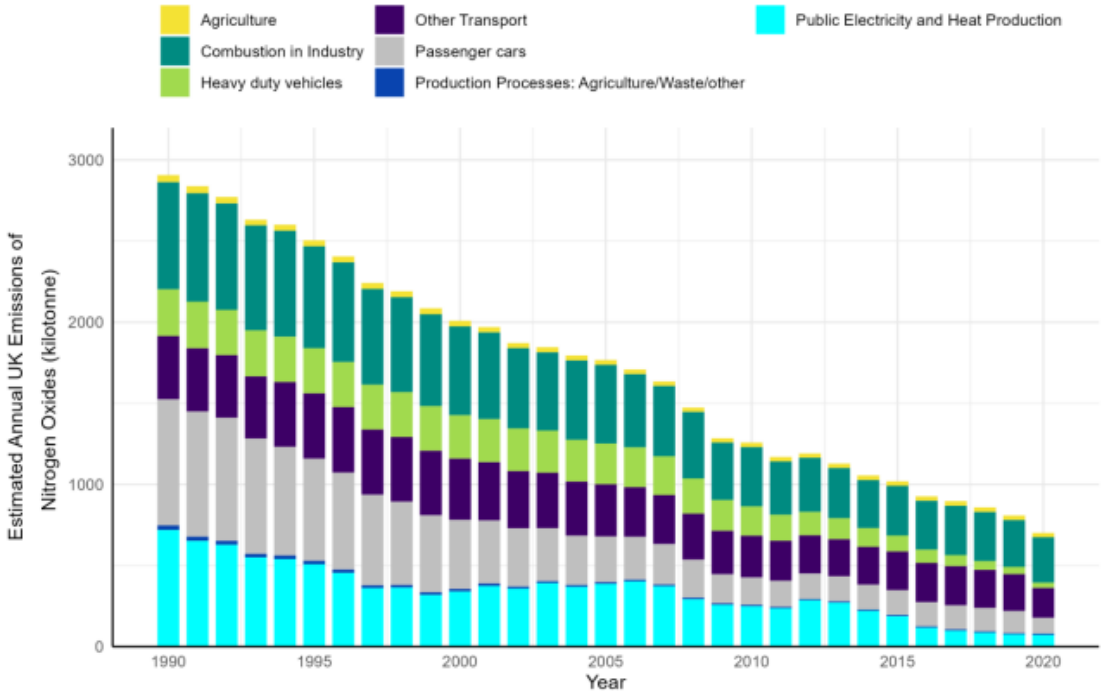

Figure 2. Annual emissions of NO_x by sectors in the UK, 1990-2021 (NAEI, 2022 cited in DEFRA, 2022)

2.1.2. Meteorological factors

It is important to understand the relationship between meteorological factors and NO_2 concentration because the forecasting model in this study is designed to use both meteorological data and traffic data to predict NO_2 concentration.

Meteorological factors play an important role in the dispersion and accumulation of NO_2, even on s hundreds of kilometres scale (DEFRA, 2004). Voiculescu et al. (2020) summarised the results of studies on the relationship between meteorological factors (wind, temperature, and wind speed) and NO_2 from various locations, as shown in Table 1 below. It shows that NO_2 concentration is negatively correlated with temperature, humidity, and wind speed in general.

Table 1. Correlations between NO_2 and meteorological factors for various locations (Voiculescu et al., 2020)

Location	Temperature	Humidity	Wind Speed	Source
Egypt, Cairo	Insignificant	Negative	Negative	Elminir (2005)
India, Surat	Insignificant	-	Negative	Verma and Desai (2008)
India, Jabalpur	Negative	Negative	-	Srivastava, Sarkar and Beig (2014)
Saudi Arabia, Makkah	Insignificant	Negative	Negative	Habeebullah et al. (2015)
Saudi Arabia, Dhahran	Negative	Positive	Negative	Gasmi et al. (2017)
China, Beijing	Negative	Positive	Negative	Zhang et al. (2019)
China, Shanghai	Negative	Negative	Negative	Zhang et al. (2019)
China, Guangzhou	Positive	Negative	Negative	Zhang et al. (2019)
Malaysia Kuala Lumpur	Positive	Insignificant	Negative	Zhang et al. (2019)
Iran, Isfahan	Negative	Negative	Negative	Dunlea et al. (2007)

Wind direction is considered to play an important role in air pollutant concentration on a local scale (Kim et al., 2015). Because of the characteristics of wind direction and air pollutant dispersion mechanism, how to incorporate wind direction into the prediction model will be discussed in Section 2.2.2 below.

Meanwhile, the study by Liu et al. (2020) showed that precipitation and wind have a scavenging effect on PM2.5 and PM10, lowering the concentrations. According to R

Kalbarczyk and E Kalbarczyk (2007), atmospheric pressure showed a positive correlation with NO_2.

2.2. Air pollutant forecasting models

2.2.1. XGBoost for prediction

Research on air quality prediction models using deep learning has been actively conducted in recent years (Tao et al., 2019). Long short-term memory neural network (LSTM) is one of the commonly used methods, as a state-of-the-art model of RNN (Liu et al., 2019; Tao et al., 2019). As can be inferred from its name, LSTM specialises in learning long-term trends from time-series data (Bouktif et al., 2020 cited in Drewil and Al-Bahadili, 2022).

Meanwhile, XGBoost is often found to be efficient for air pollutant forecasting in recent studies such as Pan (2018), Ayus, Natarajan and Gupta (2023), and Wong et al. (2021). Pan (2018) showed XGBoost algorithm outperforms RF, Multiple Linear Regression (MLR), decision tree, and SVM for hourly PM2.5 prediction. According to Ayus, Natarajan and Gupta (2023), XGBoost outperforms hybrid deep learning models such as CNN-BiLSTM, Bi-LSTM, CNN-BiLSTM, and Conv1D-BiLSTM for air quality index (AQI) prediction. For NO_2 prediction, Wong et al. (2021) showed that XGBoost outperformed RF and DNN.

In this context, this study uses the XGBoost algorithm for NO_2 concentration prediction. The details and advantages of XGBoost will be discussed in Chapter 3.

2.2.2. Incorporating wind direction into prediction models

As mentioned above, it is important to consider how to incorporate wind direction into our prediction model. The first notable approach can be found in Teng et al. (2023). For 72-hour real time forcasting of PM2.5 by DNN with aggregated neighbourhood spatiotemporal information, they calculated the angle (θ) between the wind direction and the site-to-site direction as below:

$$\theta = Abs\left(U - \left(arcsin\frac{x_i - x_j}{d_{ij}}\right)\right)$$

U: wind direction

d_{ij}: Euclidean distance between sites i and j

For any sites having θ > 90°, they put weight contribution equal to 0 as it implies the wind blows in the opposite direction from pollutant transport (Teng et al., 2023).

Yang, Fan and Zhao (2019) suggests another approach for PM2.5 prediction, Dynamic Wind Field Distance (DWFD). To predict PM2.5 concentration in Beijing using LSTM-CNN, they adjust the distance between sites i and j by the angle (θ_{ij}) between wind direction and the line between the two sites. The equation below shows the idea of DWFD (Yang, Fan and Zhao, 2019):

$$d_{ij_{wind}}(t) = \begin{cases} d_{ij}exp\left(-\frac{\left(\sin\theta_{ij}(t) - 1\right)^2}{2\sigma^2}\right), & if\ \theta_{ij}(t) \leq 90° \\ d_{ij}, & if\ 90° \leq \theta_{ij}(t) \leq 180° \end{cases}$$

d_{ij}: Geographical distance between sites i and j

It is noteworthy that they used Gaussian kernel to give a larger adjustment at lower angles and to allow the degree of adjustment to be controlled by σ. Figure 3 shows the idea of DWFD and the variation in the degree of adjustment according to each σ value.

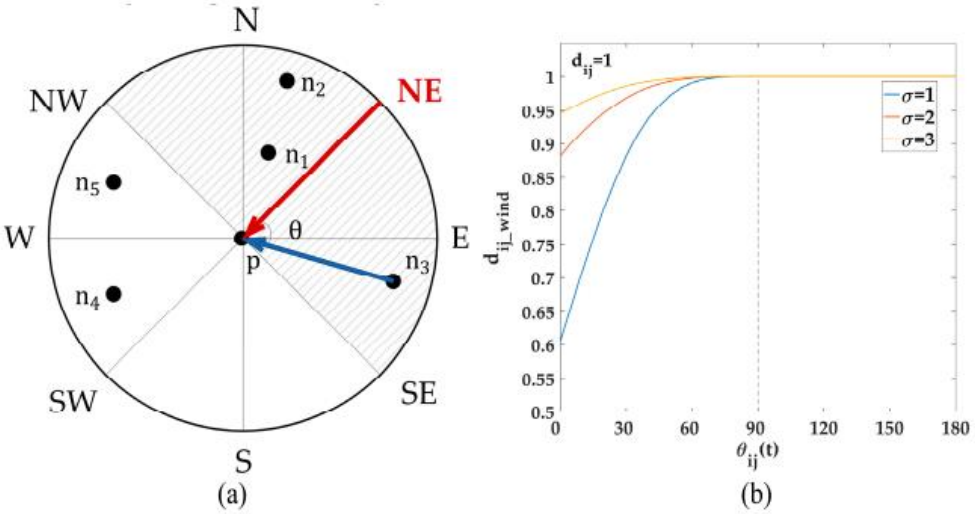

Figure 3. Illustration of the kNN-DWFD method (Yang, Fan and Zhao, 2019)

Another important thing is how they treated sites with θ > 90°. Instead of giving 0 weight to those sites as Teng et al. (2023) did, they chose to give them no adjustment in distance. Then, they did kNN-based feature selection by comparing the distances.

These two studies provide an idea of one of the key aspects of the model design in our study: the five versions of kNN-based feature selection. The first version uses the original distances for all the traffic sites regardless of wind direction, and the second version is based on the idea from Teng et al. (2023). The other three versions use adjusted distances, but the adjustment method is different from DWFD by Yang, Fan and Zhao (2019) in that it uses Euclidean distance and excludes sites with θ > 90°. The details of model design will be dealt with in Section 3.

3. METHODOLOGY

3.1. Description of the dataset

3.1.1. Meteorological data

The meteorological data of NO_2 (µg/m3), wind direction, wind speed (m/s), and temperature (°C) is recorded at Manchester Piccadilly Automatic Urban Monitoring Network (AURN) site (UKA00248), of which information can be found on the Department for Environment Food & Rural Affairs (DEFRA) webpage (https://uk-air.defra.gov.uk/networks/site-info?site_id=MAN3). The latitude of the site is 53.481520, and the longitude is -2.237881. The dataset consists of 24-hour hourly records from 1st Jan 2016 to 31st Dec 2020.

The hourly data of humidity (%), precipitation (mm), and air pressure (mb) for the same site and period as the AURN data is provided by Visual Crossing. They provide historical weather database service by processing the weather data from worldwide sources such as Meteorological Assimilation Data Ingest System (MADIS), Global Historical Climate Network Daily (GHCN-D), and Deutscher Wetterdienst (DWD). More information can be found on their webpage (https://www.visualcrossing.com/resources/documentation/weather-data/weather-data-sources-and-attribution/).

3.1.2. Traffic data

The hourly traffic data for this study is provided by Transport for Greater Manchester (TfGM). It consists of 42 Automatic Traffic Count (ATC) sites record within a 3km distance from the Piccadilly monitoring station. The ATC sites are strategically located across the Key Route Network (KRN), and the data is recorded using Induction Loop Detectors (ILDs). The period is from 1st Jan 2016 to 31st Dec 2020. Figure 4 shows the Piccadilly AURN site and the 42 traffic record points on the map.

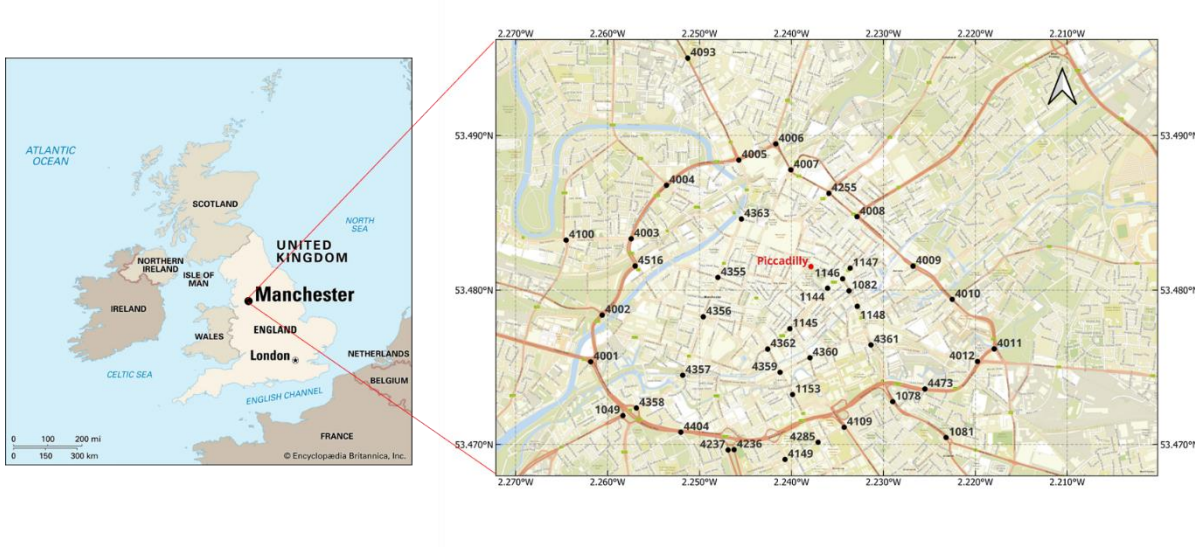

Figure 4. The location of the Piccadilly monitoring station and the traffic sites

3.2. kNN-based feature selection

In this study, there are five versions of the kNN-based feature selection approach where wind direction and distance play significant role. The concept of each variation is as below.

kNN1: Choosing the k nearest traffic sites based on distances (d) from the Piccadilly monitoring site.

kNN2: Choosing the k nearest traffic sites with the angle (θ) from the wind direction (U) is less than or equal to 90 degrees, based on distances (d) from the Piccadilly monitoring site.

kNN3: Choosing the k nearest traffic sites with θ from U is less than or equal to 90 degrees, based on adjusted distances (d′) by θ.

kNN4: The basic is the same as kNN3, but wind direction has more influence on distance adjustment.

kNN5: The basic is the same as kNN3 but it chooses only traffic sites with non-null values.

The distance (d) between the Piccadilly monitoring site and each traffic site is

calculated by the Euclidean distance measure:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

The angle (θ) between the wind direction (U) and the edge between the Piccadilly monitoring station and each traffic site for kNN2~kNN5 is calculated as follows:

$$Let\ \theta^* = Abs\left(U - atan2\left((x_i - x_j), (y_i - y_j)\right) * \frac{180}{\pi}\ \%\ 360\right),$$

$$\theta = \begin{cases} 360 - \theta^*, & \theta^* > 180 \\ \theta^*, & otherwise. \end{cases}$$

The adjusted difference (d′) for kNN3~kNN5 is calculated using Gaussian kernel.

$$d' = d * exp\left(-\frac{(\sin\theta - 1)^2}{2\sigma^2}\right), \qquad \theta \leq 90^{\circ}$$

For kNN3 and kNN5, d' is calculated with σ = 2, and σ = 1 for kNN4. As mentioned by Yang, Fan, and Zhao (2019) in Section 2.2, decreasing the value of σ leads to wind direction having more influence on the adjusted distance. Therefore, comparing kNN3 and kNN4 is to see whether putting more emphasis on wind direction can improve the model performance. The detailed algorithms for kNN1~kNN5 can be represented as follows.

As will be mentioned in the following Section 3.3, one of the advantages of XGBoost is its capability of handling missing values. kNN5 is designed to do manual imputation by choosing the next non-null value traffic sites; therefore, comparing the performance of kNN3 and kNN5 is to confirm the advantage in practice.

Below are the detailed algorithms for each model kNN1~kNN5.

kNN1 algorithm

Step 1. Calculate distance (d) between the Piccadilly monitoring site and traffic sites;

Step 2. Sort traffic sites in ascending order by distance (d);

Step 3. Select k nearest traffic sites from Step 2;

kNN2 algorithm

Step 1. The same as kNN1;

Step 2. For each hourly record sample, calculate the angle (θ) between the wind direction and the edge between the Piccadilly monitoring site and each traffic site;

Step 3. For each hourly record sample, sort traffic sites with θ ≤ 90° in ascending order by distance (d);

Step 4. For each hourly record sample, select k nearest traffic sites from Step 3;

kNN3 algorithm

Step 1~2: The same as kNN2;

Step 3. For each hourly record sample, sort traffic sites with θ ≤ 90° in ascending order by adjusted distance (d′) calculated with σ = 2;

Step 4. For each hourly record sample, select k nearest traffic sites from Step 3;

kNN4 algorithm

Step 1~2: The same as kNN2;

Step 3. For each hourly record sample, sort traffic sites with θ ≤ 90° in ascending order by adjusted distance (d′) calculated with σ = 1;

Step 4. For each hourly record sample, select k nearest traffic sites from Step 3;

kNN5 algorithm

Step 1~2: The same as kNN2;

Step 3. For each hourly record sample, sort traffic sites with θ ≤ 90° and no null traffic record value in ascending order by adjusted distance (d′) calculated with σ = 2;

Step 4. For each hourly record sample, select k nearest traffic sites from Step 3;

3.3. XGBoost regression

In this study, XGBoost will be used for NO_2 prediction using the features selected by the kNN-based feature selection method. XGBoost is a scalable, distributed gradient-boosting tree system (Chen and Guestrin, 2016). The name "XGBoost" stands for Extreme Gradient Boosting in that meaning (Pan, 2018). "Boosting" refers to an ensemble method that generates and accumulates weak learners into a collectively strong model (Pan, 2018, Price et al., 2022). And when the steps of generating weak models are on a gradient descent algorithm, we call it "Gradient Boosting" (Pan, 2018). The idea of gradient boosting can be represented as follows (Bentéjac, Csörgő and Martínez-Muñoz, 2020):

Building an approximation of F(x) as a weighted sum of functions,*

$$F_m(x) = F_{m-1}(x) + \rho_m h_m(x)$$

where $\boldsymbol{\rho_m}$ *is the weight of the m^th function* $\boldsymbol{h_m(x)}$*. F*(x) can be obtained as*

$$F_0(X) = arg\min_{\alpha} \sum_{i=1}^{N} L(y_i, \alpha)$$

It can be written using subsequent models as

$$\big(\rho_m, h_m(x)\big) = arg\min_{\rho,h} \sum_{i=1}^{N} L\big(y_i, F_{m-1}(x_i) + \rho h(x_i)\big)$$

Instead of solving the equations above directly, based on the gradient algorithm, each model $\boldsymbol{h_m}$ *is trained on a new dataset* $\boldsymbol{D = \{x_i, r_{mi}\}_{i=1}^{N}}$*, where the pseudo-*

residuals $\boldsymbol{r_{mi}}$ *are calculated as follows:*

$$r_{mi} = \left[\frac{\partial L\left(y_i, F(x)\right)}{\partial F(x)}\right]_{F(x)=F_{m-1}(x)}$$

And $\boldsymbol{\rho_m}$ *is computed along each step by solving the optimisation problem.*

XGBoost is called an "Extreme" algorithm because of the way it works by pushing the limit of computational resources (Chen, 2016 cited in Brownlee, 2021). The two main advantages of XGBoost are its speed and performance (Chen and Guestrin, 2016; Trenchevski et al., 2020; Brownlee, 2021). XGBoost is considered the 'next-generation' machine learning algorithm (Quinto, 2020), and it was the most popular algorithm among the winning solutions for the Kaggle competition in 2015 (Chen and Guestrin, 2016). Even all the top-10 winning teams in KDDCup 2015 were using XGBoost (Chen and Guestrin, 2016). The difference between XGBoost and other gradient boosting algorithms is its scalability, parallel tree learning by sparsity-aware algorithm, and out-of-core tree learning by cache-aware block structure (Chen and Guestrin, 2016).

The loss function of XGBoost can be written as follows (Chen and Guestrin, 2016 cited in Bentéjac, Csörgő and Martínez-Muñoz, 2020; Zou et al., 2022):

$$L_{xgb} = \sum_{i=1}^{N} L(y_i F(x_i)) + \sum_{m=1}^{M} \Omega(h_m)$$

$$\Omega(h_m) = \gamma T + \frac{1}{2}\lambda[\![w]\!]^2:$$

$\boldsymbol{\Omega}$*: the regularisation item to determine the complexity of the trees*

T: The number of leaves of the tree

w: The predicted values of the leaf nodes

$\boldsymbol{\lambda, \gamma}$*: Controlling factors to avoid overfitting*

Figure 5 below shows the basic structure of the XGBoost regression model (Zou et al., 2022).

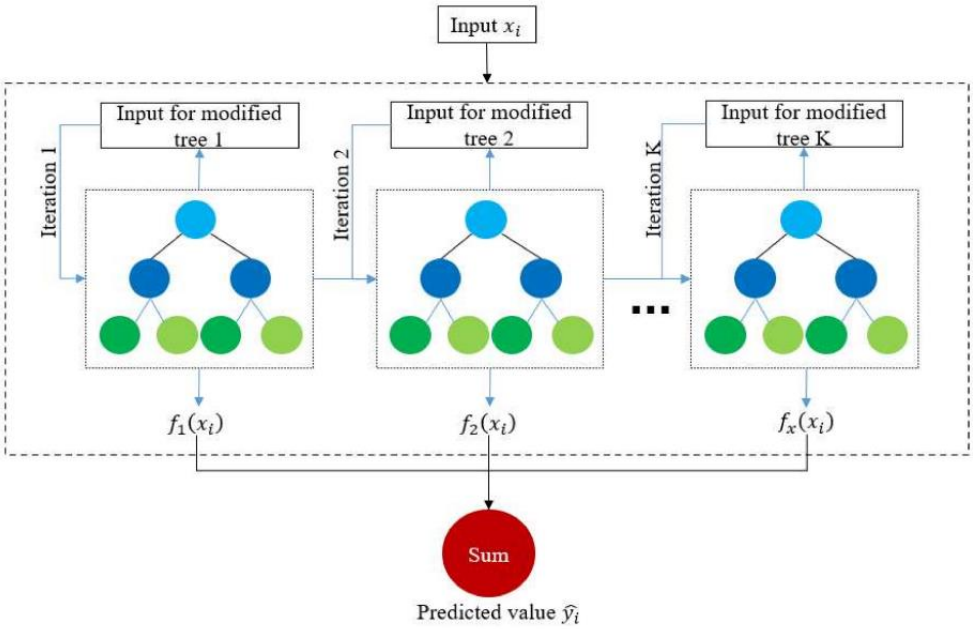

Figure 5. The basic structure of the XGBoost model (Zou et al., 2022)

XGBoost is insensitive to outliers due to its tree-based nature and provides effective ways to avoid overfitting (Zhang et al., 2019). Furthermore, the sparsity-aware algorithm leads to XGBoost's capability of handling missing values (Bentéjac, Csörgő and Martínez-Muñoz, 2020). We will compare kNN3 and kNN5 to see this advantage in practice.

3.4. Training and testing

For every possible case by the variations in k feature numbers (k=5,10,15,20,30), feature selection algorithm (kNN1~kNN5), forecast period (T+1, T+6, T+24, T+48), and the period of time-series (Entire period, before lockdown, during the 1st lockdown), RandomizedSearchCV was used for hyperparameter tuning. The hyperparameters include learning rate (eta), the maximum depth for trees (max_depth), the number of trees (n_estimators), the subsampling ratio (subsample), and the minimum sum of instances weight needed in a child (min_child_weight). The number of folds for cross-validation (cv) was set to 5, and the number of iterations (n_iter_search) was 10. Table 2 below shows the detailed setting for hyperparameter tuning.

Table 2. The hyperparameter ranges for RandomizedSearchCV tuning

Hyperparameter	Range
eta	i/1000.0 for i in range(1, 100)
max_depth	2, 3, 4, 5, 6
n_estimators	i in range(500, 5000, 100)
subsample	0.5, 0.6, 0.7, 0.8, 0.9
min_child_weight	1, 2, 3, 4, 5, 6

And Table 3 shows the optimised hyperparameters for all cases.

Table 3. The optimised hyperparameter values

Hyperparameter	Value
eta	0.097
max_depth	6
n_estimators	800
subsample	0.8
min_child_weight	1

For training each model, 70% of the dataset was used as a training set and 30% as a test (validation) set. Each model is compared by model evaluation results using the test set, based on the evaluation methods in the following chapter.

3.5. Model evaluation methods

The evaluation criteria for each model were the root mean square error (RMSE), the mean absolute error (MAE), and R-square (R^2).

$$RMSE = \sqrt{\frac{1}{n}\left(\sum_{i=1}^{n}(\hat{y}_i - y_i)\right)}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

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

RMSE is measured by calculating the differences between prediction values and actual values, and R^2 shows the proportion of the variance for the target variable explained by the independent variables (Mombeini and Yazdani-Chamzini, 2015). MAE is an averaged absolute difference between predictions and actual values (Trenchevski et al., 2020). All metrics were calculated using the test set, which was created by splitting the dataset into a training set and a test set at a 7:3 ratio.

3.6. Time-series period

The period of time-series forecasting and analysis is the same as the meteorological and traffic record (1st Jan 2016 - 31st Dec 2020). To check if the model works well even with the sudden fundamental changes, we trained and tested each model with the three different versions of the time period; 1) the entire period, 2) pre-lockdown period (1st Jan 2016 – 22nd Mar 2020), and 3) the 1st Covid lockdown period (23rd Mar 2020 – 31st May 2020). First, each model will be trained and tested for each time period cases 1) to 3) to see any possible difference in model performance. In addition, the model trained using period 2) will be tested on the records during period 3) to check if the model could have predicted NO_2 concentration well despite the sudden change due to 1st Covid lockdown.

Also, the models are trained and tested by adjusting the forecast period in hours (T+1, T+6, T+24, T+48) to see how far in time the model can show reliable performance.

3.7. Experimental design

Based on the discussions above, this study will explore all possible cases to see how

certain changes affect the model's performance. The three main factors are 1) kNN-based feature selection methods, 2) forecasting period, and 3) time-series period. As we move on to the next stage, the model increases in scale gradually to include all the cases from the previous stage. It is to catch any possible changes in the relative performance of each model or any relationship between the main factors.

Furthermore, where possible, there will be some discussion of what each result means. Figure 6 below shows the experimental design of this study.

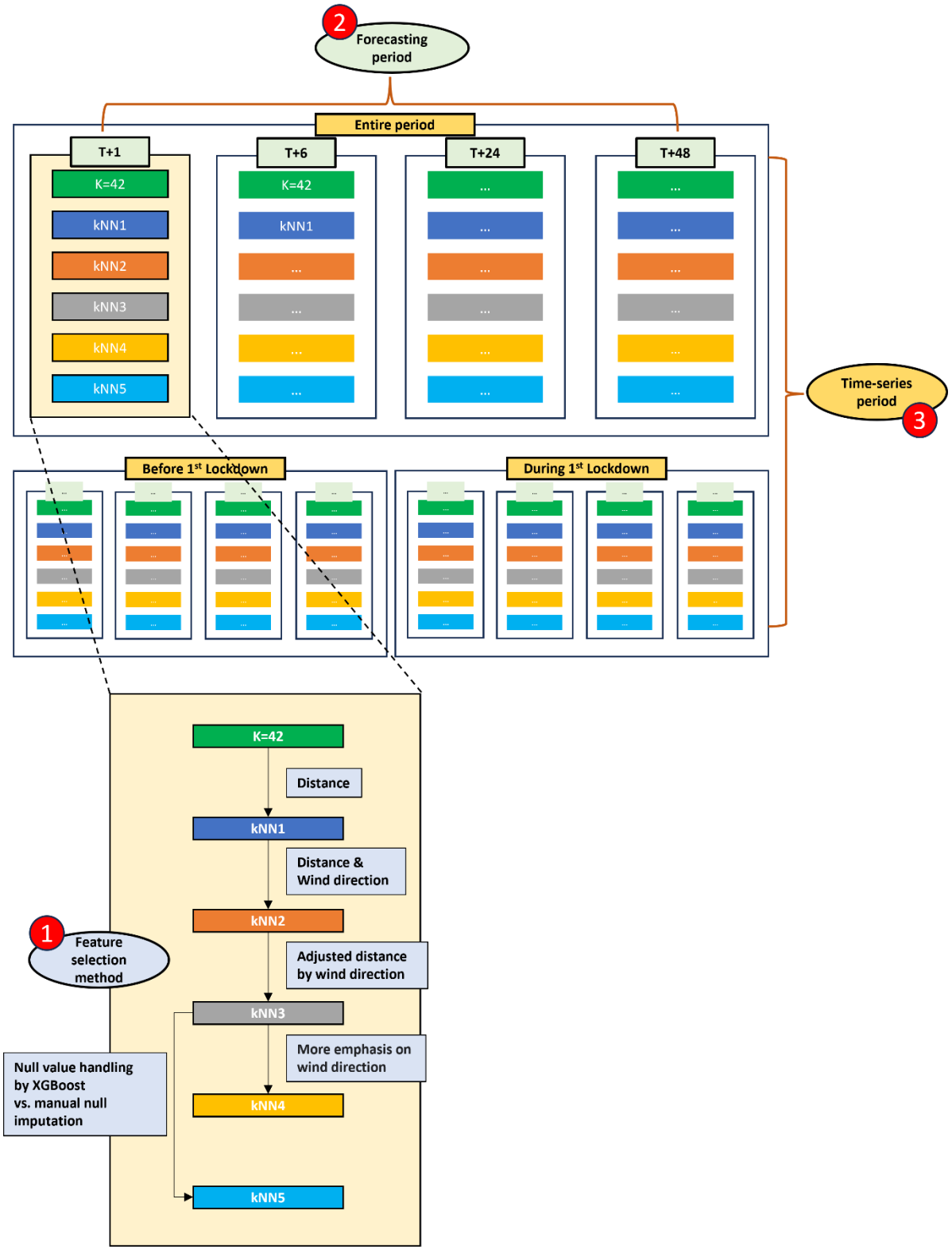

Figure 6. The experimental design of the study

4. RESULTS

4.1. Exploratory Data Analysis

4.1.1. Meteorological data

Figure 7 below shows the Pearson's correlation between each of the meteorological variables and NO_2. In general, the meteorological variables showed a weak correlation between NO_2. Wind speed and temperature were found to be negatively correlated with NO_2, which corresponds to the majority of results in Table 1 in Section 2.1.2. However, humidity is positively correlated with NO_2, which matches only two out of the nine studies in Table 1. Meanwhile, precipitation and atmospheric pressure showed a very weak correlation with NO_2, making it difficult to conclude whether these results are consistent with Liu et al. (2020) or R Kalbarczyk and E Kalbarczyk (2007). A moderate negative correlation can be found between wind speed and precipitation, and between humidity and temperature.

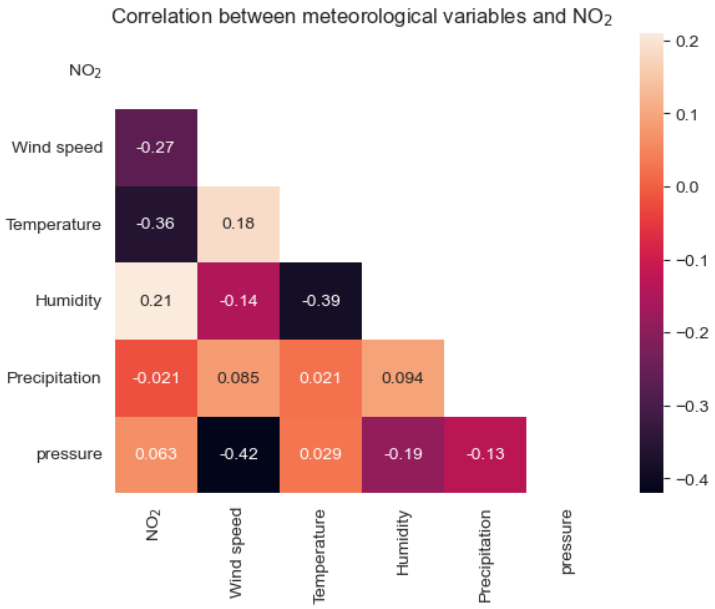

Figure 7. Correlation between meteorological variables and NO_2

We can check the distribution of wind direction by plotting wind roses. Figure 8 shows that the wind usually blew from the southwest during the entire period (1st Jan 2016 – 31st Dec 2020).

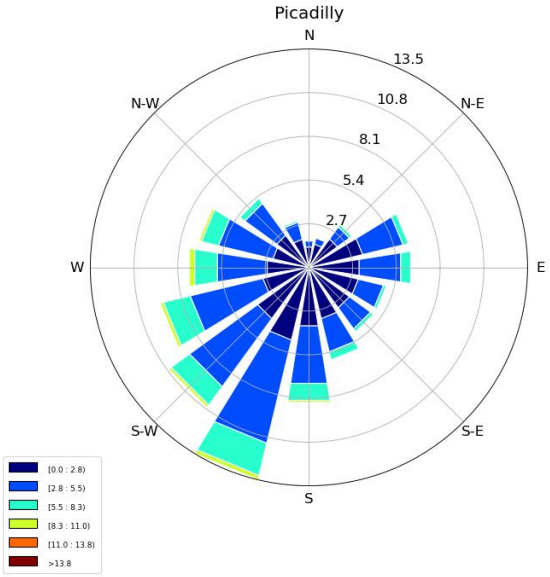

Figure 8. Wind rose for the entire period (1st Jan 2016 – 31st Dec 2020)

Wind roses between the pre-lockdown period and 1st lockdown period show different. As shown in Figure 9 below, the wind rose of the pre-lockdown period is similar to the entire period (9a), whereas there was more wind from the east during the 1st lockdown period (9b).

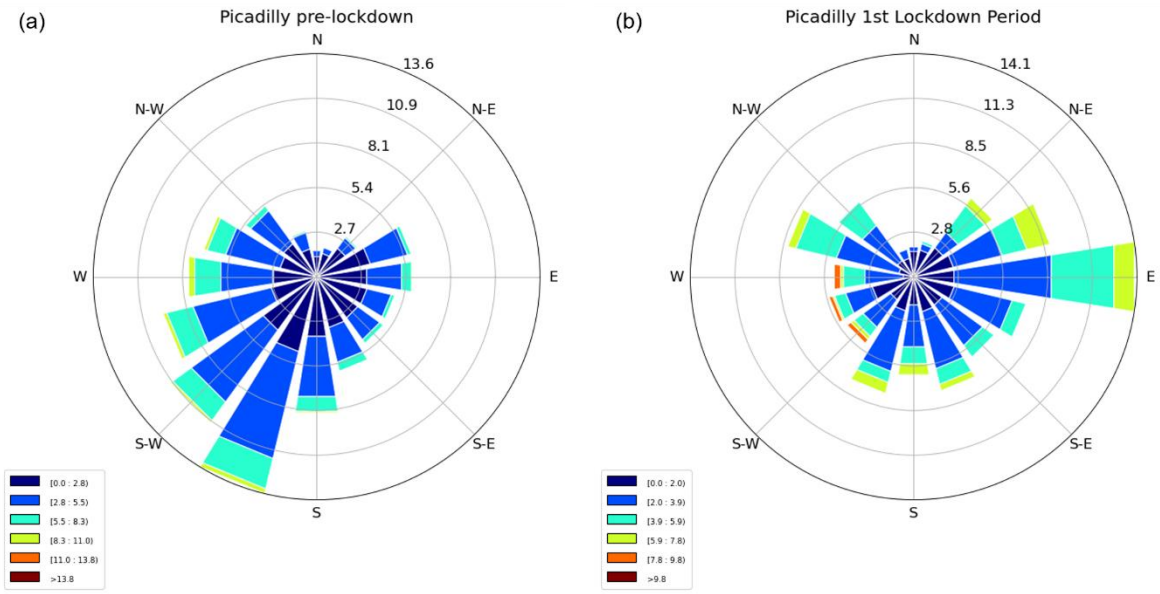

Figure 9. Wind rose of the pre-lockdown period (a) and 1st lockdown period (b)

Figure 10 shows the time-series graph of each meteorological variables and NO_2. We can find seasonal patterns in variables, especially NO_2, temperature, and humidity. The directions of the patterns match with the correlations in Figure 7.

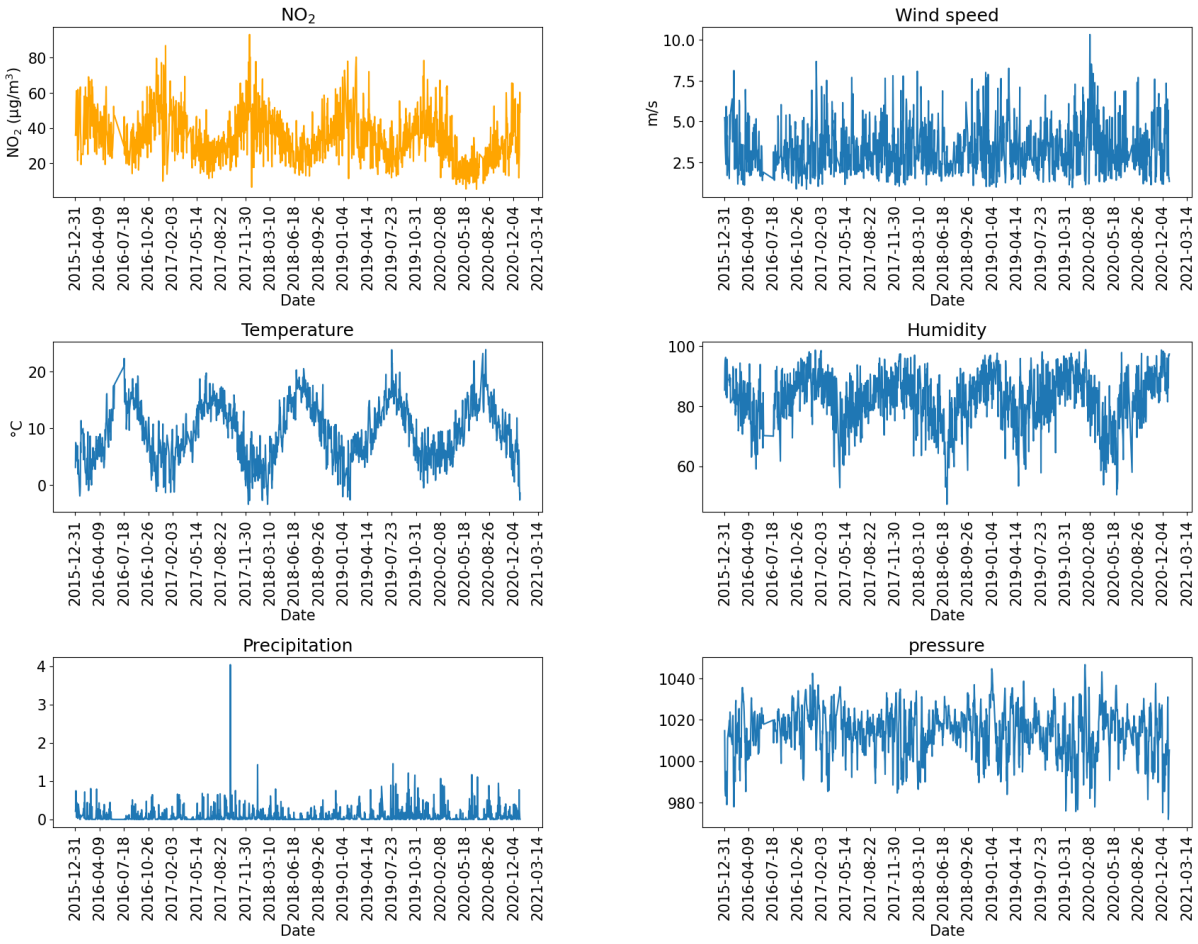

Figure 10. Time-series graphs of meteorological variables and NO_2

4.1.2. Traffic data

As shown in Figure 11, the traffic volume of each site showed a moderately positive correlation with NO_2, except for MAC000001078. Notably, the correlation does not weaken with increasing distance.

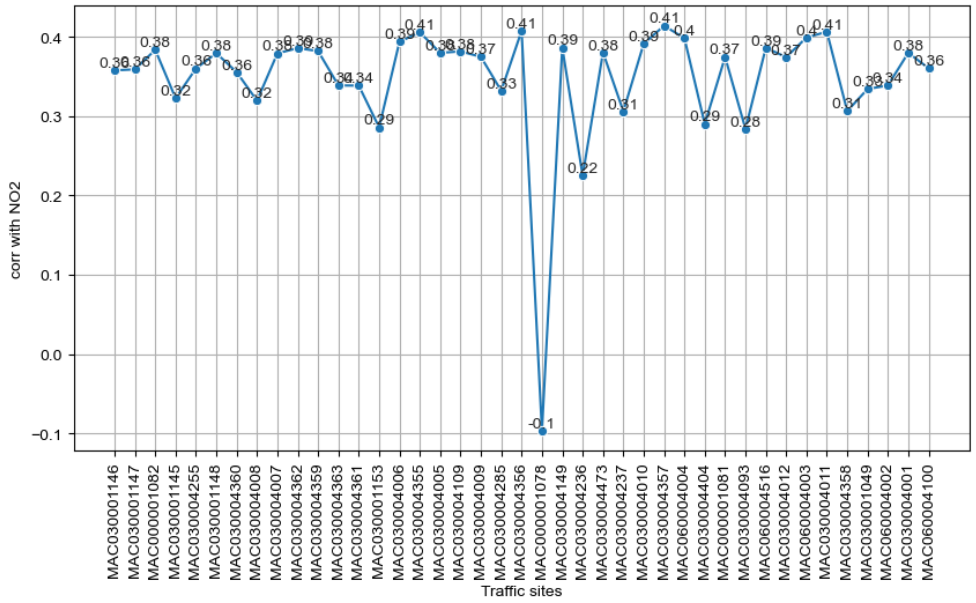

Figure 11. Correlation between traffic volumes and NO_2

Figure 12 shows the correlation between traffic site volumes. Except for MAC000001078, we can see the traffic volumes are highly positively correlated with each other. This might be because the traffic sites are located around Manchester city centre within a 3km distance from Piccadilly monitoring station. Combined with the result in Figure 11, the correlation not getting weaker with increasing distance may be due to the correlation between the traffic volumes, but this is not certain. A high correlation between traffic site volumes gives us room for testing feature selection methods to improve model performance.

We can plot traffic volumes and NO_2 on time-series graphs. Figure 13 shows NO_2 and total average traffic volume on a daily average. Excluding the period from April 2018 to March 2020, there is moderate consistency between NO_2 and traffic volume. We also can find a sharp decrease in NO_2 and traffic volume around 19th Mar 2020 due to the impact of COVID-19.

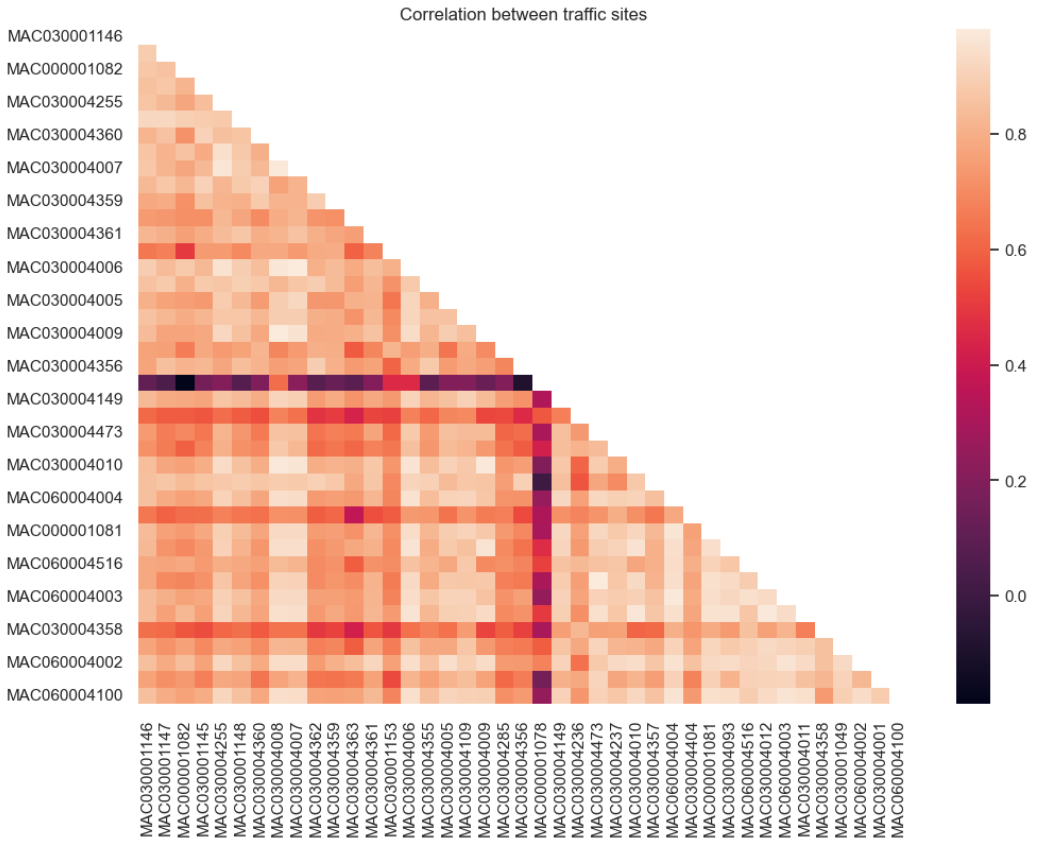

Figure 12. Correlation between traffic site volumes

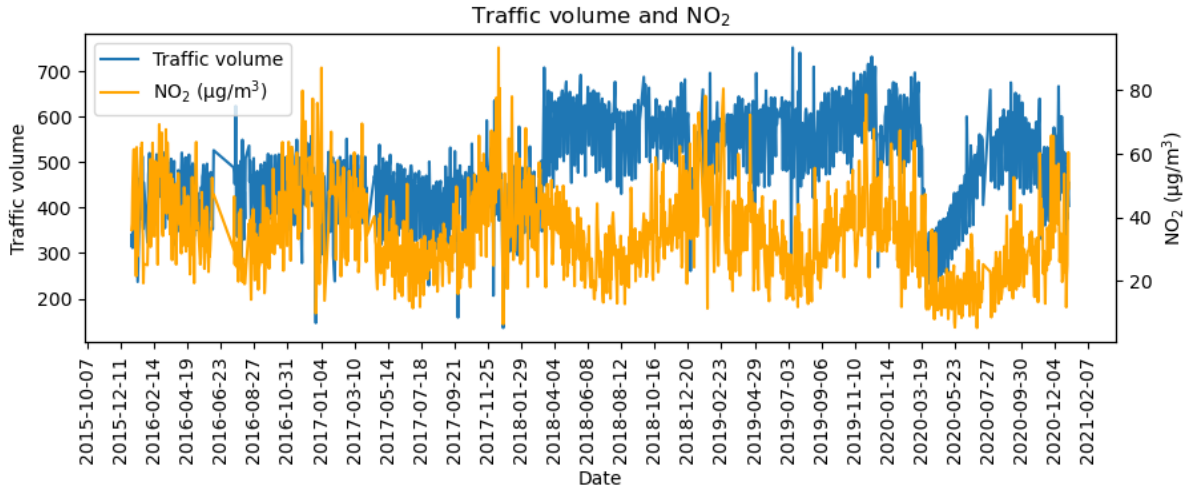

Figure 13. Time-series graph of average traffic volume and NO_2

4.2. Feature selection methods

In this section, we show the comparative results of each kNN-based feature selection model (kNN1-kNN5) based on the T+1 prediction period.

4.2.1. Entire dataset and kNN1

When we do feature selection by the distance between the Piccadilly monitoring station and traffic sites (kNN1), the model performance gradually decreases with fewer features (Figure 14). The decrease in performance was the biggest when we changed K=10 to K=5. The decrease in R2 value is less than 0.02 for each interval from K=42 to K=10. The R^2 value remained above 0.7 until the K value decreased to 15.

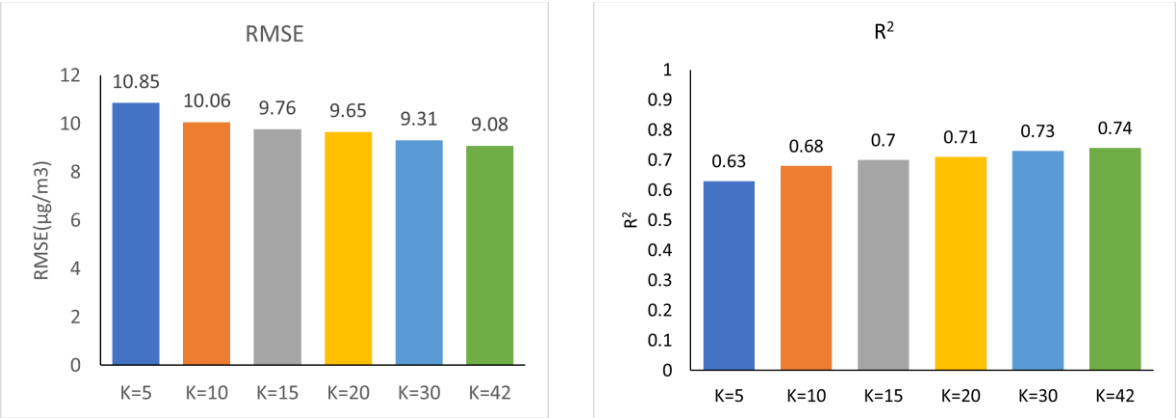

Figure 14. RMSE (left) and R^2 (right) by K values

4.2.2. kNN2 and kNN3

The difference between kNN2 and kNN3 lies in how to incorporate wind direction into feature selection. As mentioned in Chapter 3, both methods exclude any sites with θ > 90°, but kNN3 adjusts distance by θ while kNN2 uses the original distance. Figure 15 shows the RMSE and R^2 values for each case from K=5 to K=30, and we can find that kNN2 outperforms kNN1 and kNN3. Furthermore, the R^2 score of K=20 and K=30 for kNN2 were the same as the entire set (K=42). Figure 16 shows the predicted values and actual values of kNN1 and kNN2, with K=15 and T+1.

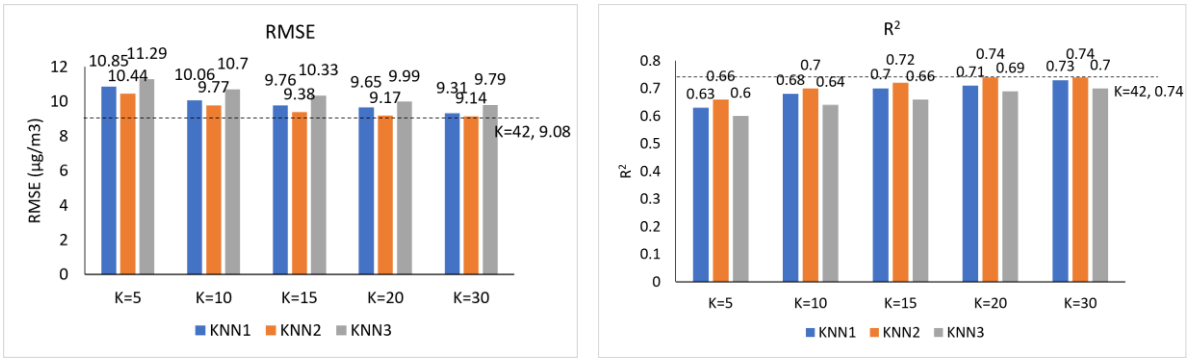

Figure 15. RMSE and R^2 of kNN1, kNN2, and kNN3

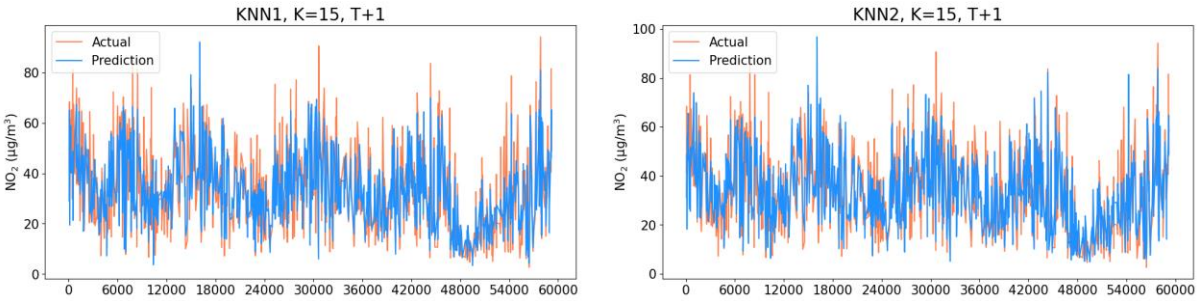

Figure 16. Prediction and actual values of kNN1 and kNN2 (K=15, T+1)

4.2.3. kNN3 and kNN4

kNN4 is another version of kNN3 with σ = 1, putting more emphasis on wind direction for adjusting distances. Figure 17 shows that kNN3 outperforms kNN4.

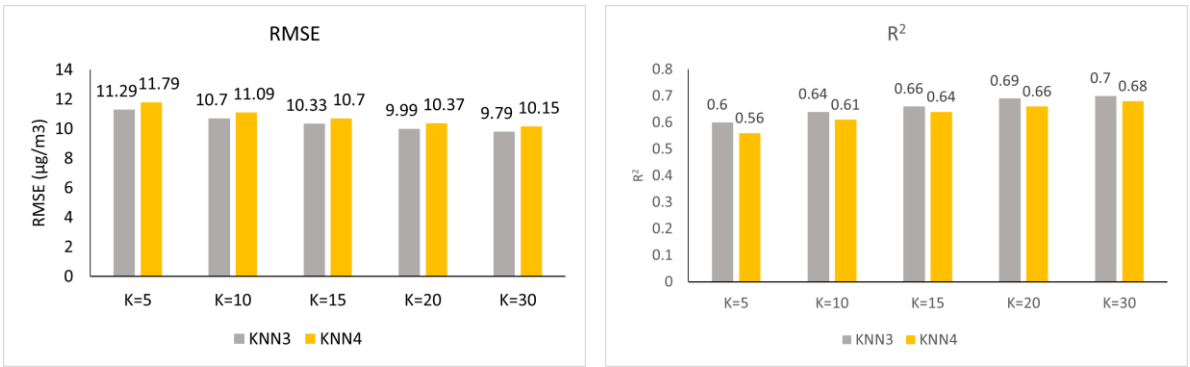

Figure 17. RMSE and R^2 of kNN3 and kNN4

4.2.4. kNN3 and kNN5

The difference between kNN3 and kNN5 is that kNN5 does manual imputation for null values. As shown in Figure 18, there was no significant difference in performance between the two models.

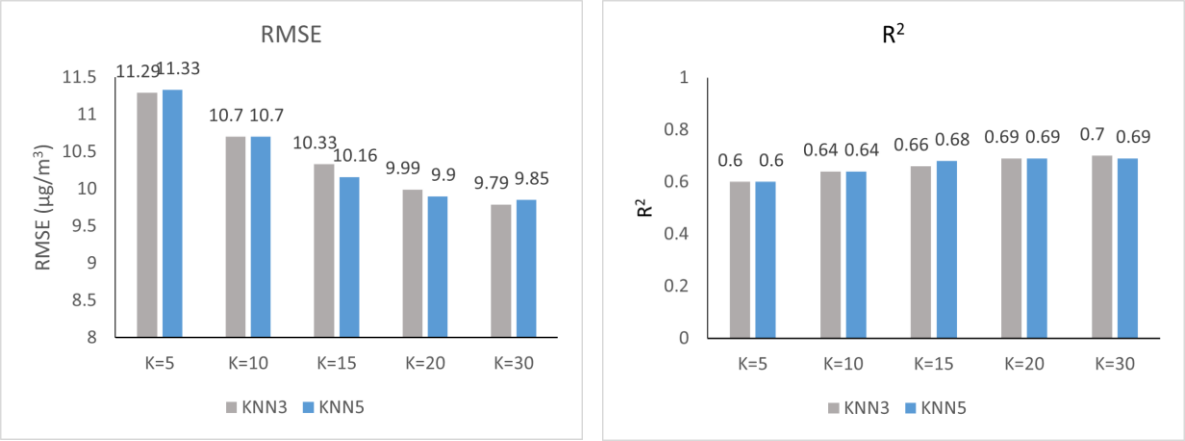

Figure 18. RMSE and R2 of kNN3 and kNN5

4.3. Forecasting period

Figure 19 shows the comparative results of the models for different forecasting periods (T+6, T+24, T+48). Based on RMSE values, the entire set model showed the best performance. And a longer forecasting period led to lower prediction accuracy.

The comparison result between kNN1 and kNN2 for K=5 with all the period cases is the same as T+1. However, for K values higher than 5, kNN1 was more accurate than kNN2 for all the period cases. The RMSE difference ranges from 0.2 to 0.16. kNN5 showed slightly lower RMSE than kNN3, and kNN3 was followed by kNN4 except for K=30 with T+48 case. In general, the RMSE score gets higher as the forecasting period gets longer from 6 to 24, and 24 to 48. Adding more features to the model resulted in more accurate predictions.

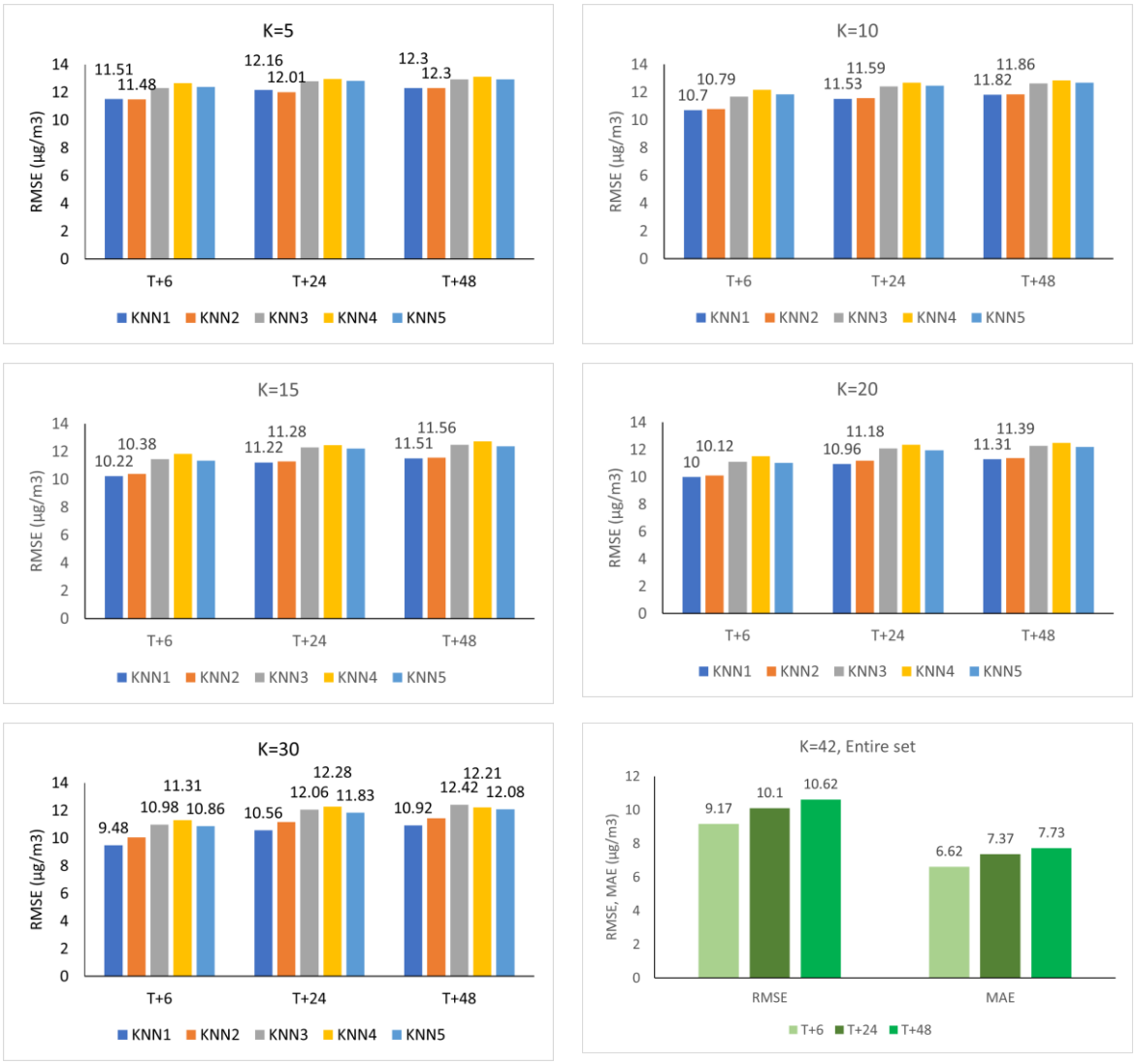

Figure 19. The model performances by different forecasting periods

4.4. Time-series period

The test for different time-series periods has been done in two ways: 1) train and test models for each of the periods independently (independent test), 2) train the pre-lockdown period model to test on the 1st lockdown period dataset (cross-test).

4.4.1. Independent test

Figure 20 shows the independent test results in R^2 values with K=5 and K=30. Comparing the entire period (1st row) and pre-lockdown period (2nd row), we can see the pre-lockdown period results show the same pattern as the entire period but with

lower R^2 values. More features (higher K) and shorter forecasting periods (T) resulted in higher accuracy (higher R^2).

However, the result of the 1st lockdown period showed different patterns from the other two periods. First, the prediction was generally more accurate than the pre-lockdown period for all the cases. Especially, the R^2 value of the K=5 case ranges from 0.7 to 0.9 for the 1st lockdown period case, while 0.48 to 0.63 for the pre-lockdown period. Second, the T+1 case showed lower accuracy than the other longer periods. T+6 case performed best, followed by T+24, T+48, and T+1. Another notable thing is that there was no significant difference in general between K=5 and K=30.

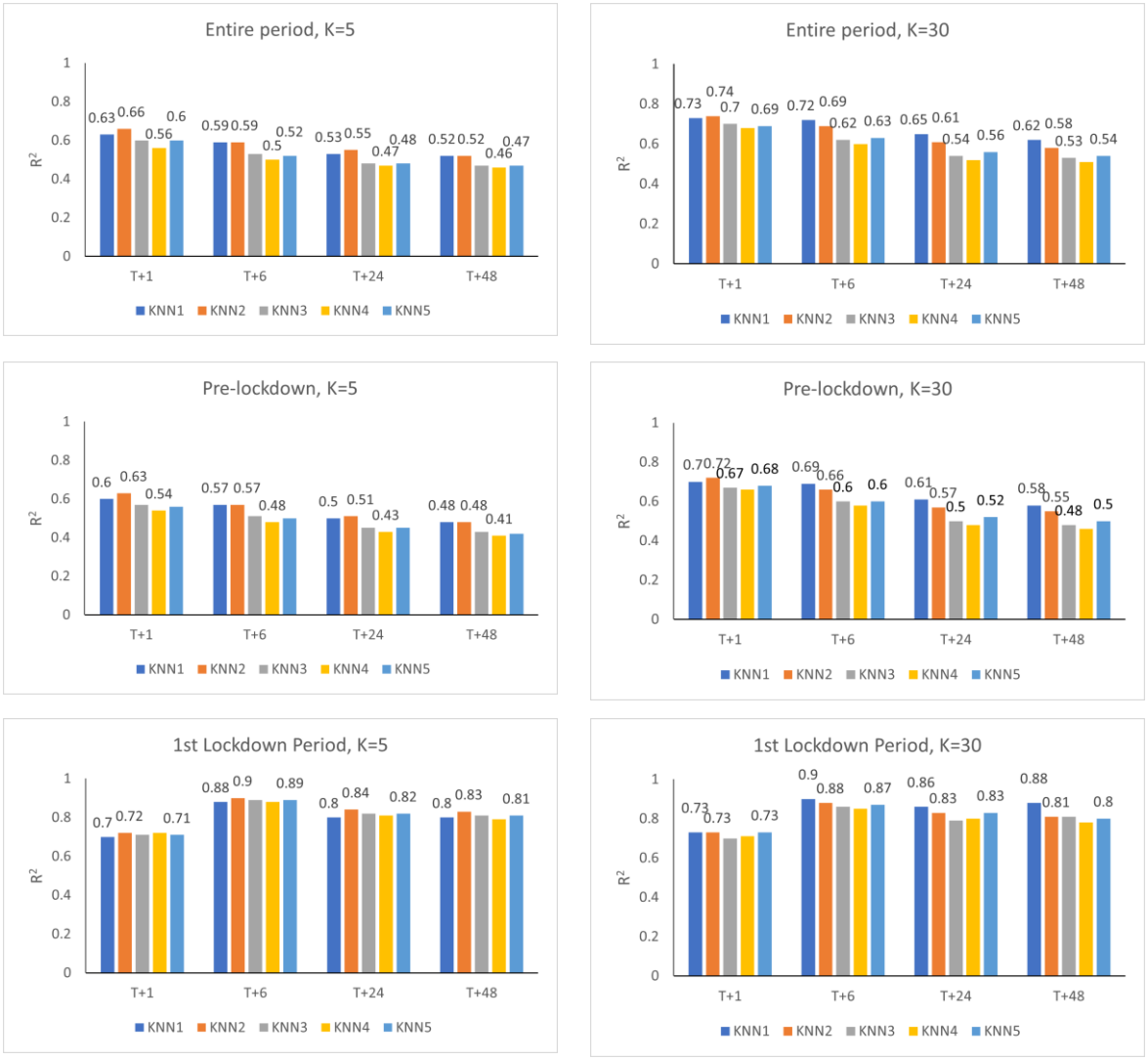

Figure 20. Independent test results for different time-series periods

Compared to the entire set result (Figure 21), kNN1 with K=30 for the lockdown period showed no significant difference in accuracy. kNN2 with K=5 for the same period also showed small differences in R^2, 0.0225 on average.

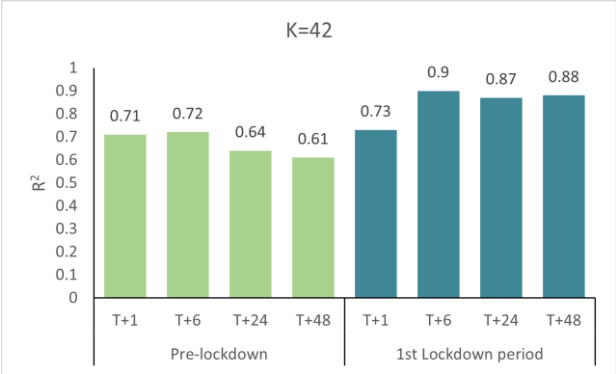

Figure 21. R^2 values for the entire dataset model by different time-series period

Meanwhile, the comparison result between kNN1 and kNN2 was the same as the other periods; kNN1 showed better performance than kNN2 except for cases with K=5 or T+1.

4.4.2. Cross-test

The test result of testing the pre-lockdown model on the lockdown period data shows poor performance of the models. As shown in Table 4, R^2 values were negative for all cases. This contrasts with the results of independent tests for the first lockdown period, which showed high accuracy with R^2 values ranging from 0.73 to 0.9. The implications of this result will be discussed in the next chapter.

Table 4. R2 values of the cross-test (Pre-lockdown period → 1st lockdown period)

R^2		Forecasting period (hour)			
		T+1	T+6	T+24	T+48
Entire (K=42)		-0.54	-1.76	-1.48	-2.3
K=30	kNN1	-0.67	-1.87	-2.63	-3.92
	kNN2	-0.91	-2.02	-2.5	-2.44
	kNN3	-1.02	-2.15	-2.34	-1.84
	kNN4	-0.84	-1.87	-2.19	-1.78
	kNN5	-0.84	-1.91	-2.66	-2.06
K=5	kNN1	-0.85	-2.12	-2.29	-1.6
	kNN2	-0.49	-2.04	-2.31	-2.13
	kNN3	-0.86	-2.37	-2.12	-1.89
	kNN4	-1.13	-2.24	-2.37	-1.98
	kNN5	-0.82	-2.18	-2.34	-2.15

5. DISCUSSION

5.1. Distance and wind direction for NO_2 prediction

The comparative results between the models from kNN1 to kNN5 tell us about the effect of distance and wind direction on NO_2 prediction performance using traffic and meteorological data. And the result differs by the forecasting period from T+1 to T+48.

5.1.1. T+1 forecasting

First, the result of kNN2 outperforming the other models for the T+1 case shows using wind direction for NO_2 prediction (when using meteorological and traffic data) can be effective. However, the results of kNN3, kNN4, and kNN5 showed that putting more emphasis on wind direction by adjusting physical distances may lead to poorer model performance. Especially from the results of kNN4 showing the lowest prediction accuracy, we can better observe this trend. This result may be due to the size of the study area. As we discussed in Chapter 3, all the traffic sites are within 3km distance from the Piccadilly monitoring station. Yang, Fan and Zhao (2019) used their kNN-DWFD model for the air quality stations in Beijing; a larger area compared to this study.

Second, it is important that the kNN2 model with K=20 and K=30 showed the same R^2 value as the entire dataset model (K=42). This shows we can effectively reduce the number of features for prediction using distance and wind direction, without affecting the prediction performance.

5.1.2. T+6, T+24, and T+48 forecasting

However, changing the forecast period from T+1 to T+6, T+24, and T+48 resulted in a different comparison from Section 5.1.2. For any longer period than T+1, kNN1 outperformed the other models using wind direction, including kNN2. This may be due to the small study area (3km zone). Wind direction can change in hours, and the effects of wind are short-lived as pollutants carried by the wind from traffic sites pass through the climate station in a relatively short time compared to large areas. Wind

speed is included as an independent variable in the models, but we may need to find a way to combine distance, wind speed, and direction and incorporate them into the model more effectively. The fact that the model using the entire dataset (K=42) outperforms the other models also points to the need for a new appropriate feature selection method.

Meanwhile, kNN2 outperforming kNN1 with K=5 implies that if we are going to choose a small number of the nearest traffic sites, considering wind direction can be more effective than only using distances for feature selection.

5.2. Model performance by time-series periods

Changing the prediction periods, such as 1) the entire period, 2) the pre-lockdown period, and 3) the 1st lockdown period, resulted in significant differences in model performance. The first is that the model trained using the record of the pre-lockdown period did not properly predict NO_2 during the 1st lockdown period. The negative R^2 values in Table 4 in Section 4.4.2 mean a prediction failure.

To explain why this happened, it is worth noting that the model using only the 1st lockdown period dataset showed higher prediction performance than the model using only the pre-lockdown period dataset. Resulting in different performance with the same model structure mean that there may be a fundamental difference in the relationship between NO_2 and the independent variables between the two periods.

In this regard, ONS (2021) gives us an important clue. According to ONS (2021), 24.3% of all industries across the UK had temporarily closed or paused trading during the 1st lockdown period. And in Figure 22 below, we can find that the 1st lockdown had a greater impact on other industries, such as arts entertainment and recreation (82.2%), construction (29.1%), and manufacturing (22.7%) than on transport and storage (7.7%).

As we discussed in Section 2.1, NO_2 can be emitted from a variety of sources, including industrial combustion, electricity and heat production, vehicles, and other transport. The greater impact of COVID-19 on other industries may have led to a relatively higher traffic contribution to NO_2 emissions. This means that our model, which predicts NO_2 concentrations using only traffic and meteorological data, can

have greater explanatory power. In the same context, it also explains why the lockdown period models showed better performance than the other period models.

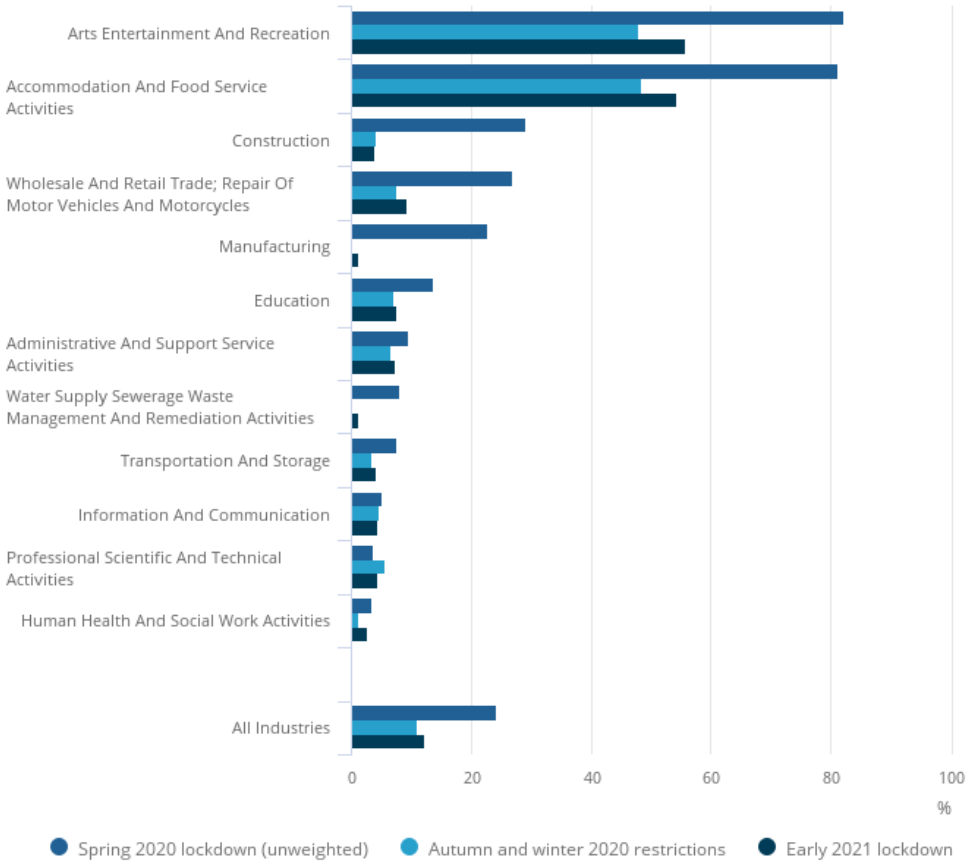

Figure 22. Proportion of UK businesses that had temporarily closed or paused trading across a two-week period in lockdown, by industry, UK (ONS, 2021)

Furthermore, the changes in the relative traffic contribution to NO_2 emission during the 1st lockdown mean that the relationship between NO_2 concentration and traffic volume has also changed. We can infer that this is why the cross-test in Section 4.4.2 resulted in poor model performance.

5.3. XGBoost in handling missing values

As shown in Chapter 4, no significant difference in model performance was found between kNN3 and kNN5. This shows how well XGBoost can handle missing values by itself.

6. CONCLUSION

Accurate NO_2 prediction is important for effective policy responses. This study trained and tested various XGBoost prediction models with kNN-based feature selection. The meteorological data and traffic data from 1st Jan 2016 to 31st Dec 2020 were used. As a result, we found the kNN2 feature selection model that uses distance and wind direction to be the most efficient model for the T+1 case or K=5 case. For the other time-series periods (T+6, T+24, and T+48) and K numbers (K=10, K=15, K=20, and K=30), the most efficient model was the kNN1 that uses only distance.

The models using adjusted distance by wind direction, such as kNN3, kNN4, and kNN5, showed lower accuracy than the other models. This seems to be due to the short study range of 3km. We can conclude that within a small range, wind direction plays only a limited role in predicting NO_2 or other pollutants. Feature selection did not lead to higher model accuracy, only meaningful in selecting a smaller number of features while maintaining similar accuracy. And kNN2 outperforming kNN1 for the K=5 case tells us that if we are going to use only a small number of features, using wind direction for feature selection can be effective.

Meanwhile, the model using the lockdown period data showed higher accuracy than the one using only the pre-lockdown data. Also, the model trained using the pre-lockdown data failed to make accurate predictions for the lockdown period. This may be due to the changes in the contribution rates by sources of NO_2 during the lockdown period. The greater impact of COVID-19 was on other industries than the transport sector, leading to relatively increased traffic contribution to NO_2. For the lockdown period, it was even possible to get the same model accuracy (R^2=0.90) with only the five nearest traffic sites (K=5) as the entire set (K=42).

This study suffers several limitations because of its nature. First, the models need to be tested with a bigger study area. As mentioned in Chapter 5, Yang, Fan and Zhao (2019) that introduced the kNN-DWFD model covers the whole area of Beijing. There is a probability of adjusted distance models such as kNN3, kNN4, and kNN4 showing better performance than kNN1 or kNN2 if trained and tested with longer distances.

The failure to adequately predict NO_2 for the lockdown period using the pre-lockdown models suggests that NO_2 prediction models may need to include more independent variables in addition to traffic data. Further studies are required to train and test models by adding more features; we may find an effective model that gives high prediction accuracy with the fewest number of variables. And such a model may continue to show high accuracy even in the face of sudden changes in circumstances, such as COVID-19.

This study considered only one climate station in Manchester Piccadilly. As mentioned earlier, since NO_2 emissions are affected by various factors, the results of this study may be different for places with different regional characteristics. Further research involving multiple regions (e.g., major urban cities of the world, or other types of regions sharing some characteristics) may be conducted in the future to draw meaningful implications.

In addition, there are other technical limitations due to the limited time and computational power. We may increase the model accuracy by testing more hyperparameters using grid search. Testing different versions of models using algorithms such as deep neural networks may help us to find the benefits of using XGBoost or other meaningful conclusions.

LIST OF REFERENCES

Ayus, I., Natarajan, N. and Gupta, D. (2023). 'Comparison of machine learning and deep learning techniques for the prediction of air pollution: a case study from China', *Asian journal of atmospheric environment*, 17(1), pp. 1–22. [Online]. Available at: https://doi.org/10.1007/s44273-023-00005-w (Accessed: 28 August 2023).

Bentéjac, C., Csörgő, A. and Martínez-Muñoz, G. (2020). 'A comparative analysis of gradient boosting algorithms', *The Artificial Intelligence Review*, 54(3), pp. 1937-1967. [Online]. Available at: https://doi.org/10.1007/s10462-020-09896-5 (Accessed: 29 August 2023).

Bouktif, S., et al. (2020). 'Multi-sequence LSTM-RNN deep learning and metaheuristics for electric load forecasting', *Energies*, 13(2), p. 391. [Online]. Available at: https://doi.org/10.3390/en13020391 (Accessed: 29 August 2023).

Brownlee, J. (2021). *XGBoost With Python: Gradient Boosted Trees with XGBoost and scikit-learn*. 1st edn. *Machine Learning Mastery*. [Online]. Available at: https://machinelearningmastery.com/xgboost-with-python/ (Accessed: 29 August 2023).

California Air Resources Board (CARB) (2023). *Nitrogen Dioxide & Health*. Available at: https://ww2.arb.ca.gov/resources/nitrogen-dioxide-and-health (Accessed: 28 August 2023).

Chen, T. (2016). *Tianqi Chen's answer to What is the difference between the R gbm (gradient boosting machine) and xgboost (extreme gradient boosting)? - Quora*. Available at: https://www.quora.com/What-is-the-difference-between-the-R-gbm-gradient-boosting-machine-and-xgboost-extreme-gradient-boosting/answer/Tianqi-Chen-1 (Accessed: 29 August 2023).

Chen, T. and Guestrin, C. (2016). 'XGBoost: A scalable tree boosting system', *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco. 13th – 17th August 2016. Cornell University Library. pp. 785-794. Available at: https://doi.org/10.1145/2939672.2939785 (Accessed: 28 August 2023).

Cooper, M., et al. (2022). 'Global Fine-Scale Changes in Ambient NO2 During COVID-19 Lockdowns', *Nature (London)*, 601(7893), pp. 380–387. [Online]. Available at: https://doi.org/10.1038/s41586-021-04229-0 (Accessed: 28 August 2023).

Department for Environment Food & Rural Affairs (DEFRA) (2004) *Nitrogen Dioxide in the United Kingdom*. Available at: https://uk-air.defra.gov.uk/library/assets/documents/reports/aqeg/Nitrogen_Dioxide_in_the_UK_2004.pdf (Accessed: 29 August 2023).

Department for Environment Food & Rural Affairs (DEFRA) (2017). *UK plan for tackling roadside nitrogen dioxide concentrations: An overview*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/633269/air-quality-plan-overview.pdf (Accessed: 29 August 2023).

Department for Environment Food & Rural Affairs (DEFRA) (2022). *Air Pollution in the UK 2021*. Available at: https://uk-air.defra.gov.uk/assets/documents/annualreport/air_pollution_uk_2021_issue_1.pdf (Accessed: 28 August 2023).

Department for Environment Food & Rural Affairs (DEFRA) (2023). *Air quality statistics background*. Available at: https://www.gov.uk/government/statistics/air-quality-statistics/background (Accessed: 28 August 2023).

Drewil, G. and Al-Bahadili, R. (2022). 'Air pollution prediction using LSTM deep learning and metaheuristics algorithms', *Measurement. Sensors*, 24, p. 100546. [Online]. Available at: https://doi.org/10.1016/j.measen.2022.100546 (Accessed: 29 August 2023).

Dunlea, E.J., et al. (2007). 'Evaluation of Nitrogen Dioxide Chemiluminescence Monitors in a Polluted Urban Environment', *Atmospheric chemistry and physics*, 7(10), pp. 2691–2704. [Online]. Available at: https://doi.org/10.5194/acp-7-2691-2007 (Accessed: 28 August 2023).

Elminir, H.K. (2005). 'Dependence of urban air pollutants on meteorology', *The Science of the total environment*, 350(1), pp. 225-237. [Online]. Available at: https://doi.org/10.1016/j.scitotenv.2005.01.043 (Accessed: 28 August 2023).

European Environment Agency (EEA) (2022). *Health impacts of air pollution in Europe, 2022*. Available at: https://www.eea.europa.eu/publications/air-quality-in-europe-2022/health-impacts-of-air-pollution (Accessed: 29 August 2023).

Gasmi, K., et al. (2017). 'Analysis of NOx, NO and NO2 Ambient Levels as a Function of Meteorological Parameters in Dhahran, Saudi Arabia', *WIT Transactions on Ecology and the Environment*, 211, pp. 77–86. [Online]. Available at: https://doi.org/10.2495/AIR170081 (Accessed: 29 August 2023).

Habeebullah, T.M., et al. (2015). 'The Interaction between Air Quality and Meteorological Factors in an Arid Environment of Makkah, Saudi Arabia'. *International Journal of Environmental Science and Development*, 6(8), pp. 576-580. [Online]. Available at: https://doi.org/10.7763/IJESD.2015.V6.660 (Accessed: 29 August 2023).

Hamra, G., et al. (2015). 'Lung Cancer and Exposure to Nitrogen Dioxide and Traffic: A Systematic Review and Meta-Analysis', *Environmental health perspectives*, 123(11), pp. 1107–1112. [Online]. Available at: https://doi.org/10.1289/ehp.1408882 (Accessed: 29 August 2023).

Kim, K., et al. (2015). 'Influence of Wind Direction and Speed on the Transport of Particle-Bound PAHs in a Roadway Environment', *Atmospheric pollution research*, 6(6), pp. 1024–1034. [Online]. Available at: https://doi.org/10.1016/j.apr.2015.05.007 (Accessed: 28 August 2023).

Liu, D., et al. (2019). 'Air Pollution Forecasting Based on Attention-based LSTM Neural Network and Ensemble Learning', *Expert systems*, 37(3), p. n/a. [Online]. Available at: https://doi.org/10.1111/exsy.12511 (Accessed: 29 August 2023).

Liu, Z., et al. (2020). 'Analysis of the Influence of Precipitation and Wind on PM2.5 and PM10 in the Atmosphere', *Advances in meteorology*, 2020, pp. 1–13. [Online]. Available at: https://doi.org/10.1155/2020/5039613 (Accessed: 28 August 2023).

Méndez, M., Merayo, M. and Núñez, M. (2023). 'Machine Learning Algorithms to Forecast Air Quality: a Survey', *The Artificial intelligence review*, 56(9) pp. 10031–10066. [Online]. Available at: https://doi.org/10.1007/s10462-023-10424-4 (Accessed: 29 August 2023).

Ministry for the Environment (MfE) (2021). *Nitrogen dioxide*. Available at: https://environment.govt.nz/facts-and-science/air/air-pollutants/nitrogen-dioxide-effects-health/ (Accessed: 29 August 2023).

Mombeini, H. and Yazdani-Chamzini, A. (2015). 'Modeling Gold Price via Artificial Neural Network', *Journal of Economics, Business and Management*, 3(7), pp. 699-703. [Online]. Available at: https://doi.org/10.7763/JOEBM.2015.V3.269 (Accessed: 28 August 2023).

National Atmospheric Emissions Inventory (NAEI) (2022). *Pollutant Information: Nitrogen Oxides*. Available at: https://naei.beis.gov.uk/overview/pollutants?pollutant_id=6 (Accessed: 28 August 2023).

National Oceanic and Atmospheric Administration (NOAA) (2010). *Nitrogen Dioxide – Science On a Sphere.* Available at: https://sos.noaa.gov/catalog/datasets/nitrogen-dioxide/ (Accessed: 29 August 2023).

Office for National Statistics (ONS) (2021). *Coronavirus: how people and businesses have adapted to lockdowns.* Available at: https://www.ons.gov.uk/economy/economicoutputandproductivity/output/articles/coronavirushowpeopleandbusinesseshaveadaptedtolockdowns/2021-03-19 (Accessed: 30 August 2023).

Pan, B. (2018). 'Application of XGBoost algorithm in hourly PM2.5 concentration prediction', *IOP conference series: Earth and Environmental Science*, 113(1), p. 12127. [Online]. Available at: https://doi.org/10.1088/1755-1315/113/1/012127 (Accessed: 27 August 2023).

Price, J., et al. (2022). 'XGBoost: Interpretable Machine Learning Approach in Medicine', *2022 5th World Symposium on Communication Engineering (WSCE)*, Nagoya, Japan. 16th – 18th September 2022. IEEE. pp. 109-113. Available at: https://doi.org/10.1109/WSCE56210.2022.9916029 (Accessed: 28 August 2023).

Quinto, B. (2020). *Next-Generation Machine Learning with Spark: Covers XGBoost, LightGBM, Spark NLP, Distributed Deep Learning with Keras, and More*. 1st edn. Berkeley, CA: Apress L.P.

Srivastava, R., Sarkar, S. and Beig, G. (2015). ‘Correlation of Various Gaseous Pollutants with Meteorological Parameter (Temperature, Relative Humidity and Rainfall)’, *Global Journal of Science Frontier Research*, 14(H6), pp. 57-65. [Online]. Available at: https://journalofscience.org/index.php/GJSFR/article/view/1452 (Accessed: 29 August 2023).

Tao, Q., et al. (2019). ‘Air Pollution Forecasting Using a Deep Learning Model Based on 1D Convnets and Bidirectional GRU’, *IEEE access*, 7, pp. 76690–76698. [Online]. Available at: https://doi.org/10.1109/ACCESS.2019.2921578 (Accessed: 29 August 2023).

Teng, M., et al. (2023). ‘72-Hour Real-Time Forecasting of Ambient PM2.5 by Hybrid Graph Deep Neural Network with Aggregated Neighborhood Spatiotemporal Information’, *Environment international*, 176, pp. 107971–107971. [Online]. Available at: https://doi.org/10.1016/j.envint.2023.107971 (Accessed: 29 August 2023).

Trenchevski, A., et al. (2020). ‘Prediction of Air Pollution Concentration Using Weather Data and Regression Models’, *Proceedings of International Conference on Applied Innovation in IT*, Koethen, Germany. 10th March 2020. Anhalt University of Applied Sciences. pp. 55-61. Available at: https://doi.org/10.25673/32749 (Accessed: 28 August 2023).

UK Parliament (2021). *Coronavirus: A history of English lockdown laws.* Available at: https://commonslibrary.parliament.uk/research-briefings/cbp-9068/ (Accessed: 29 August 2023).

United States Environmental Protection Agency (EPA) (2023a*). Basic Information about NO2*. Available at: https://www.epa.gov/no2-pollution/basic-information-about-no2#Effects (Accessed: 29 August 2023).

United States Environmental Protection Agency (EPA) (2023b). *Setting and Reviewing Standards to Control NO2 Pollution*. Available at: https://www.epa.gov/no2-pollution/setting-and-reviewing-standards-control-no2-pollution (Accessed: 29 August 2023).

Verma, S. and Desai, B. (2008). ‘Effect of Meteorological Conditions on Air Pollution of Surat City’, *Journal of International Environmental Application and Science*, 3(5), pp. 358-367. [Online]. Available at:

https://www.yumpu.com/en/document/view/48891473/effect-of-meteorological-conditions-on-air-pollution-of-surat-city (Accessed: 29 August 2023).

Vilčekova, S. (2010). 'Indoor Nitrogen Oxides', in Nejadkoorki, F. (eds.) *Advanced Air Pollution*. Rijeka, Croatia: InTech, pp. 31-50.

Voiculescu, M., et al. (2020). 'Role of Meteorological Parameters in the Diurnal and Seasonal Variation of No2 in a Romanian Urban Environment', *International journal of environmental research and public health*, 17(17), pp. 1–15. [Online]. Available at: https://doi.org/10.3390/ijerph17176228 (Accessed: 28 August 2023).

Wong, P., et al. (2021). 'Using Land-Use Machine Learning Models to Estimate Daily NO2 Concentration Variations in Taiwan', *Journal of cleaner production*, 317, p. 128411. [Online]. Available at: https://doi.org/10.1016/j.jclepro.2021.128411 (Accessed: 28 August 2023).

World Health Organization (WHO) (2010). *WHO Guidelines for Indoor Air Quality: Selected Pollutants*. Available at: https://apps.who.int/iris/handle/10665/260127 (Accessed: 28 August 2023).

World Health Organization (WHO) (2022). *Ambient (outdoor) air pollution*. Available at: https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health (Accessed: 29 August 2023).

Zaini, N., et al. (2022). 'PM2.5 Forecasting for an Urban Area Based on Deep Learning and Decomposition Method', *Scientific reports*, 12(17565), pp. 1-13. [Online]. Available at: https://doi.org/10.1038/s41598-022-21769-1 (Accessed: 29 August 2023).

Zhang, H., et al. (2015). 'Relationships Between Meteorological Parameters and Criteria Air Pollutants in Three Megacities in China', *Environmental research*, 140, pp. 242–254. [Online]. Available at: https://doi.org/10.1016/j.envres.2015.04.004 (Accessed: 29 August 2023).

Zhang, X., et al. (2019). 'XGBoost Imputation for Time Series Data', *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, Xi'an, China. 10th – 13th June 2019. IEEE. pp. 1-3. Available at: https://doi.org/10.1109/ICHI.2019.8904666 (Accessed: 30 August 2023).

Zou, M., et al. (2022). ‘Optimized XGBoost Model with Small Dataset for Predicting Relative Density of Ti-6Al-4V Parts Manufactured by Selective Laser Melting’, *Materials*, 15(15), p. 5298. [Online]. Available at: https://doi.org/10.3390/ma15155298 (Accessed: 28 August 2023).