Imputation through Clustering of Time Series Data: A case study in air pollution

Wedad Obaidallah Alahamade

A thesis presented for the degree of

Doctor of Philosophy

School of Computing Sciences

University of East Anglia

United Kingdom

31st May 2022

Imputation through Clustering of Time Series Data: A case study in air pollution

Wedad Obaidallah Alahamade

Imputation through Clustering of Time Series Data: A case study in air pollution

Wedad Obaidallah Alahamade

Abstract

Air pollution is a global problem, and air pollution concentration assessment plays an essential role in evaluating the associated risk to human health. Unfortunately, air pollution monitoring stations often have periods of missing data.

In this thesis, we investigated missing values problem in air quality data by looking at the hourly pollutant concentration Time Series (TS) of the main four pollutants included in air quality assessment: O_3, NO_2, $PM_{2.5}$, and PM_{10}. The research presented in this thesis aims to reduce the uncertainty of the air quality assessment by proposing methods for the imputation of missing values either partially or completely. Our approach uses clustering of stations based on measured pollutants to inform the imputation.

We started by testing uni-variate clustering and then developing a multivariate time series (MVTS) clustering method that considers all measured pollutants at a station by aggregating the similarity between those pollutants (through a fused distance) followed by imputation models for the whole TS.

We developed various imputation models including ensemble models which aggregate temporal similarity obtained from clustering and spatial similarity obtained by the geographical correlation between stations.

Our experimental results show that using MVTS clustering enables imputation of unmeasured pollutants in any station and produced plausible imputed values for all pollutants. Ensemble imputation models (Model 8 and 9) gave the lowest RMSE, the highest (IOA) between imputed and real values, and met the minimum requirement criteria using FAC2 for air quality modelling.

The imputation models reproduce high pollution episodes at stations within the clusters where these episodes possibly happened but were not measured, as some of them were captured by the cluster centroids. We also found two important pollutants associated with those episodes: $PM_{2.5}$ and O_3 which may require more measures or should be imputed in different locations for more realistic air quality monitoring.

Supervisors: Beatriz De La Iglesia and Iain Lake, Claire E. Reeves

Acknowledgements

Firstly, I would like to express my sincere gratitude to my primary supervisor, Dr. Beatriz De La Iglesia, for the guidance and encouragement she has provided me during this journey. Without your assistance and dedicated involvement in every step throughout the process, I would never have accomplished this research. It was a true honour to know you and work with you.

Secondly, I would like to show gratitude to Prof. Iain Lake and Prof. Claire Reeves for their support and unlimited advice. I would also like to recognize their thoughtful comments and recommendations that improved the quality of this work and made it more valuable. Thanks for helping me expand my knowledge in environmental science and opening the door for me in this research area.

I gratefully acknowledge the funding received for my PhD from Taibah University, Saudi Arabia, and all the support I received in the UK from the Saudi Cultural Bureau in London.

I would like to thank my parents and sisters for their support, patience, and prayers that have been the source of my strength. Without your love and support, I would not be where I am today.

All the love goes to my little family and my supportive and loving husband, Emad. Even though it was difficult for you to be beside me, your encouragement when times got rough are much appreciated despite the distance. My daughters Mayar, Meral, Jona, and Jwan, who accompanied me on this journey, thank you for being in my life. I dedicate my success to you.

Contents

List of Figures

List of Tables

Abbreviations

ARIMA Auto Regressive Integrated Moving Average.

ASW Average Silhouette Width

AURN Automatic Urban and Rural Network

BCSS Between Clusters Sum of Squares

CH_4 Methane

CO Carbon monoxide

Conn Cluster Connectivity

DAQI Daily Air Quality Index

Defra Department for Environment, Food and Rural Affairs

DI Dunn Index

DM Distance matrix represents the pair-wise distances between objects

DTW Dynamic Time Wrapping, a distance measure for Time-series

EC Elemental carbon

FAC2 Fraction of predictions within the factor of two

Fe Iron

FMD Fusion matrix reporting the fused distance

HC Hydrocarbons

IOA Index of Agreement

LWTs Lamb Weather Types

MAR Missing At Random

MB Mean Bias

MICE Multiple Imputation by Chained Equations

MNAR Missing Not at Random

MVTS multivariate Time-series object

NCC Normalized Cross-Correlation

NMB Normalised Mean Bias

NMHC Nonmethane hydrocarbon

NO Nitrogen oxide or nitrogen monoxide

NO_2 Nitrogen dioxide

NO_X Nitric oxides

OC Organic carbon

O_3 Ozone

$PM_{2.5}$ Particulate matter (PM) that have a diameter of less than 2.5 micrometers or less

PM_{10} particulate matter (PM) that have a diameter of less than 10 micrometers or less

R Coefficient of correlation

RMSE Root Mean Squared Error

SBD Shape-Based Distance, a distance measure for Time-series

SO_2 Sulfur dioxide

SPM Suspended particulate matter

RSPM Respirable particulate matter

Std Standard Deviation

TS Univariate Time-series object

WCSS Within Clusters Sum of Squares

Chapter 1

Introduction

1.1 Background

Air is one of the essential natural resources for humans and for all life on this planet. With the development of the economies throughout the world and population rises in cities, environmental problems involving air pollution have attracted increasing attention. Air pollution is defined as the contamination of the environment caused by some substances called *pollutants* [167]. Nowadays, air pollution is a fundamental problem in several parts of the world. Air pollution becomes one of the world's leading risk factors for death, with seven million deaths per year worldwide attributed to air pollution-related diseases [166]. Sources of air pollution are varied and include anthropogenic sources such as combustion (e.g.in power plants, motor vehicles and residential heating), agriculture and industry as well as natural sources such as vegetation, soils, and lightning [90]. There are several side effects of air pollution on health and the environment. According to Kampa and Castanas [85], there are various air pollutants that negatively affect human health and the environment such as carbon monoxide (CO), particulate matter ($PM_{2.5}$ and PM_{10}), ozone (O_3), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), etc. Air pollutants effects start from minor respiratory irritation to chronic respiratory conditions and lung cancer, acute respiratory conditions in children, heart and lung disease, or asthmatic attacks. Also, long-term exposures to high air pollutant concentration

can cause mortality, and reduced life expectancy [85]. Two factors influence the effect of poor air quality on human health: the ambient concentration of the air pollutants and exposure time [111].

The existing research on air pollution focuses mostly on analysing the effects of air pollutants on human health, meteorological conditions on air pollution, identifying air pollution sources, and predicting or forecasting air quality. The majority of these studies use statistical models [103] or traditional machine learning algorithms such as Support Vector Machines (SVM) [146], and Artificial Neural Networks (ANN) to predict either the air quality index or pollutant concentrations [173]. Other studies use unsupervised learning algorithms such as hierarchical or partitional clustering, to discover new knowledge about air pollutants behaviour and other factors[173]. Ensemble models in machine learning that combine decisions from multiple models promises more accurate predictions for complex data. However, there are limited studies that applied ensemble models to air pollution problem. One opportunity for this is to combine spatial and temporal analysis to predict air quality.

1.2 Motivation

Understanding the behaviour of certain pollutants through air quality assessment can produce improvements in air quality management that will translate to health and economic benefits. However problems with missing data and uncertainty hinder that assessment.

Recently there has been a step-change in the amount of data available for such an analysis, which now includes individual pollutants, air quality data, and meteorological data. One of the available sources of air pollutants data in the UK is the Air Pollutants Monitoring Network. This is the UK's largest automatic monitoring network for air pollution. The Network contains air pollution monitoring stations that record the air pollutant concentrations for the most important pollutants such as nitrogen oxides (NO_x), sulphur dioxide (SO_2), ozone (O_3) and particles (PM_{10}

and $PM_{2.5}$). These pollutants are measured at various monitoring stations and the measured concentrations of each pollutant become a time series (TS) requiring further transformation and analysis to produce air quality assessments. As reported by Defra in 2017 [43], there are 285 air quality monitoring sites across the UK as shown in Figure 1.1, containing several types of networks with different objectives and coverage.

Our focus will be on the automatic monitoring network called Automatic Urban and Rural Network (AURN) maintained by Defra UK. The are 167 AURN stations around the UK during our study period (2015-2018). These networks are automatic and produce hourly pollutant concentrations as well as modelled weather data in some stations. These data are collected and stored, then made directly available via the Internet (`http://uk-air.defra.gov.uk/`).

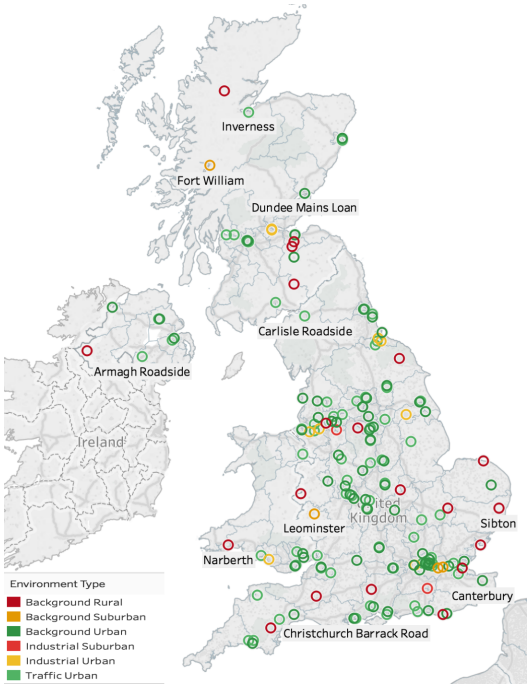


Figure 1.1: Geographical distribution of the air quality monitoring stations at the AURN network. Figure source: Figure developed by the author, information obtained from [7]

The temporal coverage is highly variable in these sites; for some it can go back to 1972 but for others it may only have been collected more recently (i.e 2018).

Spatially, the stations are distributed across regions but concentrated in urban areas within each region (see Table 2.1 in the next chapter). We will explain more about these stations and the data collected in detail in Chapter 2. The data generated in these stations are very complex; they include multivariate time series and spatial-temporal data with many dimensions, multiple outcome variables, coverage varying by location, etc. Adding to that it has many missing values.

Using these data to study air pollution is a priority area for the Government, and the research may provide information and evidence for future air quality interventions [78]. Research on air quality has been reviewed in recent years [114], and there are difficulties in handling air pollution data using basic data mining approaches. Some of these challenges are:

1. Not all the stations report all the pollutants and even if a station does, it may not measure a particular pollutant all the time due to instrument down-time.

2. Each pollutant can be emitted from various sources and be involved in different chemical reactions and so their concentrations exhibit different temporal and spatial distributions. Despite these differences, these pollutants' distributions are often related due to some common or co-location of sources, links between the chemical reactions, and weather conditions that act to either trap, disperse, or transport the pollutants [1].

3. Air quality is affected by multiple factors such as time, location, temperature, humidity, pressure, and rainfall, etc. [36, 133, 173]. These factors affect each other, making the analysis of air quality more variable and complex [11].

4. Air quality monitoring stations have different environment types such as roadside, industrial, and urban, suburban, and rural, making the readings not always generalisable to locations around them.

5. Missing data due to reasons such as failure or servicing, that require appropriately models that can deal with uncertainty.

6. Air quality data is a time series associated with sequential properties, so it is essential to consider the time effect. It has temporal dependencies on its historic values either by its recent hourly values or by the longer historic values.

7. Air quality has also spatial dependencies on the station's geographical properties such as longitude, latitude, and altitude.

8. In the UK, the daily Air Quality Index (DAQI), is forecast based on grid model of 11KM * 11KM, but the actual readings that can be contrasted with the forecast are based on available stations, which may provide inadequate geographical coverage. See Figure 1.1 to observe varying geographical cover of measuring stations. The index used in public communication is based on maximum values and may be misleading.

Together this results in high levels of complex data with many missing values and associated uncertainty. Therefore current air quality assessments are based on high levels of uncertainty resulting from calculating the air quality index even with the absence of some pollutants. This may lead to incorrect policy decisions, with further negative environmental and health consequences [79]. For example, the percentage of missing observations within the collected hourly pollutants observations from 167 stations during the three years reached 16% out of the total number of observations (i.e 13289256 observations), see Table 3.2 in the Chapter 3 for more details. In addition to all missing pollutants observations in case if the pollutants are not measured at all in a station. These challenges have motivated us to explore further how the available data can be used in combination with data mining techniques and time series to enhance the air pollution data available by imputing all missing data.

We are motivated by a real-world problem: the need to build a model to produce plausible imputations for missing measurements for air pollutants which can then assist us to calculate a that is more realistic, taking into account pollution episodes

that may not be measured at a particular station but occurred and where measured at other points. A plausible imputation may translate those episodes to stations that did not measure them and produce a different , which may give a more realistic account of air pollution. It may also be possible to infer where more or less measures may be recommended to reduce uncertainty in air quality assessment.

Our data involves multivariate time series (MVTS), and those are becoming more prominent specially as part of large and complex datasets [55]. The literature review revealed that multivariate time series have received limited attention. Hence we are also motivated by the need to generate modelling techniques for multivariate time series data, where air pollution is an example of such data. Modelling MVTS in the context of missing data and uncertainty is also therefore a desirable research challenge.

1.3 Research Aim and Objectives

The research presented focus on the problem of missing air pollutant concentrations data either because a limited set of pollutants is measured at a monitoring site or because an instrument is not operating, so a particular pollutant is not measured for a period of time.

We want to understand the relation between different pollutants concentrations and their geography. In particular, understanding such relations may enable us to impute missing data (including entire TS) where particular pollutants are not being measured. We postulate that in such cases, pollutants measurements from other stations may act as a proxy measurement for the missing pollutants (TS).

Therefore, our aim is to investigate robust models for estimating the missing values when there are no measurements of a particular pollutant at a site at all.

The following are the key objectives of the proposed work:

1. Obtain and aggregate air pollution data that can be used for better assessment of the air quality.

2. Test multiple imputation or other missing data methods, to impute missing observations within time series (pollutant concentrations) to enhance the data and reduce the uncertainty.

3. Apply univariate time series clustering algorithms on individual air pollutant datasets to cluster data collection sites (stations) in terms of their similarity in pollutant concentrations over time (i.e. temporal similarity). The results should be analysed to improve our understanding of air pollutant concentrations around the UK.

4. Develop and apply multivariate clustering algorithms to cluster stations in terms of their aggregated similarity in all pollutant concentrations over time.

5. Develop imputation models, including ensembles, that are able to impute plausible pollutant concentrations for missing pollutants (whole time series) using either uni-variate or multivariate clustering results.

6. Compare the clustering results from the different approaches to understand how clustering can aid imputation.

7. Evaluate the stability of the proposed models under different weather conditions and station environment types, thereby including some of the additional complexity of the data.

8. Validate the approach by producing new DAQI values that is based on observed and imputed values after imputing all the missing pollutants in stations and then comparing those with the observed DAQI that is based on measured pollutants only. Historical pollution events recorded can then be used to assess the potential of the imputation to capture missed events.

1.4 Scope of Research

The research covers all the UK cities that include AURN stations, and the wider regions (16 regions) as shown in Figure 1.1. These areas incorporate several environmental types that have different characteristics such as urban, rural, traffic and industrial areas. The stations are run by Defra and each station measures all or a combination of the concentrations of ozone (O_3), fine particulates ($PM_{2.5}$ and PM_{10}), nitrogen dioxide (NO_2), and sulfur dioxide (SO_2). In this research, we only focus on the main four pollutants, that influence the pollution level in the UK. These pollutants are: O_3, NO_2, PM_{10}, and $PM_{2.5}$, we ignore sulphur dioxide (SO_2) because the UK emissions of sulphur dioxide have been reduced in the recent decades due to the closure of coal plants and the restrictions on the sulphur content of fuels [66]. The UK meets the current emission ceilings for sulphur dioxide for the period 2010 to 2019. Adding to that, there are few stations in the AURN network that measure SO_2, there are only 28 stations out of 167 stations including in this research, as shown in the next chapter, in Table 2.2.

1.5 Research Questions

The main problem we are focusing on is the imputation of missing pollutant either partially or completely. Pollutants concentrations can be influenced by station location, environmental types, distance to pollutant sources, seasonality, ..., etc, which makes the imputation more challenging.

In developing our research we give some answers to the following questions:

- Does clustering helps to understand pollutant behaviours around the UK and deliver plausible imputation through clustering? (see chapter 4)
- Which clustering approach is more effective for the purpose of generating an imputation: univariate or multivariate time series clustering? (see chapter 5

and 6)

- Does the uncertainty of the missing data impact the clustering results? (see chapter 6)
- What is more important to pollutant imputation in a station, spatial similarity with its neighbours or temporal similarity with other stations? (see chapter 7)
- Which imputation model is able to give more plausible imputed values and how do we evaluate that? (see chapter 7)
- What factors affect imputation model performance? e.g. pollutant behaviour, station location or distance to pollutant sources, station environmental types, weather,.. etc. (see chapter 7)
- What additional measurements (i.e. which pollutant(s)) could help to improve air quality assessment in stations with unmeasured pollutants? e.g. if imputation captures an unmeasured ozone event in a station which appears to be real, should ozone measurement improve at that station? (see chapter 8)
- Which pollutant(s) may have most influence over the DAQI? (see chapter 8)

1.6 Contributions of Research

The research conducted in this thesis has resulted in the following productive contributions to the area of air quality modelling, and multivariate time series analysis:

- We developed a clustering imputation approach to impute whole time series for a missing pollutant in a station using k-medoids clustering. As part of this work, we conducted initial experiments to select the best imputation method and the TS distance measure that works better for our dataset. We also

compared single and multiple imputation methods to impute the missing values within the time series. For whole pollutant imputation, we proposed the following imputation models: Cluster Average (CA), Cluster Medoid (CM), First neighbour (1NN), and the Average of 2NN. This is presented in Chapter 4 and was limited to ozone dataset only, and was published in the Proceedings of Hybrid Artificial Intelligence Systems (HAIS 2020) [13]. Based on decisions taken in this work, we developed a univariate time series clustering and imputing framework based on the k-means clustering algorithm and Shape based distance (SBD) as similarity measure. This approach is based on individual pollutant clustering and imputation. This is presented in in Chapter 5.

- We developed a multivariate time series (MVTS) clustering and imputation framework based on a fusion approach that aggregates the similarity/dissimilarity of the univariate TS (pollutants) between every two MVTS (stations). This is presented in Chapter 6 and was published as a journal paper to the Neurocomputing journal [14].

- We proposed nine imputation models to impute the whole time series including the clustering solutions that represent the temporal similarity, spatial similarity, and ensembles approaches that aggregate the temporal and spatial similarity between stations. We also evaluated these models based on statistical and graphical model evaluation functions to select the best imputation model for each pollutant. This is presented in Chapter 7 and was published as a journal paper to Geoscientific Instrumentation, Methods and Data Systems [12].

- We conducted detailed analysis on the performance of the best imputation models (Median to impute PM_{10} and $PM_{2.5}$ and Median.ENV to impute O_3 and NO_2) for air quality assessment in stations with one or more unmeasured pollutants. From that, we produced an enhanced version of the air pollution

dataset that is used to calculated a new DAQI based on all four pollutants, which helps to identify where more measurements for specific pollutants may be beneficial. This is presented in Chapter 8 and will be prepared for submission as a Journal article.

1.7 Thesis Outline

The reminder of this thesis is organised as follows:

- Chapter 2: **Literature Review** provides a review of air quality assessment in the UK and some related problems from the perspective of data mining and time series analysis.
- Chapter 3: **Research Methodology Design** presents our proposed methodologies and tools used to achieve the aims of this research.
- Chapter 4: **Initial Exploratory Experiment in Clustering Imputation for Air Pollution Data** presents our initial exploratory experiment in clustering imputation approach using ozone dataset only.
- Chapter 5: **Results of Applying Univariate Time Series Clustering and Imputation** presents the results of applying univariate time series clustering to impute individual pollutants.
- Chapter 6: **Results of Applying Multivariate Time Series (MVTS) Clustering and Imputation** presents the results of applying the proposed multivariate time series clustering for imputation.
- Chapter 7: **Evaluation of the Imputation Models** includes the evaluation results for imputation models using the MVTS clustering approach.
- Chapter 8: **Model Application** includes the application of our imputation models to a real air pollution dataset, including an evaluation of DAQI values for real and imputed data.

- Chapter 9: **Conclusions and Further Work** presents conclusion of our research and suggested further work.

Chapter 2

Literature Review

This chapter reviews the relevant background materials for this research. We review the problems related to air quality from the perspective of data mining, machine learning techniques and time series analysis. We also present a review of the proposed data mining solutions used in air quality assessment and prediction.

This chapter is organised as follows: Section 2.1 includes an introduction to air quality and how it is measured in the UK; Section 2.2 is a general introduction into the data mining techniques, focusing more on partitioning clustering techniques and evaluation measures; Section 2.3 is a general introduction to time series analysis and the most common distance measures; Section 2.4 discusses missing data imputation and its evaluation techniques; Sections 2.5 and 2.6 present an extensive review of the proposed data mining, machine learning approaches, and time series analysis that have been applied to air quality data; Section 2.7 introduces some of the previous research limitations and challenges related to air quality research; finally, Section 2.8 includes a summary of this chapter.

2.1 Air Quality

The quality of air is negatively affected by some particles and gases that harm human health. Also, it is affected by several factors including location, time, and

uncertain variables [86]. The air pollutants can be classified as either primary or secondary. A primary pollutant is a pollutant that is emitted directly from its sources, e.g. carbon monoxide gas is directly emitted from the motor vehicle exhaust. In contrast, secondary pollutants are not emitted directly, but are formed by reaction or interaction to other pollutants for example ozone [114].

Air quality measures are based on pollutant's concentrations in the air; when the pollutants are raised to a high level, the air quality gets low. We have an essential measure called air quality index (AQI) which quantifies air quality in a region for a given period of time. This index is used by the Government agencies to communicate to the public how polluted the air is currently. According to Defra, the UK's air quality is measured using the Daily Air Quality Index (DAQI) based on the air pollutants concentrations recorded by the monitoring network around the country. This index represents the air quality by numbers from 1-10. Those numbers negatively represent the air quality, meaning that the highest index represents the worst air quality. Countries have different criteria for their air quality index, Section 2.1.3, addresses some of the air quality indices that are used around the world.

2.1.1 Air Pollutant Monitoring Network in the UK

The Air Pollutants Monitoring Network contains air pollution monitoring stations that record the concentrations of the air pollutant. This network has many stations around the UK. Many, however, use obsolete sensors and/or have been dismissed. The most reliable measurements are available from a network of 167 stations, which record the hourly concentration of air pollutants and modelled weather data at some locations. These are collectively called the Automatic Urban and Rural Network (AURN). As we said previously, there are 285 air quality monitoring sites across the UK, but our focus will be on the 167 AURN stations as shown in Figure 1.1. The UK has 16 regions, that are divided into 43 zones for air quality assessment. There are 28 agglomeration zones and 15 non-agglomeration zones, for reporting

and monitoring air pollution. The non-agglomeration zones match regional boundaries in England, Scotland, Wales and Northern Ireland. These regions and their monitoring stations are shown in Appendix A, Table A.1 and Figure A.1.

2.1.2 Automatic Urban and Rural Network (AURN)

There are different AURN stations based on environmental types such as typology (suburban, urban) and predominant emission sources (background, traffic, and industrial). Table 2.1 shows the number of stations under each environmental type. Each type of air quality monitoring has a specific scope; background stations are located in areas where pollution levels are not highly influenced by any single source or street; industrial stations are located nearby industrial areas to measure the pollution emitted by factories; roadside stations are located near the traffic (roads, motorways, highways) to measure the emission there; rural stations are located in small settlements with natural ecosystems or forests. As well as that, Urban/Suburban stations are located where there is more population, and they focus more on measuring ozone [6].

AURN stations are automatic, in that they produce hourly pollutant concentrations. Pollutant concentration and other data are collected from individual site (i.e AURN) by modem, which is used for data communication between the remote central station and the site logger that records data over time [21]. Data from the AURN is collected hourly from each site. Other modelled meteorological parameters may include modelled temperature, modelled wind direction, and modelled wind speed. AURN uses modelled meteorological parameters that are generated by some models at the station as a useful alternative to commercially available meteorological observations which as those may be measured some distance from the local air quality monitoring station [22]. As noted, not all stations report all parameters, as this mainly depends on the purpose of the monitoring station. Even when a station reports a particular parameter, it may not always be recorded, resulting

in high levels of missing data. These data, along with archived data from current and defunct monitoring sites, is then made directly available via the Internet [6].

Table 2.1: Number of AURN station based on environmental type.

Station environment type	**Number of stations**
Background Rural	20
Background Urban	62
Background Suburban	4
Industrial Urban	9
Industrial Suburban	2
Traffic Urban	70
Total number of station	167

The main objective for the AURN stations is to monitor the air pollutant concentrations that are used to asses air quality; those five air pollutants mentioned earlier used to calculate the air quality index in different areas. However, one of the challenging parts of our research is that not all the AURN stations measure the same set of pollutants. Table 2.2 shows the number of stations that measure each pollutant. There are only 16 out of 167 stations that measure all five pollutants, and 26 stations that measure the four main pollutants that we focus on, in this research (O_3, NO_2, PM_{10}, and $PM_{2.5}$). The majority of the stations measure NO_2, followed by those that measure particulate matter $PM_{2.5}$ and PM_{10}, then O_3.

Table 2.2: Number of AURN stations that measure each pollutant .

Pollutant	**Number of stations**
O_3	70
PM_{10}	75
$PM_{2.5}$	77
SO_2	28
NO_2	157
O_3, NO_2, PM_{10}, $PM_{2.5}$, and SO_2	16
O_3, NO_2, PM_{10}, and $PM_{2.5}$	26

2.1.3 Air Quality Index and Forecast

In general, the Air Quality Index (AQI) is an important indicator for the public to understand the air quality and when it may have an impact on their health. It is an indicator of air pollution levels, based on the highest concentration of air pollutants such as carbon monoxide (CO), nitrogen dioxide (NO_2), sulphur dioxide (SO_2), ozone (O_3), particles $< 2.5\,\mu m$ ($PM_{2.5}$), and particles $< 10\,\mu m$ (PM_{10}).

An air quality forecast is an effective way to provide early warnings when the air pollutants are predicted to be higher than normal levels. This has had great attention recently due to the detrimental effects of air pollution on human health and the environment [89]; as a result, several countries have an early warning system to report the air quality. For example, the ir Quality Index (AQI) is used in the USA, China, and India, the Air Pollution Index (API) is used in Malaysia, the Common Air Quality Index (CAQI) is used in Europe, and the Daily Air Quality Index (DAQI) is used in the UK. The DAQI will be our main focus in this study. These air quality indices are different in the set of the monitoring pollutants they use, the pollution standards, and the index categories. The AQI used in USA and China monitors six pollutants, namely, PM_{10}, $PM_{2.5}$, NO_2, SO_2, CO, and O_3. The AQI transforms the air pollutant concentrations into scores from 0 to 500, divided into six levels of air quality (Excellent, Good, Lightly Polluted, Moderately Polluted, Heavily Polluted, and Severely Polluted) [124]. These levels reflect the air quality impact on human health. However, there are differences in their standards; for example, the maximum 24-hour average for $PM_{2.5}$ exposure is 35 μgm^{-3} in the USA, and 75 μgm^{-3} in China [17]. Malaysia Air Pollution Index (API) is based on the United States' Environmental Protection Agency (USEPA). It focuses on the same set of air pollutants, according to related documentation [19]. The AQI in India uses the same AQI quality levels, but it monitors eight air pollutants: PM_{10}, $PM_{2.5}$, NO_2, SO_2, CO, O_3, Ammonia NH_3, and Lead Pb, according to the relevant documentation [18]. Finally, the Europe Common Air Quality Index (CAQI) uses

different air pollutants concentration scores; it is from 1 to 100, associated with five levels of air quality called: Very Low, Low, Medium, High, and Very High. Also, it uses some mandatory components for the index (NO_2, PM_{10} and O_3) and optional pollutants ($PM_{2.5}$, CO and SO_2) and based on the environment type either it is background or traffic, according to the available documentation [31]. Canada Air Quality Health Index (AQHI), uses the combination of pollutants of O_3, $PM_{2.5}$ and NO_2 to determine the final index. The air quality health-related risks divided into a scale of 1 to 10+ with four categories of health risks as Low, Moderate, High, and Very High [124].

The common rule between all these air quality indices is that the final air quality index is determined by the highest level among the indices of individual pollutants.

2.1.4 Daily Air Quality Index (DAQI)

In the UK, the Daily Air Quality Index (DAQI) represents air pollution levels in the UK. This index is reported based on the highest individual DAQI derived for each of the five major air pollutants (O_3, NO_2, PM_{10}, $PM_{2.5}$, and SO_2) based on their concentrations. If concentration data for not all of these pollutants are available, the DAQI is based those pollutants for which data are available. The DAQI is used to provide an indication of the air quality, and some associated information that may used by at-risk groups as well as the general population [5]. The DAQI is numbered from 1 to 10, and divided into four bands; 'low' (1–3), 'moderate' (4–6), 'high' (7–9) and 'very high' (10). The air quality is negatively correlated with DAQI index, meaning that a higher DAQI index represents worse air quality. Figure 2.1 shows the DAQI indexing levels and related air pollutant concentrations according to Conolly et al. [41]. Table 2.3 explains each of these air pollutants and how they are measured. Each pollutant is reported hourly, but the hourly calculation may be measured differently according to Defra [43]. For example, as hinted in Figure2.1 ozone is measured hourly from a running 8 hourly mean, nitrogen dioxide is measured from an hourly mean, sulphur dioxide is measured

hourly based on the maximum 15 minute mean, $PM_{2.5}$ and PM_{10} is measured hourly based on the running 24-hour mean. All measurements are given in μgm^{-3} and reported hourly.

DAQI forecasts are issued on a national scale in the UK; they are produced by the Met Office in the morning for the current day as well as for the next four days. The forecast is improved by incorporating the recent observations of air quality recorder at the AURN stations.

Band	Index	Ozone Running 8 hourly mean µgm^{-3}	Nitrogen Dioxide hourly mean µgm^{-3}	Sulphur Dioxide 15 minute mean µgm^{-3}	$PM_{2.5}$ Particles (EU Reference Equivalent) 24 hour mean µgm^{-3}	PM_{10} Particles (EU Reference Equivalent) 24 hour mean µgm^{-3}
Low	1	0-33	0-67	0-88	0-11	0-16
	2	34-66	68-134	89-177	12-23	17-33
	3	67-100	135-200	178-266	24-35	34-50
Moderate	4	101-120	201-267	267-354	36-41	51-58
	5	121-140	268-334	355-443	42-47	59-66
	6	141-160	335-400	444-532	48-53	67-75
High	7	161-187	401-467	533-710	54-58	76-83
	8	188-213	468-534	711-887	59-64	84-91
	9	214-240	535-600	888-1064	65-70	92-100
Very High	10	241 or more	601 or more	1065 or more	71 or more	101 or more

Figure 2.1: Daily Air Quality Index. Source:[41]

It is important to note that pollutants have different behaviours. PM has many varied sources, both primary (emitted directly into the atmosphere) and secondary (produced in the atmosphere via chemical and physical processes). Whilst PM concentrations are often greater at the roadside [2], the particles can have lifetimes of several days in the atmosphere, meaning that they can be distributed widely. The larger particles are subject to greater loss via sedimentation, so $PM_{2.5}$ is more evenly distributed than PM_{10} [8].

Particulate matter, which includes soot and dust comes from different sources that burn fossil fuels such as metal processing, and traffic; this includes $PM_{2.5}$ and PM_{10}. Both can travel large distances in the atmosphere [8]. According to Defra [2], the concentrations of PM at the locations near to roadsides are much higher than those in background locations. Also, $PM_{2.5}$ has a longer lifetime so is more widely spread and distributed more evenly [4].

Table 2.3: Air pollutants explanation.

Air pollutants	Detailed Explanation	Measure
$PM_{2.5}$	$PM_{2.5}$ Particles are air pollutants, that refers to atmospheric particulate matter (PM) that have a diameter of less than 2.5 micrometers or less. These air pollutants come from different sources that burn fossil fuels such as metal processing, and traffic.	The daily mean concentration, latest 24 hour running mean for the current day.
PM_{10}	PM_{10} Particles are air pollutants, that refers to atmospheric particulate matter (PM) that have a diameter of less than 10 micrometers or less.	The daily mean concentration, latest 24 hour running mean for the current day.
SO_2	Sulphur Dioxide is a toxic gas with a burnt match smell. It is produced by-product of the burning of fossil fuels, coal, oil and many industrial processes.	The 15-minute mean concentration.
NO_2	Nitrogen dioxide is one of a group of highly reactive gases. Its primary source comes from the burning of fuel such as cars, trucks and buses, power plants, and off-road equipment.	The hourly mean concentration.
O_3	Ozone is a pale blue gas with pungent smell, it makes 0.6 ppm of the atmosphere. Also, ozone is not emitted directly into the atmosphere, but is a secondary pollutant generated from the reaction between nitrogen dioxide, hydrocarbons and sunlight.	The running 8-hourly mean.

The primary source of NO_2 comes from fuel burning such as cars, trucks and buses, power plants, and off-road equipment. However the NO_2 concentrations are influenced by the traffic density, road locations, and meteorological conditions, which cause variation of NO_2 concentration from one roadside location to another. Adding to that NO_2 is shorter lived than other pollutants and shows greater spatial variability, with concentrations being strongly influenced by the environment type (e.g. roadside, urban background, rural). This gives NO_2 a local pattern, that changes from one location to another based on the environmental type [34].

Ozone is complex as is not directly emitted into the air, but it is formed as a secondary pollutant by rough chemistry involving nitrogen oxides (NO_x), the sum

of NO_2 and nitric oxide (NO) and volatile organic compound (VOC) in the presence of sunlight [48]. So, the ozone formation depends on the VOC–NO_x ratio [107]. This chemistry is non-linear and newly emitted NO can reaction with O_3 leading to reductions in O_3 concentrations close to sources of NO (e.g. in urban ares and in particular close to roads). Urban areas are often lower in ozone than those in rural areas [71], due to this chemistry.

O_3 and NO_2 are strongly anti-correlated, indicating that the O_3 is strongly depressed by high NO_x [100, 62]. Furthermore, ozone can have a lifetime of days to weeks [93], meaning that ozone at a specific site may have been produced by NO_x and VOCs emitted from other distant locations.

2.2 Introduction to Data Mining

This section is a general introduction to data mining techniques, focusing more on clustering algorithms and it's evaluation metrics.

Data mining is the process of discovering hidden knowledge from large datasets, using a set of techniques, methods and algorithms to help decision makers [150]. Data mining have been applied to several fields such as sciences [37], medicine [92], social sciences [77], healthcare [120], and many others. There are many tasks to which data mining algorithms can be applied. Data mining tasks can be classified into two broad categories: predictive and descriptive tasks. Predictive task are used to forecast unknown or future values; this includes classification, regression and time series analysis. Descriptive tasks are used to discover and describe new information and patterns from the available dataset; this includes clustering, association, and summarising [172].

2.2.1 Classification and Regression

Classification is a supervised learning method used in data mining to predict a class label. There is an implication therefore that the data under analysis has been labelled (by a 'supervisor' hence the name of supervised learning) with a categorical label. The classification algorithms learns first from the training dataset which is a labelled dataset; a model is generated which can predict the class label based on the input of other variables. A new data point can then be classified based on the training set model [142] in order to test the accuracy/goodness of the model; labels are also required for the so-called test set. However, once the model has been accepted as of good quality according to its evaluation on the test set, it can be used on new data without labels to produce a prediction. Regression uses a similar process to classification, but instead of predicting the class label, the algorithm predicts a numeric value associated with the data item. There are many methods and techniques used for classification such as k-nearest Neighbour (KNN) [147], Support Vector Machines (SVMs) [173], Decision Trees (DT) and Random Forest [145], Naive Bayes algorithms [65], and Neural Networks (NNs) [173].

2.2.2 Clustering

Clustering is an unsupervised technique used in knowledge discovery to explore new relations between data items. Its main task is to group a set of data items (observations) into related groups or clusters based on some similarity or dissimilarity measure without having any previous knowledge about the class labels [75]. There are different types of clustering techniques have been proposed for cluster analysis. However, we may distinguish three main types of clustering techniques: hierarchical, fuzzy, and partitional.

Hierarchical clustering seeks to build a hierarchy of clusters/groups. The final result of the hierarchical clustering is a tree-based representation of the objects called a dendrogram [135].

Fuzzy clustering divides the data points into partitions/clusters and allows data points to belong to more than one cluster [32].

Partition clustering, in general, divides the data points into K non-overlapping partitions/clusters (the number of clusters K has to be pre-specified). Given a set of N unlabeled data points, a partitioning clustering constructs K partitions each partition is a cluster. The most heuristic methods of partitioning algorithm are the k-means and k-medoids. In the k-means algorithm [105], each cluster is represented by the mean value of its objects, while in the k-medoids the cluster is represented by one of the object in the cluster [87]. It has also been suggested that partitioning algorithms are somewhat less sensitive to outliers than hierarchical clustering algorithms [122].

One of the main components of a clustering algorithm is a measure of distance or similarity/dissimilarity between objects. The similarity is the measure that reflects the strength of association between the data objects. There are a variety of clustering distance measures. The most commonly used distance measures are the Euclidean distance (ED) and Manhattan Distance (M), as known as a city block distance. The Euclidean distance (ED) is the ordinary distance between two data points and is shown in Equation. 2.1. It represents the straight line distance between two points in a space, while in Manhattan Distance (M), the distance is the sum of the absolute differences of their cartesian coordinates, as shown in Equation. 2.2; in other words, it is the total sum of the difference between the x-coordinates and y-coordinates [118]. Since the similarity is fundamental of the clustering process, the distance measure must be chosen carefully [83].

$$ED(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2.1}$$

$$M(X,Y) = \sum_{i=1}^{n}|(x_i - y_i)| \tag{2.2}$$

In these equations, $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots y_n)$ represent n-dimensional vectors of data points for objects X and Y, and n can be the number of data points in each object.

In this research, we investigate the k-medoids and the k-means clustering algorithms with air pollution data. Hence, in the following sections, we focus on the process and the advantages of these clustering algorithms.

2.2.2.1 K-means Clustering Algorithm

The k-means is one of the most widely used clustering algorithms, proposed by McQueen et al. [105]. K-means starts by randomly choosing K centroids from the dataset, then assigns all the data points to the closest centroid based on the selected distance metric. This will create K clusters. Then, centroids are recalculated for each cluster using the mean of all the data points in the cluster. These two steps will keep repeating until there is no change in the cluster centroids or the iteration reaches its maximum. K-means attempts to minimise within-cluster distance and maximise between-cluster distance. The general schema of the k-means algorithm is as follows:

1. Randomly select K data point as initial centroids.
2. Assign each data point to the closest centroid based on the selected distance metric.
3. Calculate the centroid of each cluster by averaging all the data points within the cluster.
4. Repeat steps 2 and 3 until there is no change in the cluster centroids or the iteration reaches its maximum.

2.2.2.2 Partitioning Around the Medoids (PAM/k-medoids)

The k-medoids, also called Partitioning Around Medoids (PAM) was proposed by Kaufman and Rousseeuw [87]. The cluster medoids are the cluster centers, which are the most representative objects of a cluster. The average dissimilarity between medoids and all data points in the cluster is minimal. The concept of cluster medoids is similar to cluster centroids, but medoids are always members of the data set whereas centroids may not correspond to real objects. Medoids are not necessary located in the center of a cluster.

This algorithm requires the number of clusters K to be known. K-medoids starts randomly by selecting K data points that represent the initial cluster medoids, then assigns each data point to the nearest medoid based on the distance metric. After assigning all the data points to clusters in the first iteration, it evaluates the clusters by calculating the total sum of distances between all data points and its medoid. In the next iteration, it tries to decreases the total sum of distances in each cluster by swapping the non-medoid data point with the medoid. The iterations stop, when there is no change happening. For each cluster, there is only one medoid, which is the data point with lower sum of square error to other data points within the cluster [160]. The general schema of the K-medoid algorithm is as follows:

1. Randomly select K data points as initial medoids.
2. Assign each data point to the closest medoid based on the distance metric.
3. Calculate the total cost of the cluster, which is the average dissimilarity of the cluster medoid to all data points in the cluster.
4. Calculate a swapping cost between each data point and its medoid. If the swapping between the data point and the cluster medoid decrease the total sum of distances within the cluster, the swap will be confirmed; otherwise, the medoid will not change.

5. Repeat steps 2, 3, and 4 until there is no change in the medoids assignments or the iteration reaches its maximum.

The main advantage of PAM/k-medoids over other partitioning algorithm such as k-means is that it is more robust and it deals with outliers and noisy data [87]. Also, Clustering LARge Applications (CLARA) [87] and Clustering Large Applications based on RANdomized Search (CLARANS) are two improved versions of k-Medoid algorithm for handling very large data based on a sampling method.

2.2.3 Clustering Evaluation Measures

There exists a wide range of cluster quality measures to evaluate the clustering solutions to select clusters that are compact and well separated. These measures are called Cluster Validity Indices (CVI). The compactness is a cluster homogeneity measure that reflects how close are the objects within the cluster by measuring the intra-cluster (within-cluster variation), while the separation is the degree of separation between clusters. It measures how well separated a cluster is from other clusters by measuring the inter-cluster (between-cluster variation) [44].

Cluster quality measures may be external, internal or relative [126]. The basis of the external and internal validity indices are statistical testing, while the relative validity indices are based on relative criteria and do not involve statistical tests [73]. We will discuss each category of these indices in more details in the following sections.

Notation: Let us define a dataset D that contains N objects D=x_1, x, ..., x_N. A partition or clustering in D is a set of K separated clusters C=c_1, c, ..., c_K. C_c is a set of those clusters' centers, where C_c=C_{c1}, C_c, ..., C_{cK}. Similarly, the center of the data set is the mean vector of the whole dataset D_c, where $D_c = \frac{1}{N}\sum_{i=1}^{N} x_i$. To measure the distance between two objects x_i and x_j in the clustering process,

using any distance measure is denote $d(x_i, x_j)$. P is a previously defined partitions (ground truth) of a dataset D.

2.2.3.1 External Validation Index

External measures require prior data or ground truth data (the optimal clusters) to evaluate a clustering algorithm's results. Among the most well-known external criteria are the Rand Index [125], the Jaccard coefficient [82], and Fowlkes-Mallows [60].

To evaluate the clustering solution C based on P, these external indices are based on comparing the placement of each two objects (x_i and x_j) in C and P, using the frequency of placement each pairs x_i and x_j considering the following cases:

- a**:** Number of object pairs assigned to the same group in C and P;
- b**:** Number of object pairs assigned to the same group in C, but different group in P;
- c**:** Number of object pairs assigned to to different groups in C, but the same group in P;
- d**:** Number of object pairs assigned to different groups in C and P.

Now we follow with the definition of the various indices:

1. **Rand Index (RI):** The Rand Index was proposed by Rand [125]. This index measures the similarity between two partitions C and P. The value of this index is within the range $[0,1]$, where 1 indicates that these two partitions are very similar and 0 means partitions are very different. RI is defined as:

$$RI = \frac{a+d}{a+b+c} \tag{2.3}$$

2. **Jaccard Index (J):** the Jaccard Index was proposed by Jaccard [82]. The value of this index is within the range [0,1] a value closer to 0 indicates different partitions and a value closer to 1 indicates similar partitions. J index is defined as:

$$J = \frac{a}{a+b+c} \tag{2.4}$$

3. **Fowlkes-Mallows Index (FM):** The Fowlkes-Mallows Index was proposed by Fowlkes and Mallows [60]. The value of this index defines the similarity between C and P, which means higher values indicate higher similarity and vise versa. FM index is defined as:

$$FM = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}} \tag{2.5}$$

2.2.3.2 Internal Validation Index

Internal measures are based on intrinsic properties of the clustering solution such as compactness, separation, and connectedness of the cluster partitions, so they are based on measurable aspects of a clustering solution [20].

First, we need to identify two types of distance that are used to measure the compactness and the separation of the clusters. The inter-cluster and intra-cluster distance. In the following definition:

intra-cluster distance:
This distance measure is used to evaluate the compactness of a cluster. It is the distance between objects within a cluster that reflects the cluster variation. A good clustering solution should have high intra-cluster similarity, which means low variance among the cluster members.

inter-cluster distance:
This distance measure is used to measure the separation between clusters. A good

clustering solution should have low inter-cluster similarity, which means high variance among different clusters.

Internal validation indices based on these criteria are the Silhouette index (Sil), Dunn index (DI), the Davies–Bouldin index (DB), the SD index and SDbw index. More details on these indices are given below:

- **Silhouette Width (ASW)**:

  The Silhouette Width measure was proposed by Rousseeuw [127]. It combines the compactness and separation of the cluster into a single score. Silhouette width is a measure to assess how well an object fits into a cluster rather than its neighbour cluster by measuring the average distance between an object and all other objects within the same cluster, and the average distance between an object and other objects in its neighbour cluster. The higher silhouette average indicates a good clustering solution. This measure is one of the measures that helps to select the optimal number of clusters and also to evaluate the clustering results. The maximum of the Silhouette index over a range of possible values for K indicates the optimal number of clusters and a good clustering solution.

  For each data point/object x_i, the silhouette width Sil_i is calculated as described in [127]:

  1. Calculate the average dissimilarity between x_i and all other data points in the same cluster of x_i (let's say c_i).

  $$a_i(x_i) = \frac{1}{|c_i|} \sum_{y_o \in c_i, o=1}^{|c_i|} d(x_i, y_o) \tag{2.6}$$

  Where $|c_i|$ is the number of objects in cluster c_i.

  2. Calculate $d(x_i, c)$, which is the minimum average dissimilarity between x_i and all other data points that belong to other clusters c. Then, select

the cluster that gives the minimum average dissimilarity to x_i (called the neighbour cluster of x_i), defined as:

$$b_i(x_i) = \min_{c_j \in C, c_i \neq c_i} \{\frac{1}{|c_j|} \sum_{y \in c_j} d(x_i, y)\} \tag{2.7}$$

3. Finally the silhouette width of object x_i is defined by the formula:

$$Sil(x_i) = \frac{b_i(x_i) - a_i(x_i)}{max(a_i(x_i), b_i(x_i))}. \tag{2.8}$$

The Average Silhouette Width (ASW) for a clustering solution C is defined as:

$$ASW(C) = \frac{1}{N} \sum_{c_i \in C} \sum_{x_i \in c_i} \frac{b_i(x_i) - a_i(x_i)}{max(a_i(x_i), b_i(x_i))} \tag{2.9}$$

or,

$$ASW(C) = \frac{1}{N} \sum_{i=1}^{N} Sil(x_i) \tag{2.10}$$

- **Dunn index (DI)** :

The Dunn index was introduced by Dunn [53]. It is the ratio of the smallest inter-cluster distance between two objects from different clusters to the largest intra-cluster distance [54]. It uses to identify those cluster sets that are compact and well separated. Dunn index has a value between zero and ∞, this value should be maximized to represent a good clustering solution. The Dunn index is defined as:

$$DI = \min_{1 \leq i \leq K} \left\{ \min_{i+1 \leq j \leq K} \left\{ \frac{dist(c_i, c_j)}{\max_{1 \leq l \leq K} diam(c_l)} \right\} \right\} \tag{2.11}$$

Where $dist(c_i, c_j)$ is the minimal distance between two objects that belong to different clusters, which is the inter-cluster distance for cluster c_i and c_j, it is defined as:

$$dist(c_i, c_j) = \min_{x_i \in c_i, x_j \in c_j} d(x_i, x_j) \tag{2.12}$$

Where x_i is an object in cluster c_i and x_j in cluster c_j.

$diam(c_l)$ is the maximum distance between two objects in the same cluster, which is the intra-cluster distance, for a cluster c_l, it is defined as:

$$diam(c_l) = \max_{x_i, x_j \in c_l} d(x_i, x_j) \tag{2.13}$$

- **Davies-Bouldin index (DB)**:

  The Davies-Bouldin index (DB) was proposed by Davies and Bouldin [44]. It is based on the ratio of a measure of the between-cluster and within-cluster distances. A lower DB index indicates better separation between the clusters. DB index is defined as:

$$DB = \frac{1}{K} \sum_{i=1}^{k} \max_{j=1,\ldots,k, i \neq j} \frac{diam(c_i) + diam(c_j)}{dist(c_i, c_j)} \tag{2.14}$$

  Where $d(C_{ci}, C_{cj})$ is the distance between the center of the cluster c_i and c_j, and $diam(c_i)$ and $diam(c_j)$ are the cluster diameters, which is the average distance between the cluster objects and its center.

- **SD Validity Index**:

  The SD index is proposed by Halkidi et al. [74]. It defines two quantities: the average scattering $Scatt$ of the objects within the clusters and the total separation between clusters Sep. SD validity index is defined as:

$$SD = Sep(C_{max}).Scatt(C) + Sep(C) \tag{2.15}$$

  $Scatt(C)$ is the average scattering for the clusters, which is the measure of the homogeneity of the objects within the clusters. $Scatt(C)$ is defined as:

$$Scatt(C) = \frac{1}{K} \sum_{i=1}^{k} \frac{||\sigma(c_i)||}{||\sigma(D)||} \tag{2.16}$$

  Where $\sigma(D)$ is the variance of a data set, and $\sigma(c_i)$ is the variance of cluster c_i.

The variance of a data set D is the sum of the squared distances between the centroid of the data set D_c and each objects in the data set x_i, $\sigma(D)$ is defined as:

$$\sigma(D) = \sum_{i=1}^{N} \sigma(D_c, x_i)^2 \tag{2.17}$$

The total variation of a cluster c_i is defined as:

$$\sigma(c_i) = \sum_{j=1}^{|c_i|} \sigma(C_{ci}, x_j)^2 \tag{2.18}$$

$Sep(C)$ is the total separation of the clusters or the between-cluster distance, it is based on the maximum and the minimum distance between cluster centers. $Sep(C)$ is defined as:

$$Sep(C) = \frac{\max_{i\,j}(d(C_{ci}, C_{cj}))}{\min_{i\,j}(d(C_{ci}, C_{cj}))} \sum_{i=1}^{K} \left(\sum_{j=1, j\neq i}^{K} d(C_{ci}, C_{cj}) \right)^{-1} \tag{2.19}$$

Where $\min_{i\,j}(d(C_{ci}, C_{cj}))$ is the minimum distance between cluster centers, $\max_{i\,j}(d(C_{ci}, C_{cj}))$ is the maximum distance between cluster centers among all the K clusters centers (C_c).

$Sep(C_{max})$ in SD is a weighting factor that is equal to the separation of the clusters in case of maximum number of clusters. Lower value of SD indicates better compacted and separated clustering solution.

- **SDbw Validity Index**:

  SDbw validity index is an enhancement index for SD validity index [73]. In this index the density of the clusters is considered instead of the weighting factor $Sep(C_{max})$. Lower value of SDbw indicates dense and well separated clusters. SDbw index is defined as:

$$SDbw = Scatt(C) + densBW(C) \tag{2.20}$$

  $densBW$ represents the inter-cluster density, that indicates the average number of points in the region among clusters in relation with density of the

clusters in the clustering solution. A small value of $densBW(C)$ indicates well-separated clusters. $densBW$ is define as follows:

$$densBW = \frac{1}{K(K-1)} \sum_{i=1}^{K} \sum_{j=1, j \neq i}^{K} \frac{density(u_{i,j})}{max\{density(C_{ci}), density(C_{cj})\}} \quad (2.21)$$

$$density(u) = \sum_{l=1}^{n_{ij}} f(x_l, u) \quad (2.22)$$

where n_{ij} is number of tuples that belong to clusters c_i and c_j. The value of $f(x, u)$ is equal to 0 when the distance between point x anf u ($d(x, u)$) is larger than the average standard deviation of the clusters, and 1 otherwise.

$$stdev = \frac{1}{K} \sqrt{\sum_{i=1}^{K} \parallel \sigma(C_{ci}) \parallel} \quad (2.23)$$

- **Connectivity (Conn)**:

  The connectivity measure was introduced by Chen and Wang [35] to measure the degree to which neighbouring objects have been placed in the same cluster by calculating penalties for each object. The connectivity for clustering solution C, that contains K separated clusters gives a value between zero and ∞. Thus, connectivity should be minimized.

  The connectivity for clustering solution C can be computed by:

$$Conn(c) = \sum_{i=1}^{N} \sum_{j=1}^{L} P_{i,nn_{i(j)}} \quad (2.24)$$

  where L is a parameter giving the number of nearest neighbours to use (L is user selectable, we used L=10 in our calculation), $nn_{i(j)}$ is the jth nearest neighbour of object i, and $P_{i,nn_i(j)}$ are the penalties of an object i to its jth neighbour. It can be zero if i and j are in the same cluster and $\frac{1}{j}$ otherwise.

Additional clustering quality measurements:

- **Within-Cluster Sum of Squares (WCSS).**

  Within Cluster Sum of Squares (WCSS) measures the compactness based on the distance between all the objects within a cluster to the cluster's center; this is one way to calculate the intra-cluster distance. To calculate WCSS, we first calculated the squared fused distance between a cluster center and all objects within the cluster. Then, sum these distances for each cluster for the clustering solution. The within cluster sum of square across all clusters is the measure of the cluster variation. A within-cluster sum of square for cluster c_j, is calculated as:

$$WCSS(c_j) = \sum_{i=1}^{|c_j|} d(x_i, C_{cj})^2 \tag{2.25}$$

  where x_i is an object of the cluster c_j.

  The total within cluster variation for a clustering solution is the average of the clusters variation, it is calculated as:

$$WCSS(C) = \frac{1}{K} \sum_{i=1}^{K} WCSS(c_i) \tag{2.26}$$

$$WCSS(C) = \frac{1}{K} \sum_{j=1}^{K} \sum_{i=1}^{|c_j|} d(x_i, C_{cj})^2 \tag{2.27}$$

  In general, WCSS measures the compactness, cohesion or homogeneity within each cluster. A cluster that has a small sum of squares (WCSS) is a cluster with small variation, which is more compact than a cluster that has high variation.

- **Between-Cluster Sum of Squares (BCSS).**

  Between Clusters Sum of Squares measures how far apart the centers of the clusters in the final partition are from one another. It is a measure of clusters heterogeneity. This measures the average distance between all centers.

The BCSS for a single cluster center is the sum of all distances from that cluster center to all other cluster centers. The average of all BCSS from all the clusters represents the total cluster BCSS.

The BCSS for cluster c_j is calculated as:

$$BCSS(c_j) = \sum_{i=1, j \neq i}^{K} d(C_{ci}, C_{cj})^2 \tag{2.28}$$

$$BCSS(C) = \frac{1}{K} \sum_{i=1}^{k} BCSS(c_i) \tag{2.29}$$

A small value of BCSS indicates clusters that are close to each other, while a high value indicates clusters that are spread out, which representa good clustering solution.

2.2.3.3 Relative Validation Index

The relative measures are based on comparing the clustering solutions produced from the same algorithm, but under different parameter values (e.g. changing the number of clusters or parameters of the algorithm).

The fundamental idea is to select the best clustering solution of a set of pre-obtained solutions by different parameters. Let consider P_{alg} the set of parameters associated with a specific clustering algorithm, and K is the number of clusters. There are two cases whether the number of clusters K is included in P_{alg} or not [73]:

- P_{alg} **without the number of clusters** K**:**
  In this case, the selected clustering algorithm is run for a wide range of its parameters values P_{alg} to identify the largest range of these parameters for which K remains constant. Then the optimal parameter values are selected based on the middle of the range of these parameters.

- P_{alg} **with the number of clusters** K**:**
  In this case, the best clustering solution is selected by running the cluster-

ing algorithm with different number of clusters K, then these solutions are evaluated based on a validity index q.

For each clustering solution, the algorithm runs r times using different set of values of parameters in P_{alg}. Then the best values of the selected validity index q obtained for each K are plotted to identify the best clustering solution.

2.3 Introduction to Time Series

Time Series(TS) is a sequence of observations that a variable takes over time, such as (t_1, v_1),...,(t_i, v_i),...(t_m, v_m), where t_i is the time step and v_i is the observation. The order in the time series data is important since the values are based on time.

A time series can be univariate when the time series represents a sequence of the same variable collected overtime, or multivariate when several variables are observed and recorded simultaneously, this becomes a multivariate time series (MVTS) [58]. A large variety of real-world applications use Time Series (TS) analysis such as weather forecasting [32], earthquake prediction [47], system engineering [165], or human activity recognition [141]. Multivariate Time Series (MVTS) are becoming more prominent specially as part of large and complex datasets being produced [55].

2.3.1 Time-Series Distance Measures

Distance is the degree of similarity/dis-similarity between points, groups, or TS, however the similarity of the time series is not calculated, it is estimated and it is not based on the exact match as in traditional clustering methods [117]. There are many distance metrics for time series data, we only considered the most common two types of distance metric in this research:

- **Dynamic Time Warping (DTW)**:

  DTW is a non-linear similarity measure that is used to find the optimal align-

ment (shortest path) that minimizes the sum of distances between two TS. It was proposed by Sakoe and Chiba [136]. DTW is a popular distance measure in the speech recognition community. Also, it has been widely applied on TS clustering, classification, and anomaly detection. It is an extension of the Euclidean Distance measure (ED) that offers a non-linear alignment between two series. Figure 2.2 shows the optimal alignment between two different TS. DTW is an elastic distance measure that compares one-to-many or one-to-none points in two TS based on their minimal distance [136]. As we can see in Figure2.3 a single point "observation" in the first TS matches to one or more points in the second TS. The DTW distance metric takes into account the movement and the distortions in the TS much better than ED.

Consider two TS to compare X=x_1,x_2,x_3,...,x_n and Y=y_1,y_2,y_3,...,y_m. To compare X and Y, a point-wise distance matrix $M(n * m)$ is created, where every element in this matrix corresponds to the distance between two points $i \in X$, and $j \in Y$, as follows:

$$M_{i,j} = (x_i - y_j)^2 \tag{2.30}$$

To find the optimal alignment between X and Y, a warping path

$$W = w_1, w_2, w_3, \ldots, w_k,$$

in matrix M_i is constructed, where $w_k = (i, j)_k$ indicates the alignment and matching relationship between i and j.

The DTW distance between X and Y is calculated as follow:

$$DTW(X, Y) = min\left\{\frac{1}{K}\sqrt{\sum_{k=1}^{K} W_k}\right\}. \tag{2.31}$$

- **Shape-Based Distance (SBD):**

  The Shape-Based Distance measure is a faster alternative to DTW, and it is based on the cross-correlation with coefficient normalization (NCC_c) sequence between two series. The cross-correlation is a statistical measure that

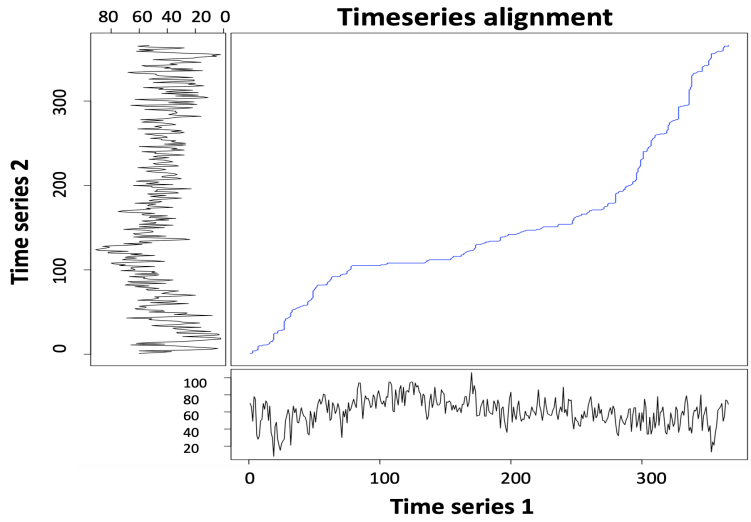


Figure 2.2: Visual comparison of optimal alignment between two time series based on DTW.

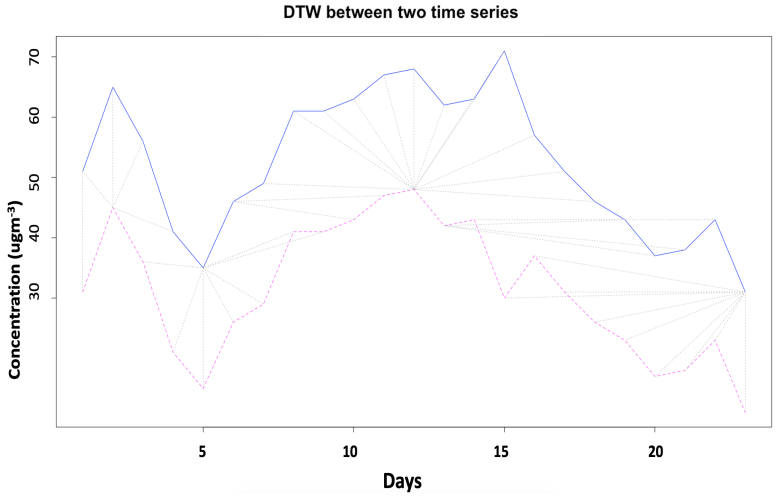


Figure 2.3: Visual comparison of matched points between two time series based on DTW.

can determine the similarity of two sequences (time series) even if they are not properly aligned, while NCC_c uses the Fast Fourier transform (FFT) to compute the cross-correlation sequence between two series. The SBD was proposed as part of the k-Shape clustering algorithm by Paparrizos et al. [119].

As recommended by Paparrizos et al. [119], TS data should have appropriate amplitudes, or be z-normalized in order to get better results using SBD

metric. The SBD distance is calculated by the following formula:

$$SBD(X,Y) = 1 - \frac{max(NCC_c(x,y))}{\|x\|_2\|y\|_2} \tag{2.32}$$

where $\|.\|_2$ is the l_2 norm of the series. SBD range lies between 0 and 2, with 0 indicating perfect similarity [138]. The l_2 norm also known as the Euclidean norm represents the shortest distance from one point to another. It is calculated as the square root of the sum of the squared vector values [164].

Some of the univariate TS similarity measures cannot handle missing data or TS of different lengths [162]. The difficulties are increased when more than one time series is involved (i.e. in a multivariate TS environment); then the clustering problem becomes more challenging. The major limitation of DTW is of handling TS of different lengths. So, the existing DTW is also not sufficient for clustering multivariate time series data.

2.3.2 Time Series Analysis

There are different tasks for time series data mining, such as querying time series, classification and clustering. Querying of the time series is the most frequent task. It is based on retrieving a set of solutions that are most similar to a query provided by the user. The query can be identified as 'whole series matching' retrieving only the time series that matches the whole query, or 'sub-sequence matching' retrieving every sub-sequence of the series matching the query [58].

Time series clustering is the process of finding the most homogeneous time series that are as distinct as possible from others, while time series classification uses the labels to each time series and creates a model that differentiates them. Both, time series clustering and classification, can be processed using the whole or a sub-sequence series. To measure the similarity/dissimilarity between two time series,

the most common used metric is DTW [136], as well as, Euclidean distance (ED), which is one of the first generic dissimilarity measures that was proposed for time series, but it has some limitations that can be solved using DTW measure as that is an elastic dissimilarity measures [140]. DTW provides better results for many application areas of time series, however many alternatives to DTW have been proposed.

Existing time series clustering algorithms can divide into three types depending on whether raw data are used directly or indirectly [9, 98]: Raw-data-based [55, 64], feature-based [162, 61], and model-based [175, 56].

- **Raw-data or observation based clustering**. In this approach, clustering is based on a comparison of the observed time series. This approach is useful when the time series are not very long [55].
- **Feature-based clustering**. This approach is used when time series data is very long, with noisy time series, and when the relevant knowledge about a problem domain is available [61]. The clustering process is based on features extracted from the time series [106].
- **Model-based clustering**. This approach is based on the assumption that a time series can be generated by a statistical model. It is assumed that a set of time series generated from the same model would most likely have similar patterns [56, 175].

2.4 Data Imputation Methods

The main classification of missing data mechanism are data missing at random (MAR), missing completely at random (MCAR) or missing not at random (MNAR) based on Rubin and Schafer [131, 139]. MAR means that the probability for a data point, in our case a time series, to be missing is not related to the missing data, but it is related to other observed data. For example, if we have two variables, age

and income, and the probability that income is missing is not related to income itself, but may be related to age, then the data are MAR. MCAR means there is no relation between the missing data point and any other values in the dataset, observed or missing. As an example, if we have biosamples collected for genotyping and some results are missing because the instrument failed for one batch of samples then the data is MCAR. The last mechanism known as MNAR, also called "non-ignorable", means there is a relation between the missing data point and the observed values [128]. An example may be rainfall amount that cannot be measured due to extreme raining that affects the inference related to predicting extreme precipitation.

Based on a number of studies [115, 121, 57] the missing data mechanism of air quality data is MAR. To impute the missing data, there are two main methods available: single imputation and multiple imputation [80]. These methods take incomplete dataset and create a complete dataset.

2.4.1 Single Imputation Techniques

The single imputation method is a method where each missing value is imputed by only one estimated value. This method is more common than multiple imputation methods. The main advantage of single imputation is that simple statistical methods such as mean [101], regression interpolation [45], ... etc can be applied to replace the missing data. However, the main drawback of this method is that replacing the missing value by a single value and then consider it as if it were a true value [129] ignores the uncertainty in the imputation. As a result, single imputation does not reflect the uncertainty and necessary variation due to missing data [130]. The fastest way to impute the missing value is by replacing the missing value by the mean value [15].

2.4.2 Multiple Imputation Techniques

Multiple imputation is a statistical technique, that was proposed by Rubin (1987) [129]. This technique replaces each missing value with a set of imputed (n) values. The results of the multiple imputation methods are (n) datasets [132]. The differences between these datasets reflect the uncertainty of the missing values [156]. Multiple imputation has some advantages over other missing data approaches; it involves filling each missing value multiple times, creating multiple datasets, and it takes into account the uncertainty [26].

One of the most effective multiple imputation methods is Multiple Imputation via Chained Equations (MICE), is also known as Sequential Regression multiple imputation [155, 123]. This method assumes that the missing data are Missing at Random (MAR), which means that the probability for a data point, in our case a time series, to be missing is not related to the missing data of that point, but it is related to some of the other observed data.

This method imputes missing data based on Fully Conditional Specification (FCS), meaning that each incomplete variable is imputed by a separate model a variable-by-variable basis. This means that each variable can be modeled according to its distribution, for example, continuous variables modeled using linear regression, binary variables modeled using logistic regression and polytomous regression for categorical data [156, 26]. For numeric continues data (time series), this method uses predictive mean matching (PMM) in the imputation process [30].

The MICE imputation process starts by initially filling the missing values using simple imputation, such as imputing the mean, and is performed for every missing value in the dataset. Then, for each variable, a univariate regression model is built by considering each variable as a dependent variable, and all the other variables are independent variables in the regression model. The missing values in each variable are replaced with predictions from the regression model. This process continues for each variable with missing data until all specified variables have been imputed.

The imputation process for each individual variable should be run several times (iterations) until it appears that convergence has been met. At the end of each iteration, the observed data and the final set of imputed values would then create one complete data set. The whole process is repeated to create the second, third, ..., etc. dataset [26].

2.4.3 Imputation Methods Evaluation

There are a wide range of statistical model evaluation metrics that can be used to calculate the magnitude of the errors between the actual and the imputed data. We review some of these metrics that is used in air quality model evaluation as well.

We define each one of these valuation statistics as in [33], in the following definitions, O_i represents the ith observed value and M_i represents the ith modelled value for a total of n observations.

- **Fraction of predictions within the factor of two (FAC2):** This measure is defined as the percentage of the predictions/modelled within a factor of two of the observed values that satisfy:

$$0.5 \leq \frac{M_i}{O_i} \leq 2.0 \tag{2.33}$$

- **Mean Bias (MB):** The mean bias is the mean error that indicates if the mean over or under estimate of predictions. The MB is defined as:

$$MB = \frac{1}{N}\sum_{i=1}^{N}(M_i - O_i) \tag{2.34}$$

- **Mean Gross Error (MGE):**

$$MGE = \frac{1}{N}\sum_{i=1}^{N}|M_i - O_i| \tag{2.35}$$

- **Normalised Mean Bias (NMB):** This measure is the same as the MB, but normalized to use for comparing objects with different scales, in our case pollutants that cover different concentration scales.The NMB is defined as:

$$NMB = \frac{\sum_{i=1}^{N}(M_i - O_i)}{\sum_{i=1}^{N} O_i} \tag{2.36}$$

- **Normalised mean gross error (NMGE):**

$$NMGE = \frac{\sum_{i=1}^{N}|M_i - O_i|}{\sum_{i=1}^{N} O_i} \tag{2.37}$$

- **Root Mean Squared Error (RMSE):** This measures the average magnitude of the errors between the observed and the modelled values. The RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(M_i - O_i)^2} \tag{2.38}$$

- **Coefficient of correlation (R):** Pearson's correlation coefficient (R) is a measure of the strength of linear relationship between two variables/ TS. The values of the Pearson's correlation coefficient take a range from +1 to -1. A value of 0 indicates that there is no correlation, a value less than 0 indicates that there is a negative correlation, and values greater than 0 indicate a positive correlation between variables. Pearson's correlation coefficient (R) is defined as:

$$r = \frac{1}{(N-1)}\sum_{i=1}^{n}\left(\frac{M_i - \overline{M}}{\sigma_M}\right)\left(\frac{O_i - \overline{O}}{\sigma_O}\right) \tag{2.39}$$

- **Index of Agreement (IOA):** The Index of Agreement is commonly used in model evaluation to measure the degree of model prediction error. It is proposed by Willmott [168]. Its value spans between -1 and +1. Values near to +1 represent better model performance, while values near to -1 can

mean that the model-estimated variations are poor estimates of the observed variations; but, they can also mean that there is little observed variability. IOA = 0 implies that the sum of the error-magnitudes is equivalent to the sum of the observed-deviation magnitudes [169]. The IOA is defined as:

$$IOA = \begin{cases} 1 - \dfrac{\sum_{i=1}^{n} |M_i - O_i|}{c \sum_{i=1}^{N} |O_i - \overline{O}|}, when \\ \sum_{i=1}^{N} |M_i - O_i| \leq c \sum_{i=1}^{N} |O_i - \overline{O}| \\ \dfrac{c \sum_{i=1}^{N} |O_i - \overline{O}|}{\sum_{i=1}^{N} |M_i - O_i|} - 1.0, when \\ \sum_{i=1}^{N} |M_i - O_i| > c \sum_{i=1}^{N} |O_i - \overline{O}| \end{cases} \tag{2.40}$$

Where $c = 2$, for better scaling, as identify by the author [169].

If one model works best for one measure and another for another measure, we select the model that is best for the majority of these measures.

2.5 Data Mining Application to Air Pollution

Data mining encompasses a set of tools to analyse and discover knowledge in many fields, including that of air pollution; it is beneficial when dealing with big, complex, multi-dimensional data. In recent years, data mining techniques have been increasingly applied in air pollution studies to achieve several goals, as their ability to take data from a variety of sources enables us to model air quality systems that are dynamic, spatially expansive, and heterogeneous, and they can process data from a variety of sources [114].

In our investigation of the related work that focuses on applying data mining techniques and time series analysis to air pollution problems, we found several

studies that used different models to predict or forecast air quality [145, 72, 76], to identify air pollution sources [145, 38], to explore the relation between air pollution and climate change [157, 36, 23], and to assess the association between air pollution and some other health concerns [63, 153, 52, 65, 148].

We divided the previous proposed air quality data mining and machine learning models into two categories: Prediction paradigm covered in Section 2.5.1 and knowledge discovery paradigm covered in Section 2.5.2.

2.5.1 Prediction Paradigm

There is an urgent need for accurate air quality prediction, as it can help to reduce peak pollution levels. Predicting air quality is very important for both the public and the authorities. If the authorities can identify areas with high pollution levels or predict episodes of high levels of ozone or other air pollutants, they can act to safeguard public health.

Traditional approaches for prediction and analysis of air quality use mathematical and statistical models in which data are coded with mathematical equations and processed using a physical model. However, these methods can be inefficient and suffer from disadvantages.They also provide limited accuracy [114]. Advancements in technology and research have led to proposals of alternative methods for forecasting atmospheric pollutant levels and air quality; these use single or hybrid models, such as Artificial Neural Networks (ANN) [173], Support Vector Machines (SVMs) [146], Fuzzy Logic (FL) [32], and Genetic Algorithms [146].

The main objective of the predictive models is to predict a value for a decision or output variable based on the values of other variables. The predicted value can be categorical or numerical, depending on the user's objectives. Several supervised learning algorithms are used in forecasting and predicting air pollution [28].

We categorize the studies under the prediction paradigm based on their objective into three categories.

2.5.1.1 Machine Learning for Air Pollution Epidemiology

Research under this category focuses on applying data mining techniques to generate new hypotheses to better understand the relationship between air pollution and adverse health conditions [28].

Several work have been done to study the associations between air pollutants exposures and asthma [63, 153], which is one of the respiratory disease that effected mostly by air pollution. While [52], studied the association between exposure time and asthma. Others [157, 104, 152] have used data mining techniques to identify the relation between multiple air pollutants exposure and mortality rates in the UK.

Gass et al. [63] have used Regression Trees (CART) to analyze how the associations between mixtures of exposures (to CO, NO_2, O_3 and $PM_{2.5}$) are connected with pediatric admissions for asthma in Atlanta, USA. Similarly, Toti et al.[153] used association rule mining to identify the same relation in Texas, USA. Researchers found that using CART and association rule mining helps with understand of single and multiple simultaneous exposures in the field of air pollution epidemiology studies [63, 153].

Ding et al. [52] focused on studying the link between short-term exposure to air pollutants and children's hospital visits for asthma in Chongqing, China. Using conditional logistic regression to analyze the data. They found that short-term exposure to multiple air pollutants (i.e. PM_{10}, $PM_{2.5}$, SO, NO, and CO) could result in a rise in the number of hospital visits. The analysis also showed that the strongest effect on asthmatic children related to NO_2, while O_3 had no effect. Conditional logistic regression (CLR) helped to determine the association between some pollutants and asthma in children. The major limitation of Logistic Regression is the assumption of linearity between variables. Also, CLR is usually employed when case subjects correlated to particular conditions.

Vitolo et al. [157] identified the multi statistical dependencies between air pollu-

tion, climate, and health using a Bayesian Network (BN) graphical probabilistic model. This work is the only previous work that covered a wide area of the UK, from 1981 to 2014. They used multivariate data that includes environmental factors like topography and weather variables (Temperature, Wind Speed, Wind Direction, Solar Radiation) provided by ERA-interim, health outcomes in the form of mortality rates for England, and the exposure levels, which consisted of the concentration of air pollutants (O_3, NO_2, SO_2, PM_{10}, $PM_{2.5}$, and CO) provided by Defra. As a result, the model identified and predicted the dependencies between variables that predict exposure to pollutants and population health outcomes. They used air pollution data predictions with missing data. In this paper, they introduced the problem of incomplete air pollution data set (not all the pollutants are measured in every stations and the station does not measure the pollutants all the time). To solve that, an expectation-maximization algorithm was used in order to make use of partial observations as part of the Bayesian networks implementation.

Others [104, 152], have analysed mortality and emergency hospitalisations associated with atmospheric particulate matter episodes in London, UK, particularly in the spring of 2014. As recorded that the UK experienced widespread of high levels of particulate air pollution in March-April 2014, where the hourly mean observations of $PM_{2.5}$ reached up to 83 μgm^{-3} at urban background sites. These studies proved that long-term exposure to air pollution has a more significant potential effect on public health than short-term air pollution episodes do. In this context, Thornes [152] examined the relationship between morbidity and weather conditions on some days of high pollution associated with Saharan dust. He used three datasets: UK Met Office; Daily Air Quality Index (DAQI); and morbidity data from the London Ambulance Service. The analysis includes two air pollution episodes across the UK that led to a 'High' or 'Very High' DAQI during March and early April 2014. The high DAQI was associated with increasing levels of $PM_{2.5}$ and PM_{10}. After analysing the weather conditions for these periods, the study found that the sources of air pollution were likely to have been imported from Northern

Germany, Holland, and North Africa. The second episode, in early April, had a more significant health impact on hospitalisation on (Saharan) dust storm days, particularly for cardiovascular causes, than the March episode did. The study also found that the effect of Saharan dust depends on episode length. McIntyre et al. [104] analysed the same episodes, and they agreed that long-term exposure to poor air quality had been associated with premature deaths. They found that over the two episodes, the estimated total mortality burden attributable to short-term exposure to $PM_{2.5}$ was around 600 deaths brought forward summed across the UK over ten days.

2.5.1.2 Machine Learning for Air Pollutants Prediction

Other researchers have worked on predicting the concentrations of some air pollutants which are of significant concern to people's health when their levels in the air are relatively high, for example, fine particulate matter $PM_{2.5}$ and ozone.

Most of the air pollution studies under this section focus on predicting the air quality index [145, 72, 76], predicting a specific pollutant such as ozone level [40], $PM_{2.5}$ level [42, 163, 134, 88], or more than one pollutant ([159, 32, 176, 76]).

Tapiwas and Ditsela [40] estimated actual daily ozone levels in μgm^{-3} in Johannesburg, South Africa using neural networks (ANN) and a linear regression approach. Two approaches for the estimation are used, the cross-correlation between readings from the station, and the spatial correlation between neighbouring stations. The highest accuracy achieved to estimate ozone level at a particular site was 79% using ANN with spatial correlated features. Wei [163] proposed a binary classification to predict the level of the $PM_{2.5}$ into 'High' ($>115\ \mu gm^{-3}$) and 'Low' ($<115\ \mu gm^{-3}$) in order to solve the uncertainty of the specific $PM_{2.5}$ value. The approach used Support Vector Machines (SVMs) to predict the $PM_{2.5}$ level based on a dataset consisting of daily weather, the air quality of the previous day, and traffic parameters in Beijing, China. In a similar study, Rybarczyk and Zalakeviciute [134] used

J48, ZeroR, and Naive Bayes to predict fine particulate matter $PM_{2.5}$ given a set of weather conditions such as wind speed and rainfall in Quito, Ecuador. A decision tree algorithm was used to classify the concentrations of $PM_{2.5}$, into two categories ($>15\mu gm^{-3}$ vs. ($<15\mu gm^{-3}$), from a limited number of parameters such as the level of precipitation and the wind speed and direction. The proposed model was able to classify 65% of the concentrations correctly. Another study at Quito [88] aims to identify the meteorology effects on $PM_{2.5}$. It used regression modelling to predict $PM_{2.5}$ values based on meteorological data for six years, that include wind speed, wind direction, and precipitation. The results showed that the regression model can predict $PM_{2.5}$ concentrations up to $20\mu gm^{-3}$, however the model accuracy was improved in conditions of strong winds and high precipitation. This can indicate the correlation between $PM_{2.5}$ concentrations and extreme weather conditions.

Zhu et al. [176], proposed refined models to predict the hourly air pollution concentration for SO_2, O_3, and $PM_{2.5}$ based on the historical meteorological and air pollutant data of previous days by formulating the prediction of 24 hours as a multi-task learning problem and comparing it to several traditional regularisation such as standard Frobenius norm and nuclear norm regularisation. Vong et al. [159] used SVM to forecast air quality for NO_2, SO_2, O_3, SPM from pollutants observations and meteorological data in Macau, China for the period of 2003-2006.

Corani and Scanagatta [42] applied multi-label classification based on Bayesian Networks to the problem of predicting multiple air pollution variables. This study included three case studies; predicting the $PM_{2.5}$ in Shangai and the ozone in Berlin and Burgas. In each study, they predicted the level of the air pollutants for the current day and the next day given observations of the air pollutants and the meteorological parameters from the previous day. Each time they determined the essential variable for the prediction. For example, they found that the $PM_{2.5}$ level from the previous day is the most important variable to predict the $PM_{2.5}$ for the next day.

Han et al. [76] used air quality data from eight major cities in different states in the USA and five machine learning algorithms to forecast the air quality index for NO_2 and O_3 (using 1- to 5-steps ahead forecasting). The dataset contains historical data for 16 years that included NO_2, O_3, Air Quality Index (AQI), and information on the location of sample collection and dates. In this study they used a Linear Regression Algorithm (LR), Multi-Layer Perceptron (MLP), Radial Basis Function Network (RBFN), Radial Basis Function Regressor (RBFR), and Support Vector Machine Regressor (SMOR) for forecasting the NO_2 and O_3 Air Quality Indices. The results suggested that the SMOR outperforms all other algorithms followed by RBFR and the MLP.

All the previous work discussed in this section focuses on a limited number of pollutants, mostly $PM_{2.5}$ or O_3, and may or may not be applicable to all other air pollutants. Studies that used machine learning prediction techniques to predict air pollutant concentrations are based on meteorological conditions without taking any consideration of the relation between different pollutants and how they react to each other [176, 134, 159]. Also, often they just focus on the spatial correlation between neighbouring stations [40] to predict pollutant concentrations in some locations. Another drawback is that prediction is limited to short periods of time, such as predicting air pollution concentrations for the next 24 hours [176].

2.5.1.3 Machine Learning Prediction with Ensemble

Other recent studies, [145, 72] applied an ensemble learning method to model the complex relationships between a set of independent and dependent variables to increase the accuracy of a single model. On the other hand, some studies [174, 102, 147, 161] found that spatial distance plays an important role in air pollution, and with temporal relations can help to model changes more accurately.

Singh et al. [145] used ensemble learning (EL) model-based classification and regression functions for predicting air quality and identifying air pollution sources.

The classification is achieved using an Ensemble Decision Tree (DT) that consists of a Single Decision Tree (SDT), a Decision Tree Forest (DTF), and a Decision Tree Boost (TB) to classify the ambient urban air quality based on the seasons, and a regression function to predict the AQI, and Combined AQI (CAQI) in the city of Lucknow, India. Also, they used principal component analysis (PCA) to study the correlation between air pollutants and meteorological factors and to identify the air pollution source. This study found that vehicular emissions and fuel combustion are the primary air pollution sources in the city, and the air quality indices indicate that during summer and winter the urban air quality is unhealthy to humans. Chen et al. [39] used ensemble Neural Network to forecasts AQI one day ahead in 16 urban cities in China, using meteorological data (temperature, wind speed and direction, precipitation, humidity, and Sunshine duration) and air pollutant data that included NO_2, SO_2, O_3, CO, PM_{10}, $PM_{2.5}$ for three years. In this model, they used a feature selection method called Partial Mutual Information (PMI), that measures the degree of predictability of the output variable based on the input variables to select the best predictors for each city.

Hajek et al. [72] predicted the common Air Quality Index (AQI) classes of three monitoring stations in the Czech Republic based on data measured by the three monitoring stations. Their dataset includes the average daily meteorological variables and the maximum daily emission variables (NO_2, NO, NO_x, SO_2, PM_{10}, O_3), 'working day' which is a variable with a value of 1 or 0, and station location. They used a model to predict the air quality indices for each air pollutant separately, then the common air quality index was predicted using several computational intelligence methods including adaptive neuro-fuzzy inference system (ANFIS) and support vector regression (SVR). Liu et al. [102] used a Multi-dimensional collaborative support vector regression (SVR) model to forecast the air quality (AQI) in one city using the data from its neighbours or the cities around. The model performance is especially improved when the surrounding cities' air quality information is included. The dataset includes the air pollutants ($PM_{2.5}$, PM_{10}, SO_2, CO,

NO_2,and O_3), weather parameters (temperature, wind speed, and wind direction), and the AQI from the previous day. This study considered three cities in China Beijing, Tianjin, and Shijiazhuang for the period (2014-2016). Based on a recent review [28], Classification and Regression Trees can help overcome challenges in analysing multiple chemical interactions. Also, machine learning algorithms in general are effective for integrating knowledge and practices and producing reliable forecasts for air pollution.

Other studies have focused on urban environments and take in complex variables, including spatial-temporal features. Zheng et al. [173] proposed a semi-supervised learning approach based on a co-training framework to predict the air quality in a city at different locations in Beijing, China. They use two types of classifiers. The first is a temporal classifier using linear-chain conditional random field (CRF) to predict air quality in a different location based on temporal features such as meteorological data, human mobility, and traffic related data. The second is an spatial classifier using an artificial neural network (ANN) to predict the dependency of air quality in the location based on geo-spatial features such as Point Of Interest (POI) features, road network, location, concentrations of six air pollutants: NO_2, SO_2, O_3, CO, $PM_{2.5}$ and PM_{10} for three years, and the AQI label for neighbours locations. They found that the AQI depends on its previous status, and reflects its status on its neighbours. On a later study [174] to solve some limitations of their previous model, the authors used a data-driven method to forecast the air quality over the next 48 hours in 43 cities in China. Their model is based on a hybrid predictive model that combines the temporal and spatial prediction using current meteorological data, weather forecasts, and air quality data of the station and that of other stations within a few hundred kilometres. They used a linear regression-based temporal predictor that predicts the air quality based on local/temporal data such as AQI for the past few hours and the weather forecast, a neural network-based spatial/global predictor, that predicts the air quality based on neighbours data, and a dynamic aggregator to combine the predictions of the

spatial and temporal predictors according to meteorological data using Regression Tree (RT).

Soh et al. [147] predicted the air quality index for the $PM_{2.5}$ in seven stations in Taiwan using the information of adjacent stations or stations with similar temporal patterns. Their model is based on the spatial-temporal analysis using the K nearest neighbours in geographical distance using the Euclidean distance measures and similarity in time using DTW measure. The used the hourly $PM_{2.5}$ observations from 7 stations for two years. They found that using the kNN-DTW is better than using the kNN-ED. Wang and Song [161] proposed a deep spatial-temporal ensemble (STE) model that contains three components that combine the spatial, temporal, and weather data to predict the air quality. Their model measures the spatial correlation using Granger causalities among stations to generate spatial data that represent the relative stations and relevant areas. Also, they used temporal predictor for a long-term and short-term air quality predicting based on deep Long Short-Term Memory (LSTM). The dataset includes 35 monitoring stations in Beijing, China, for 5 years, that combines the observations of six air pollutants: CO, NO_2, SO_2, O_3, PM_{10}, $PM_{2.5}$, and weather forecasts parameters.

2.5.2 Knowledge Discovery Paradigm.

In this type of data mining paradigm, unsupervised learning algorithms are used to aggregate a particular set of objects based on their characteristics, so they are grouped according to their similarities using clustering algorithms such as k-means, hierarchical clustering, and fuzzy clustering. In the field of air pollution, there are some efforts under the knowledge discovery paradigm and since the air pollution is a problem that cannot be treated independently from the climate; most of the studies focus on studying the relation between air pollution and meteorological conditions [36, 23].

Chen et al. [36] studied the relation between air pollutants and meteorological data

around the high-tech industrial locations in Taiwan using hierarchical clustering to group the air monitoring data and the corresponding meteorological data to find the correlation between air pollutants and the changes in meteorological data.

Saithan et al. [135] used Agglomerative Hierarchical clustering (AHC) to cluster the relevant meteorological variables and air quality pollutants into groups. They used 10 air pollutants (CO, NO, NO_2, NO_x, SO_2, HC, CH_4, NMHC, PM_{10}, and O_3), and seven meteorological variables (wind speed, wind direction, humidity, temp, pressure, solar radiation, and rain) for the industrial areas of the East of Thailand. Then, they used these clusters of objects to categorise the days having high ground ozone concentration into three groups based on time of day (dusk time, noon to afternoon and evening interval). Clustering analysis was applied to determine the spatial patterns for the days where the ozone exceeded the normal limit (33 days only in this study).

Austin et al. [23] used k-means and hierarchical clustering to cluster days for six years (2004-2009) in Boston, USA, based on multi-pollutants profiles. The dataset includes the daily observations of O_3, Fe, NO_2, and the components of $PM_{2.5}$ such as elemental carbon (EC) and organic carbon (OC). Also, it includes some weather parameters such as temperature, wind speed, and water vapor pressure. The clustering algorithm identified distinct pollution events at a given site and clustered the days into five distinct groups, that are differed in their chemical, physical and meteorological characteristics. As a result, they found that the days with similar meteorological conditions also have similar pollutant concentration relationships. Then, in a later study, Austin et al. [24] they used the k-means algorithm to identify spatial patterns in air pollution data to cluster the USA cities based on their similarity on $PM_{2.5}$ composition profiles, then characterise these clusters based on chemical characteristics, emission profiles, geographic locations and population density. Their dataset consists of the observations of the components of $PM_{2.5}$. The data are collected from 109 monitoring sites, including different environmental types (Urban, Suburban, and Rural) during the period (2003–2008)

which included less amount of missing data. As a result, they clustered these stations into 31 distinct clusters.

Barrero et al. [27] studied the variations of PM_{10} concentrations at 43 stations in Basque country using k-means clustering to classify stations based on characteristic temporal variations of PM_{10}. Tuysuzoglu et al. [154] applied different clustering algorithm such as k-means, Expectation Maximization, and Canopy for each air pollutant in the dataset (NO, NO_2,SO_2, PM_{10}, and O_3), then aggregate the clustering results based on majority voting to identify one clustering solution to similar regions in terms of air quality. This study used the observations from 49 stations that measure all the pollutants included in the study. The data contains the daily means of each pollutants, and the missing data imputed using the mean value.

Ignaccolo, Ghigo and Giovenali [81] classified the air quality monitoring network in Northern Italy using Partitioning Around Medoids algorithm (PAM) considering three air pollutants, namely NO_2, PM_{10}, and O_3. They transform the time series of pollutants daily observations into functional form to smooth the time series, then run the PAM algorithm to cluster each pollutant individual to characteristics at each cluster/ region.

In China, Ye et al. [171] used spatial clustering characteristics based on the comprehensive air quality index (CAQI), that covers and integrates six pollutants ($PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, O_3) to investigate the spatial and temporal distributions of air quality at the city level for year 2016 in 338 Chinese cities.

Chen et al. [38] proposed Environmental Pollution Clustering (EPC) to cluster large number of sample data points based on their similar characteristics, including concentrations of pollutants and sources of pollution. This algorithm can be used for rich, high-dimensional datasets that may include outliers. According to the authors, the proposed algorithm solves some of the limitation of conventional methods of source apportionment such as the principle component analysis and positive matrix factorisation.

2.6 Time Series Analysis Application in Air Pollution

Time series analysis are mostly based on statistical techniques. They are useful tools to extract meaningful characteristics of the data over time. Also, forecasting can be used to predict future values based on historical data, with good accuracy [112]. One of the most powerful methods in time series forecasting is the ARIMA model proposed in 1970 by Box and Jenkins [29]. ARIMA stands for the Auto-regressive Integrated Moving Average model [144]. ARIMA models are a class of complex linear models that are capable of representing stationary and non-stationary time series [29].

2.6.1 Statistical Models for Time Series Analysis

Statistical models are widely used to estimate or forecast concentrations of air pollutants. These models do not consider the physical and chemical processes and use historical data in predicting. The simplest statistical approaches include Time Series [94, 143] and ARIMA models [112, 103, 113].

Time series analysis has recently been used in many applications in air quality forecasting. ARIMA models have been applied in air quality forecasting to predict the AQI [112, 103, 94, 143] or the future values of air pollutants concentrations[113]. These works, used ARIMA model independently or with other data mining models.

Anu [112] used ARIMA to analyse and forecast the air quality index in a small area in Kerala, India. This analysis is based on a dataset that contains NO_2, SO_2, Suspended particulate matter (SPM), and Respirable particulate matter (RSPM) recorded from four stations for the period of 2012-2015. Also, Liu et al. [103] used a set of ARIMA models with numerical forecasts (ARIMAX) to forecast the AQI up to three days ahead at both day and hour level in 15 monitoring stations in Hong Kong, using O_3, NO_2, and $PM_{2.5}$. Even though the results show significant improvements in the forecast of air pollutants, the models are built for limited steps

(such as 3 days or 7 hours) ahead. Another limitation in their models is that they work for limited station types.

Time series analysis was also used to predict the AQI for each air pollutant such as NO_2, PM_{10}, O_3, CO, SO_2 and $PM_{2.5}$ of six cities in China one hour ahead using Fuzzy logic to extract the conceptual features [94]. While Sharma et al. [143] used a time series regression model to analyse the pollution trends of 2016 to 2017 and predict the future values of the air quality and pollution trends til 2022 based on the historical data. They used a dataset that include the following air pollutants: NO, NO_2, Toluene, NO_x, O_3, PM_{10}, $PM_{2.5}$, SO_2 and CO for one station in Delhi.

Cabajal et al. [32] used the auto-regressive model (AR) to predict air quality concentrations for PM_{10} and O_3 at Mexico city based on a fuzzy inference system for classification, which classifies the pollution level into excellent, good, regular, bad and dangerous. This model used the historical observations of the air pollutants to predict their future values. Ni et al. [113] proposed a hybrid model that implements a NN and ARIMA to forecast the concentration of $PM_{2.5}$ for a few hours ahead in Beijing, China. They analyse the correlation of $PM_{2.5}$ and different source types and potential related factors including humidity, wind speed, CO, NO_2, PM_{10}, SO_2, and social media data (microblog data, specifically the daily number of specific microblog entries).

2.6.2 Data Mining for Time Series Analysis

Due to the increasing availability of time series data and the demand to analyse them, clustering time series has attracted growing research interest in recent years [55, 97, 70, 98, 162, 175, 95]. However, most of the existing clustering methods are for univariate TS data, while clustering multivariate time series remains a challenging task [98].

The main problem with multivariate time series is dimensionality, and the majority of the existing researchers have proposed methods for dimensionality reduction.

Commonly, PCA *similarity factor* has been proposed to measure the similarity between multivariate time series (MVTS) [59, 95]. Feature extraction that transform the TS into set of features is used to measure the similarity as well [162, 97], and other used statistical models to measure the similarity [175], an ensemble model [108] to aggregate the similarity of univariate TSs. In this section, we will discuss the recent researches into MVTS clustering either using simple or ensemble clustering techniques .

2.6.2.1 Simple Clustering for Time Series

There has been some research into similarity within MVTS. One effective way is to extract the features for time series, then apply a simple data mining approach such as clustering. For example, Fontes et al. [59] proposed a MVTS clustering method based on extracted features from the univariate TS. PCA is used to measure the similarity between MVTS, then fuzzy k-means is used to cluster these TS. This clustering approach was used for fault detection in a gas turbine. Li [95] proposed a multivariate time series clustering based on common principal component analysis (CPCA) to construct a projection coordinate space and to lower the dimension of the data for the clustering process. The proposed clustering approach has two main stages, first is assigning every MVTS to a cluster based on the similarity to the projection coordinate space (i.e. cluster prototype) and the other is to construct a new prototype of a cluster based on CPCA. Recently, Li et al. [96] transformed the MVTS into a network (i.e. called component relationship network (CRN)) to reflect the relationship of the MVTS data, then an improved version of DTW is used to measure the similarity for each component and cluster the MVTS data.

While others improved the existing clustering algorithm to work with MVTS data. Wu et al. [170] transformed MVTS data into Independent Components (IC) using Independent Component Analysis (ICA) to find the independent patterns for each TS. Then they proposed a new clustering algorithm called ICACLUS to cluster these patterns according to the extracted ICs instead of the traditional k-means.

In this algorithm, the similarity between TS is measured based on the number of matching ICs. This algorithm starts by computing k ICs for each TS. Then, the algorithm sorts the ICs based on the load and selects the most dominant ICs, which are the ICs with the most negative loadings and the ICs with the highest positive loadings. After that, the similarity between TS is measured based on the number of matching ICs. If this number gets to the identified threshold, then these TS are grouped into a cluster. Zhou et al. [175] developed a model-based multivariate time series clustering algorithm that first discovers the temporal patterns in each TS using confidence value to represent the relationship between different variables. Their algorithm is based on the k-means and aims to group MVTS based on the degree of patterns discovered into the same cluster.

2.6.2.2 Ensemble Clustering for Time Series

Ensemble have been applied to time series clustering, Mikalsen et al. [108] proposed a method called Time series Cluster Kernel (TCK) to learn the similarities between multivariate time series with missing data without using any imputation methods. This method uses an ensemble learning procedure that combined the clustering results of several Gaussian mixture models (GMM) from the final kernel to deal with uncertainty. The main drawback of this method is that it works only on datasets of equal length, also it needs ensemble learning with numerous learning datasets.Recently, Li et al. [97] proposed a multivariate time series clustering of weighted fuzzy features based on two distance measurement methods DTW and SBD. They first pick initial cluster centers by fast search and find of density peaks (DPC), then a fuzzy membership matrix is generated by performing DTW on each diminution (univariate time series), then SBD is utilised to measure distances within each dimension and generate fuzzy membership matrices which is used with the fuzzy c-means clustering algorithm.

2.7 Limitations

As discussed in this chapter, air pollution is a major problem. Several studies in the field of air quality applied data mining and time series analysis approaches to investigate the correlation between air quality, meteorological conditions, location, time and human health or to predict the air quality index or some of the air pollutant concentrations.

Based on the literature above, it can be seen that the models used to solve the problem have some limitations. For example, prediction models are trained on limited data for few years, data for one city only, a small number of monitoring sites, or specific environment types such as urban or industrial sites. The proposed models are used to predict/forecast limited air pollutants, and may not be suitable for all air pollutants. Also, most of the proposed air quality models do not deal with missing data. The missing data is basically removed in the pre-processing step. Studies that include the spatial-temporal analysis gave accurate results for the air quality index but they are limited.

Data mining techniques have been widely applied to study the air pollution data; the existing research on pollutant forecasting is limited to using artificial neural networks or support vector machines [11]. In addition, most of this research focuses only on a single pollutant (univariate TS), while clustering multivariate time series remains a challenging task [98].

2.8 Summary

In this chapter, we reviewed all the relevant material to this research. We started by introducing the UK's air quality, including the monitoring network, assessment process, and air pollution data problems. Then, we presented the most well-known data mining approaches, focusing more on the clustering process and its evaluation measures to select the best clustering solutions, including a short introduction

to the time series analysis and distance measures. Also, we investigated missing data imputation models and the evaluation metrics that measure how plausible the imputed values are. At the end of this chapter, we presented an extensive review of the proposed data mining, machine learning approaches, and time series analysis that have been applied to air quality data and some of their limitations.

Chapter 3

Research Methodology Design

This chapter covers the research methodologies we used to solve problems related to air pollution data. The main focus is the problem of missing data, including missing observations within TS and missing pollutants in a station (whole TS). Section 3.1 explains the strategy we used throughout the thesis to solve missing data problems and test the effectiveness of our proposed clustering and imputation methods.

Section 3.2 is the first stage of our framework, which covers the data pre-processing stage including imputation of missing values within the time series. Section 3.3 is the second stage, which contains clustering and evaluation processes applied on the training dataset (i.e. data for years 2015-2017). Section 3.4 is the third stage of our framework. This stage includes applying all the proposed imputation models for missing pollutant first, then evaluating these imputation results to select the best imputation model that gives the most plausible values. Section 3.5, provides a description of the air pollution dataset and Lamb Weather Types (LWTs) dataset that includes a daily classification of the UK weather, which is used to evaluate the proposed imputation models. Finally, Section 3.6 is a summary of this chapter.

3.1 Air Pollution Clustering and Imputation Framework

As we mentioned previously, the daily air quality index (DAQI) is calculated with a high level of uncertainty in some stations, as we have high level of missing observations and pollutants because stations do not measure all pollutants or do not measure a pollutant all of the time. Hence, we have two levels of missing data in time series: missing observations within the time series (TS), that represents the hourly concentrations of a pollutant which were not captured, and missing pollutants at a station (missing a whole TS). We focus on the four main pollutants that influence the UK's pollution levels. These four pollutants are nitrogen dioxide (NO_2), ozone (O_3), particles $< 2.5\,\mu m$ ($PM_{2.5}$), and particles $< 10\,\mu m$ (PM_{10}).

In this research, we apply different approaches to solve these problems, and we evaluate them to select the most reliable approach that gives the most plausible imputed values.

Our proposed solution contains three main stages, as shown in Figure 3.1: the first stage is the pre-processing stage that includes missing observations imputation to create a complete dataset. Then the second stage is to group/cluster stations based on their temporal similarity, which means using the stations similarity in pollutant concentrations through the k-means clustering algorithm, as we explained in Section 2.2.2.1, with the selected distance measure (SBD, as discussed in Section 2.3.1). The third stage is to impute and evaluate whole missing pollutant TS using the proposed imputation models including the clustering information from the previous stage.

We selected the clustering technique in our proposed solution as clustering is an unsupervised learning task, which helps to group similar data points together to find the underlying structure of the dataset without any prior knowledge about the data (i.e. ground truth) which are not available for the air pollution data we used.

Adding to that, we do not have any previous knowledge about the relation between stations or pollutant concentrations to impute missing values. Other analytical approaches such as regression, etc. are not immediately applicable to this problem because we have the input data but no corresponding output data. Additionally, trying to regress the values for one station based on others would be difficult as we would not know which stations to use and using them all could give us misleading values and would not give us the information about the relationship between stations that we get from clustering. Hence, we use the clustering as an initial step because the information we derive with this helps us to understand the relationships between stations and the values they measure.

Clustering time series data with TS distance measures such as DTW or SBD can discover the temporal similarity between time series and help to group the monitoring stations based on the temporal similarity of the air pollutants concentrations. The clustering is then helpful in understanding which stations have similar measurements and can therefore be used in the imputation. TS clustering can be applied to both linear [16] and non-linear [99] relations. For these reasons, we selected the clustering approach to discover the correlation of the pollutant measurements in the stations and to understand the behaviour of each pollutant around the UK so that we could also infer the imputation.

In this research, we focus on partitional methods such as k-means and k-medoid because they were widely applied in the literature to time series clustering [59, 170, 175] as discussed in Chapter 2, Section 2.6.2, due to their simplicity and efficiency in different domains. Other clustering methods such as density-based methods (DBSCAN), where clusters are defined as regions of high and low-density regions, are input parameter-dependent (i.e. defining the neighbourhoods around a data point (radius) and the minimum number of neighbours). Also, these methods have high computational complexity compared to partitional methods and cannot be used with datasets with altering densities [10]. For that reason, density-based methods cannot be applied to the air pollution dataset with multiple pollutants, as

some pollutants have a stronger spatial correlation (distribution) than others such as NO_2.

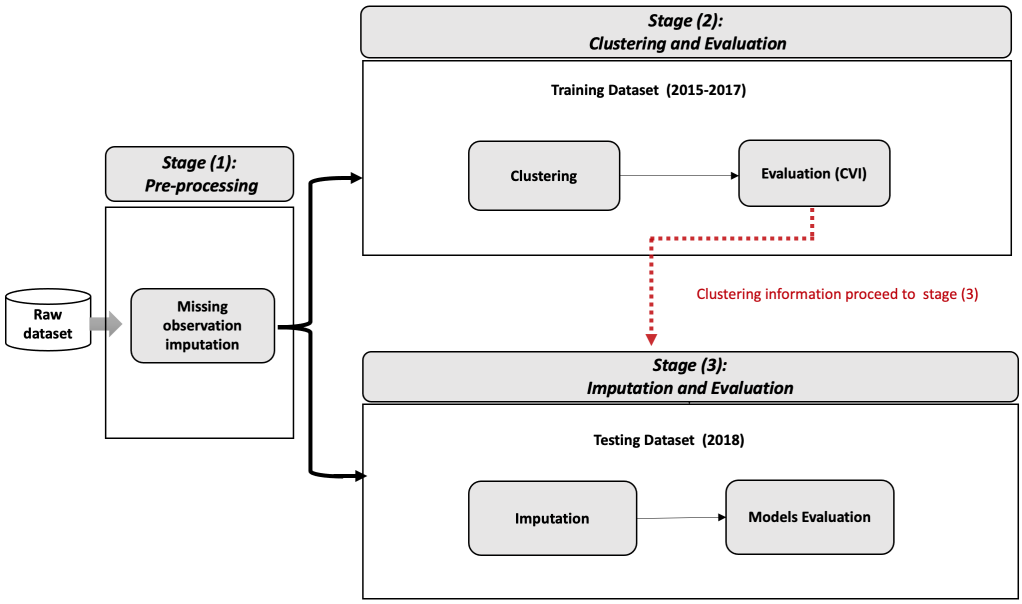


Figure 3.1: The overall proposed framework represents the main stages: the first stage is pre-processing which includes missing observations imputation; the second stage is clustering stations to similar groups based on the training dataset (data of years 2015-2017), and the third stage is imputation of missing pollutants based on test dataset (data of year 2018).

All the clustering algorithms and imputation models included in this framework were implemented in R, version (3.5.2).

3.2 Stage (1): Pre-processing

Our dataset, contains the hourly pollutants concentrations from 167 stations for a period of four years (2015-2018) as explained in Section 3.5.1. In this stage, we apply a multiple imputation method using MICE (Section 2.4.2), to impute the missing observations which is the hourly pollutant concentrations within the time series for each of the four pollutants separately.

MICE is selected based on our initial exploratory experiments. Those examined some imputation methods and how they perform in combination with clustering

algorithm for the purposed of pollutant imputation, as explained in Chapter 4. Our results showed that using MICE in combination with the clustering algorithm and the SBD measure to impute the TS missing observations is better than using some single imputation methods such as Simple Moving Average (SMA). Since MICE is a multiple imputation method, it generates multiple completed datasets (n=5 in our experiments, as five is the default number of imputations). We combine all completed MICE datasets into one by averaging each n imputed values for each individual observation creating one value. We used the average to aggregate the imputed dataset as a simple method of aggregation which still includes some uncertainty in the imputation.

This stage generates a complete dataset that can be used in the following clustering and imputation processes (in stages 2 and 3). To provide a more robust testing scenario we separate the 'model building' stage for the imputation (stage 2) from the testing stage (stage 3). Since we are going to impute the next period based on the past, it is normal to use the older data to create the model and the newer data to validate. We use an initial data period of three years (2015-2017) as a training set to build the imputations model, including the clustering results, and then impute on the next year (2018) of the TS to evaluate the goodness of fit. Using a number of years to train the model will include the pollutants seasonal variations and the resulting imputation does take account of that. Also, In the imputation models, we used the pollutants concentrations of the same time (hour) to impute the missing pollutants concentrations, hence the imputed air pollutant concentration will have similar pollutant levels. We explain each stage in the following sections.

In case of the change of the pollution level

Using a number of years to train the models will include the seasonal variations and the resulting imputation does take account of that. Also, In the imputation models, we used the pollutants concentrations of the same time (hour) to impute the missing pollutants concentrations, hence the imputed air pollutant concentration will have similar pollutant levels.

3.3 Stage (2): Time Series Clustering and Evaluation

In this stage, we use the training dataset (2015-2017) to cluster the stations based on their temporal similarity using SBD. We apply the k-means clustering algorithm to cluster the stations based on their hourly pollutant concentrations (raw data).

As it is well known in k-means that its clustering results can be influenced by the initial random seeding [149], to overcome this, we did run the algorithm several times (five times with all experiments) with different random starting points to check if the clustering solutions are stable. We experiment with two approaches: a univariate time series clustering based on individual pollutant and a multivariate time series clustering by applying an intermediate fusion approach to cluster stations based on the four pollutants. Then, we evaluate the clustering results from these approaches to select the best clustering solution using clustering validity indices.

We use the "raw" data approach directly for clustering rather than feature extraction approach, for many reasons. First, the feature extraction approach is useful when we need to reduce the dimensionality of the data, however, our aim is imputation of missing data. If we would have used feature extraction, the cluster centroids will not have captured the shape of the series, however, it is very important in our data to capture any peak of pollution episodes and those make the analyse more challenging. Visual analysis for air pollution data is very important to identify trends in data. With feature extraction, the time series data may lose interpretability, which is very important to understand the pollutant behaviour in this research, and additionally, the transformation may be expensive [25]. Adding to that, the feature- and model-based approaches are domain-dependent while raw-based approaches are suitable for almost every domain [51]. So, due to the drawbacks of the feature based approaches, we follow a raw-based approach in this research.

As we lack a reference or ground truth clustering solution, i.e. we do not know the optimal clustering solution. In this case, we have to use the internal indices, where the results are based on the clustered objects themselves. We used different internal indices to evaluate the quality of the clustering solutions. We selected Silhouette Width [127] and Dunn index [53], that measuring the cluster compactness and separation and the connectivity measure that reflects how connected objects are within the clusters [35], as explained in Section 2.2.3. In the following sections, we describe each approach in detail.

3.3.1 Approach 1: Univariate TS Clustering

In this approach, in the second stage, we use the basic k-means clustering algorithm to cluster the stations based on each pollutant independently using SBD to measure similarity between TS. So in this case, each pollutant is used to derive its own clusters and then imputation is based on that clustering solution so independent for each pollutant. These clustering results are fed to the next stage (stage 3) to impute pollutant concentrations using the proposed imputation models.

3.3.2 Approach 2: Multivariate TS Clustering (MVTS)

In this approach, we attempt an intermediate fusion approach to cluster our dataset/stations based on four TS, in our case concentrations of the air pollutants O_3, PM_{10}, $PM_{2.5}$, and NO_2. In fusion clustering, early fusion refers to an amalgamation of some of the features to create a unified subset of features for analysis; intermediary fusion refers to algorithms that somehow use fusion to operate (e.g. by fusing distances); whereas late fusion is an approach in which each modality (or in this case TS) is clustered and then the results are fused [177].

The intermediate fusion approach we use was previously proposed by Mojahed et al. [109] who used a k-medoids clustering algorithm to cluster objects that were represented by different data types (e.g. text, images and TSs). First, they se-

lected a suitable similarity measure for each data type to measure the similarities between the component of that type independently; for example, they selected cosine calculation [137] to measure the similarity between text, DTW to measure the similarity between TSs, and for images they used GIST [116]. Using those distance measures, a similarity matrix (SMs) is created for each data type. Then, those SMs are fused/merged using a weighted linear scheme based on the importance of the object into one fused similarity matrix called FM. They considered two types of uncertainty that affected the fused distances: incomplete objects that did not have recordings for a particular type of data, generating missing values in the fused matrix, and the disagreement between SMs judgments.

We take a similar approach, though as our objects are all TS we always use SBD to measure distance. From each pollutant measured in two stations we can generate a distance measure, which is then fused in the FM. We have similar uncertainty as some stations may not measure a particular pollutant, so we generate missing values in the FM. Also stations can be similar according to one pollutant but dissimilar according to another generating a second type of uncertainty for the fused distance values.

In the following sections, we explain how we combine the distance matrices of all the univariate TS into one matrix that represents the similarity between multiple time series (MVTS).

3.3.2.1 Calculation of The Fused Distance Matrix (FDM)

A Distance Matrix (DM) represents the pairwise distance between objects in a dataset. In our dataset, the distance between two stations A and B is the distance between hourly pollutant concentrations (TS) from these stations using SBD. Since we only focus on four air pollutants, we will have four DMs. We define the entries

of DM for pollutant P_i, DM_{Pi}, as follows.

$$DM_{P_i}(A, B) = SBD(A_{P_i}, B_{P_i}) \tag{3.1}$$

where A and B are two stations, and SBD is use to measure the distance between concentrations of pollutant P_i in stations A and B.

Then, we use the min-max normalization to normalize each matrix independently before fusing them. The normalization process is used to re-scale the distances with a distribution value between 0 and 1. With this re-scaling of the DM matrix, the minimum value is transformed to 0, and the maximum value to 1. We use the following formula to normalize each entry in DM_{Pi}:

$$DM_{P_i}(A, B) = \frac{SBD(A_{P_i}, B_{P_i}) - min(DM_{Pi})}{max(DM_{Pi}) - min(DM_{Pi})}. \tag{3.2}$$

where $\min(DM_{Pi})$ and $\max(DM_{Pi})$ are the minimum and the maximum distances in the matrix for pollutant P_i.

The FDM contains an entry for each pair of stations. This is the aggregated distance calculated as the simple average distance of all pollutants (O_3, PM_{10}, $PM_{2.5}$, and NO_2) when they are measured or of only those pollutants that are measured at the stations. For example for stations A and B, and supposing these stations measure all the pollutants, the fused distance between these stations is:

$$FDM(A, B) = \frac{\sum_{i=1}^{p} SBD(A_{P_i}, B_{P_i})}{p}. \tag{3.3}$$

where p is the number of pollutants. In our case $p = 4$ if all the pollutants are measured at station A and B. $SBD(A_{P_i}, B_{P_i})$ is extracted from DMs for each pollutant.

For our air pollution problem, the total number of the stations is 167. However, 10 stations are removed because they measure no common pollutant, so they cannot

deliver any information to the fused matrix. After removing these stations there are 157 stations left to construct the FDM. The removed stations are allocated to the final cluster based on their similarity to the cluster centroids.

3.3.2.2 Uncertainty

Along with FDM, uncertainty needs to be addressed. This includes two main uncertainty problems:

1. Missing pollutants: stations with missing pollutants create some missing values in the DMs. For example, if a station measures PM_{10}, $PM_{2.5}$, and NO_2, but does not measure O_3, that creates an uncertain fused distance between this station and other stations in the dataset as the O_3 distances cannot be considered in the fused value. If the station measures only one pollutant then the uncertainty of the fused distances increases.

2. Disagreement between pollutant DMs: stations can be similar in one pollutant (DM), but dissimilar in another. If all distances are say large or small, we would be more certain that the fused distance is correct, or a good representation of the distance between stations. However, if the distances between different pollutants are widely divergent then the uncertainty of the fused calculation increases. For example, we may have two stations A and B that are very similar in terms of PM_{10} and $PM_{2.5}$, but they are very different in NO_2. The divergence between these stations causes uncertainty in the fused matrix.

Below is our way to represent the uncertainty associated with the fused matrix (FDM):

Uncertainty of missing pollutants ($\mathbf{U}_P$): We represent the uncertainty of missing pollutants as a companion matrix to the FDM. To measure the uncertainty

of missing pollutants for a fused value, each value in this uncertainty matrix is associated with a value in the fused matrix. This is calculated as the proportion of missing pollutants (missing values in the DMs) associated with each fused value. We used the following formula to calculate this matrix (U_P) :

$$U_P(A,B) = \frac{1}{p}\sum_{i=1}^{p}\begin{cases} 1 & DM_{Pi}(\text{A,B}) = \text{null}, \\ 0 & DM_{Pi}(\text{A,B}) \quad \text{otherwise.} \end{cases} \tag{3.4}$$

Where p is the number of contributed pollutants in the fused matrix with a maximum value of 4. Let's say, we have two stations, A and B. Suppose station A measures four pollutants (O_3, PM_{10}, $PM_{2.5}$, and NO_2), and station B measure one pollutant (NO_2). In this case, there is one common pollutant we can measure between these stations, while the distance of all other pollutants (O_3, PM_{10}, $PM_{2.5}$) are missing. When we fuse the distance between A and B, we find that there are three missing values which are the distance of O_3, PM_{10}, and $PM_{2.5}$. The uncertainty associated with the fused distance is equal to the number of the missing pollutants divided by the total number of the pollutants, which is U_P(A,B)= 3/4.

Disagreement between pollutant DMs (U_d) : This uncertainty can also be another companion matrix to the DFM, U_d. Each value in this matrix is the standard deviation of the similarity values associated with each fused distance between two stations. Let's say, A and B are two stations that measure all four pollutants, so in each DM there is one value representing the distance between these stations for a specific pollutant. In total, we have four different values. To measure the disagreement between these values we used the standard deviation. The standard deviation therefore becomes the value of U_d(A,B) for station A and B. We used the following formula:

$$U_d(A,B) = \frac{\sqrt{\sum_{i=1}^{p}(DM_{P_i}(A,B) - \overline{DM_p(A,B)})^2}}{p}. \tag{3.5}$$

$\overline{DM}_p$(A,B) is the mean of theses values. Then we normalized the U_d using the following formula:

$$U_{dnorm}(A,B) = \frac{U_d(A,B) - min(U_d)}{max(U_d) - min(U_d)}. \tag{3.6}$$

A combined Uncertainty Matrix (UM): This uncertainty matrix represents the total uncertainty of the fused matrix which is the average of the two previous matrices (U_P and U_d). Small values of uncertainty indicate that the fused distance between two stations has a low level of uncertainty and vice versa. In the UM, the minimum uncertainty represents a case where two stations measured the four pollutants and low disagreements between DMs (from our uncertainty matrix the lowest value is 0). The maximum uncertainty represents a case where some missing pollutant where present and/or the disagreements between DMs is high (from our uncertainty matrix the highest value is 0.75). We used the average value from the uncertainty matrix (UM) as a threshold to identify if a station is certain or uncertain, with our dataset the threshold is 0.340.

For a given station S, the uncertainty is defined in relation to all other stations. We calculated the total uncertainty average associated with each station and we considered a station as uncertain if the uncertainty average exceed that threshold. For example, if only one pollutant is measured between two stations, U_P=0.75 and U_d=0 (which is the Std), UM which is the average of the two uncertainty values will be UM =0.375.

In the following sections, we introduce the modified versions of the k-means clustering algorithms we used to cluster the MVTS dataset based on calculated fused distance. We use different versions of the k-means algorithm to cluster the stations based on four air pollutants. Then, we compare these algorithms based on clustering validity indices (CVIs) to select the best clustering solution to use in the imputation process.

3.3.3 Basic k-means MVTS Clustering:

We incorporate the uncertainty of the fused matrix into various versions of the k-means clustering algorithm. For example, we devise a weighted k-means algorithm which uses the captured uncertainty of a given station to weight the selection of cluster centroids. We also devise a two-phase algorithm which first clusters certain objects (stations) and then assigns uncertain objects.

For our MVTS data which consist of 4 pollutants a cluster centroid is also a MVTS. To calculate a cluster centroid, a new TS is created for each pollutant using the average of the TS from all the stations within the cluster. Associated with that, a new FDM is calculated by measuring the distance between all the objects (stations in our application) and the centroids for each cluster, then fuse these distances using the simple average as explained in Section 3.3.2.1.

We apply the basic k-means to cluster the objects (stations) based on the fused distance matrix without using the uncertainty of the objects. We start the process using randomly selected stations (this would be a medoid in clustering) but after the first iteration we compute proper centroids. The processes of running the basic k-means on the fused distance matrix (FDM) is as follows:

Initialisation:

1. Randomly select k objects/stations as the initial centroids to start with.
2. Assign all objects to the nearest centroids based on the initial fused distance matrix (FDM).

Repeat:

1. Calculate the centroid of each cluster. The centroid will now be the average of the TS from all the stations within the cluster. Since we have 4 pollutants, every centroid will have 4 time series, one for each pollutant.

2. Calculate distances between all the stations and the new centroids, therefore creating a new FDM to represent distance between the new centroids and all the data objects.
3. Re-assign the objects to the closest centroid based on the new calculated FDM (from the previous step).

Until: No change in the cluster centroids.

3.3.4 Weighted k-means Clustering Algorithm.

In this algorithm, we use the average of the uncertainty as the weight of the object/station to the cluster centroid. As we explained in Section 3.3.2.2, the average uncertainty for each station is calculated from the Uncertainty Matrix (UM). These values represent the total uncertainty of each station to all other stations in the dataset. With the clustering algorithm, the average uncertainty of an object is calculated within the cluster. In this case, the uncertainty average for an object is changing when there is a change in the cluster members, it may increase or decrease based on the members of the cluster. The average uncertainty for all the objects is updated therefore with every iteration.

Objects with low average uncertainty (more certain objects) will contribute to the cluster centroid more than objects with a higher average of uncertainty, and that means that the stations with low uncertainty influence the centroids more than stations with high uncertainty. In the following steps, we explain the process of the weighted k-means clustering algorithm:

1. Randomly, select k objects as the initial centroids.
2. Assign all objects to the nearest centroids based on the initial fused distance matrix (FDM).
3. Calculate the uncertainty average for each object based on its cluster.

4. Re-calculate the centroid of each cluster, using the uncertainty average of the objects as the weight. The objects will contribute to the calculation of the centroid based on their weight.

5. Calculate the distance between all the objects and the recent weighted centroids for each individual pollutant and then average them, to create a new fused distance matrix (FDM) between the new centroids and all the data objects.

6. Re-assign the objects to the closest centroid based on the new calculated FDM (from step 5).

7. Repeat steps (3 to 6) until no change in the cluster centroids or the iteration reach its maximum.

3.3.5 Two Phases k-means Clustering Algorithm.

This clustering algorithm is a mixture between the basic k-means and the weighted k-means. In the first phase, we start by clustering certain objects only, which are the objects with average uncertainty less than 0.340 in this specific application. These objects will construct the clusters first and contribute to the calculation of the centroids equally. At the end of this phase, since these objects are certain, we give each object weight of 1 for the next phase.

In the second phase, the uncertain objects are assigned to nearest centroids based on the calculated fused distance between the centroids and all these objects. The average of the uncertainty of these objects will be their weight to contribute to the centroid calculation. The processes of this clustering algorithm is as follows:

Phase 1:

Running the basic k-means clustering algorithm using certain objects only (objects with average uncertainty < 0.340). The number of the certain objects used in this phase is 86 objects/stations out of 167 stations.

Phase 2:

1. Calculate the distance between all the uncertain objects (objects with average uncertainty >= 0.340) and the recent centroids for each individual pollutant and then average them, to create a new fused distance matrix (FDM) between the new centroids and all the uncertain objects.

2. Assign the uncertain objects to the nearest centroids based on the calculated fused distance matrix (FDM).

3. Calculate the average of uncertainty for all objects (certain and uncertain) within the cluster, then assign weight of 1 to the certain objects from the first phase.

4. Re-calculate the weighted centroid of each cluster using the weight of the objects.

5. Calculate the fused distance matrix (FDM) between the new centroids and all the objects.

6. Re-assign all objects to the nearest centroids.

7. Repeat steps (3 to 6) until no change in the cluster centroids or the iteration reach its maximum.

3.4 Stage (3): Imputation and Evaluation Models for Missing Pollutants

In this stage, we impute the missing pollutant, which is a whole TS in the test dataset (data of year 2018). We propose different imputation models under two main similarity criteria: the similarity using clustering solutions and the similarity using geographical distance. Adding to that, we use an ensemble model which calculates the median of all the previous imputation models; this model aggregates the temporal and the spatial imputation using both the time series clustering and the geographical location similarity.

Then, we compare our imputed/modelled values to those observed values in order to evaluate and select the model that gives the highest similarity to the observed values. For any given station, j, to impute the values of missing pollutant $P_i^j, 1 \leq i \leq 4$, in our experimental set up we take each existing Time Series (TS) for a given pollutant and station, P_i^j in turn, and impute it by the various imputation models to obtain an imputed TSs, PI_i^j. This enables evaluation with a 'ground truth'.

3.4.1 Imputation models.

3.4.1.1 Imputation models based on time series clustering:

Once a clustering of our stations is obtained, we can use the clustering solution to impute missing TS (pollutants). If station j belongs to cluster C_x, $(1 \leq x \leq k$, where k is the number of clusters) given the measured pollutants over time, then, to impute pollutant P_i based on the clustering results, we use three models:

1. **Cluster Average (CA):** We impute the average of pollutant P_i in cluster c_i, which is the hourly average of pollutant P_i in all the stations that fall in this cluster.

2. **Cluster average and station environment type (CA+ENV):** We impute the average of pollutant P_i in cluster c_i, but using only stations with the same environment type to station j within the cluster, such as 'Background Rural', 'Background Urban','Traffic', or 'Industrial'. This is in recognition that the type of station may be important and result in closer measurements of pollutant concentrations.

3. **Cluster average and station region (CA+REG):** We impute the average of pollutant P_i in cluster c_i, for stations that belong to the same region. As defined by Defra [5] there are 16 regions in the UK for air quality assessment, such as Eastern, North Wales, East Midlands, and the other UK regions (More details in Table A.1).

3.4.1.2 Imputation models by similarity using geographical distance:

First, we measure the geographic distance using Haversine metric (Equation. 3.7), which calculates geographic distance on earth based on longitude and latitude. We calculate the distance between station j and all other stations that measure pollutant P_i, then we use stations nearest neighbours in this imputation, to impute pollutant P_i for station j. Haversine distance d between two points can be computed as:

$$\text{d= 2.R.arcsin}\left(\sqrt{\sin^2(\tfrac{\varphi 2-\varphi 1}{2}) + \cos(\varphi 1)\cos(\varphi 2)\sin^2(\tfrac{\lambda 2-\lambda 1}{2})}\right) \tag{3.7}$$

Where φ represents latitude in radians , λ represents longitude in radians, R is earth's radius (mean radius = 6,371 KM), and *arcsin* is the inverse of *sine* function.

1. **The nearest neighbour (1NN):** We impute the missing pollutant P_i in a station j using its nearest neighbour based on the Haversine distance to station j.

2. **The nearest neighbour with same station type (1NN.ENV):** We impute the missing pollutant P_i in a station j using its nearest neighbour with same environment type of station j.

3. **The average of the two nearest neighbours (2NN):** We impute the missing pollutant P_i in a station j using the average of the two nearest neighbours/ stations to station j.

4. **The average of the two nearest neighbours with same station type (2NN.ENV):** We impute the missing pollutant P_i in a station j using the average of the two nearest neighbours/stations with same environment type of station j.

3.4.1.3 Imputation model by ensemble:

In this model, for a given station j, to impute pollutant P_i , we use the median value of all the imputed values from the previous models. Hence, the environment type of the station plays an important role in pollutant imputation. We have two ensemble models under this approach:

1. **Median:** We impute the missing pollutant P_i in a station j by taking the median value of cluster average (CA), cluster average considering the station type (CA+ENV), cluster average considering the region (CA+REG), first nearest neighbour (1NN), and the average of the two nearest neighbours (2NN) for station j.

2. **(Median.ENV):** We impute the missing pollutant P_i in a station j by taking the median value of (CA), (CA+ENV), (CA+REG), with nearest neighbours imputation models that consider station environment type (i.e. 1NN.ENV and 2NN.ENV).

3.4.2 Imputation Model Evaluation.

Model evaluation functions are beneficial when more than one model is involved in the comparison, and help us understand why a model does not perform very well. Hence, as we propose multiple imputation models, to evaluate the modelled values, we compare them with the observed values using different statistical and graphical models evaluation functions. These functions are part of Openair Package, a freely available air quality data analysis tool based on R statistical software [33]. After selecting the model that gives the best imputed values based on the statistical and graphical functions, we further evaluate the modelled values based on the air quality index value (DAQI) and then measure the model perform under different weather types.

We select several different error measures, based on the literature which are the most common measures in air pollution studies. We do not relay on statistical measures alone, instead we use other graphical model evaluation functions that are mainly used to evaluate air pollution modeling.

For statistical model evaluation metrics, we use Root Mean Squared Error (RMSE) (2.38), Coefficient of correlation (R) (2.39), Index of Agreement (IOA) (2.40), Mean Bias (MB) (2.34), and Normalised Mean Bias (NMB) (2.36) as defined in Section 2.4.3. These measures are used to evaluate the temporal variation of air pollutants between imputed/modelled and observed concentrations. These metrics are part of 'modStats' function that belongs to Openair Package. Based on these statistic metrics, the model that gives the lowest average of error, the highest correlation and the highest degree of agreement between imputed and observed concentrations for all stations (i.e. imputed TS) will be considered the best model. Note that the best model may change from one pollutant to another and may be affected by other factors such as station type (e.g. urban background, rural and roadside)

Graphical model evaluation functions help to analyse the graphical variations, correlation, and the distribution of imputed and observed air pollutant concentrations. We use three functions that are part of Openair Package as well:

$'TimeVariation'$, which plots the diurnal, day of the week and monthly variation between modelled and observed concentrations, typically pollutant concentrations. It produces four plots: day of the week variation, mean hour of day variation and a combined hour of day, day of week plot and a monthly plot. Also shows the 95% confidence interval for the mean.

'$ConditionalQuantile'$, which calculates conditional quantiles for the modelled concentrations and represents how well modelled concentrations agree with observations.

'$TaylorDiagram'$, this function plots a diagram that shows three model performance statistics; the correlation coefficient (R), the standard deviation (sigma), and

the root mean square error (centred). These statistics can be plotted on one (2D) graph which can be represented through the Law of Cosines [151].

After selecting the model that gives the best imputed values, we calculate DAQI from the modelled concentrations and compare it with the observed DAQI. We use this as a performance tool to evaluate our imputation model on its ability to reproduce the daily air quality index.

First we calculated the daily DAQI value using the observed data for each station. This is because the DAQI value is not saved as part of the historical data available so we need to calculate it from the downloaded data. Defra has published a guide for the implementation of DAQI [46], which explains how the value is calculated and we follow that guidance. To calculate DAQI, each air pollutant is calculated as follows:

- Ozone: O_3 is measured hourly. To determine the DAQI we need to calculate the daily maximum 8-hourly running mean concentration. First, for each hour we calculate the running 8-hourly mean from the previous hours. Then we find the maximum value of these 8-hourly running means. For this calculation 75% of the data must be captured to calculate the 8-hourly mean.

- Nitrogen dioxide: NO_2 is measured based on hourly mean. We calculate the daily NO_2 contribution to the DAQI by taking the maximum observation in 24 hours every day from 0:00 to 23:00.

- Particles PM_{10} and $PM_{2.5}$: are measured hourly. The DAQI is based on the 24 hours mean, which we calculate by taking the mean value from the hourly observations. For these pollutants 75% of the daily observations must be captured to calculate the mean, otherwise, the pollutant is considered as missing that day.

- We define the daily index for each pollutant separately. Then, for a station, we take the highest air pollutant index to be the value of the DAQI at that

station.

We called the DAQI that is calculated based on observation 'observed DAQI', and the DAQI that is calculated based on imputation 'imputed DAQI'. In this stage, we compare the imputed DAQI with the observed DAQI based on RMSE, and the number of days where there are agreements and disagreements.

In our evaluation, we use RMSE as an initial step to measure how close the imputed/modelled data is to the real time series in the training set so that we can select the best imputation model that gives the lowest average of errors with the support of other statistical measures that agree with the RMSE such as Coefficient of correlation, index of Agreement..etc. Importantly, in the evaluation process, we don't rely on the RMSE alone to select the best imputation method, but we used several graphical and statistical evaluation functions that represent the uncertainty between modelled and observed TS such as: '$ConditionalQuantile'$ and '$TaylorDiagram'$. In addition, the DAQI has been used as an evaluation measure. After selecting the best imputation model, a new DAQI index is calculated based on the observed + imputed data (i.e called imputed DAQI) and then we compare this with the DAQI index calculated from observed data (i.e called observed DAQI). The comparison is based on the number of days where there is an agreement or disagreement between imputed and observed DAQI. In this comparison, we focus on analysing any variation between these indices that may indicates some pollution episodes associated with unmeasured/imputed pollutants.

We also analyse the performance of the best imputation model based on the daily modelled concentrations under different weather types using Lamb Weather Types (LWTs) to see how the models perform under different weather conditions. This analysis is based on statistical and graphical model evaluation functions.

LWTs are a daily classification of the UK weather over the British Isles, proposed by Lamb [91]. Later, a new objective scheme containing 27 Lamb weather type classifications was added by Jenkinson and Collinson [84], as shown in Table 3.1. The

main classifications are from 0 to 28. In this research, we followed the same grouping as in a recent study by on Graham et at. [67]. In Jenkinson and Collinson's dataset we used, there are 11 LWTs: NE, E, SE, S, SW, W, NW, N, Anticyclonic, Neutral, and Cyclonic. Within this classification, there are also eight wind types: NE, E, SE, S, SW, W, NW and N, shown in the left columns of the table, and two circulation types (anticyclonic, cyclonic). Neutral days, are where wind speed and shear were too low to allow classification. There is also one LWT (-9: non-existent day) that is not used.

Table 3.1: Lamb weather classification based on Jenkinson and Collinson [84]

Wind directions	**Anticyclonic**		**Neutral**		**Cyclonic**	
			-1	UC	-9	non-existent day
-	0	A	-		20	C
NE	1	ANE	11	NE	21	CNE
E	2	AE	12	E	22	CE
SE	3	ASE	13	SE	23	CSE
S	4	AS	14	S	24	CS
SW	5	ASW	15	SW	25	CSW
W	6	AW	16	W	26	CW
NW	7	ANW	17	NW	27	CNW
N	8	AN	18	N	28	CN

We proposed this approach rather than using a simpler approach, because one major issue is that the multivariate time series objects (stations) in our dataset (i.e. the UK's air pollution data) do not have equal dimensions as not all pollutants TS are recorded in the stations. Multivariate time series comparisons are possible [158] but only if the multiple time series have the same dimensionality. These methods allow us to compare two multivariate time series as long as objects/stations contain the same number of variables/pollutant concentrations. In that case the only method that has been applied is to fuse the similarity of multiple pollutants between stations that are based on the average similarity of the univariate time series.

The proposed MVTS clustering approach enables us to impute missing (unmeasured) pollutants in any station, because the clustering assigns the station into a

cluster based on the available data (i.e. the other measured pollutants). In contrast, the individual clustering approach cannot assign a station to a cluster if it does not measure that pollutant.

On the other hand, the ensemble model can achieve better performance than any single model, as it is a multiple imputation models that aggregate the spatial and temporal similarity to capture different characteristics of data from different methods and to make better predictions for the four pollutants.

3.5 Datasets:

3.5.1 Air Pollutants Concentrations Dataset

The datasets generated at AURN stations include multivariate TS data that show the hourly concentrations of air pollutants. Our dataset for the period of four years (2015-2018). It contains data from 167 monitoring sites cross the UK. Table 2.2 shows the number of stations that measures each pollutant in this dataset. These stations categorized under six environmental types (Background Rural, Background Urban, Background Suburban, Industrial Suburban, Industrial Suburban, and Traffic Urban) as shown previously in Table 2.1. The data can be obtained from `https://uk-air.defra.gov.uk/data/data_selector`. The observations we download from each station include:

- Date;
- Time;
- Hourly observations for all measured pollutants by the station;
- Status (include R =Ratified; P=Provisional; P*=As supplied);

The total number of observations collected from 167 stations for the selected period is 11,040,012 observations for the measured pollutants. The table below (Table 3.2)

shows the basic statistic of the collected dataset and the percentage of the missing observations for each pollutant separately.

Table 3.2: Statistical presentation of air pollutants concentrations dataset for the period of four years (2015-2018).

Pollutant	**O_3**	**NO_2**	**PM_{10}**	**$PM_{2.5}$**	**Total**
Total number of observations	2454480	5505048	2629800	2699928	13289256
Total number of recorded observations	2294202	4568399	2031643	2145768	11040012
Missing observations	160278	936649	598157	554160	2249244
Percentage of missing observation (%)	6%	17%	%22	%20	-
Min	-3.2	-2.1	-4.8	-6	-
Max	224	397	971	472	-
Mean	47.3	25.3	16.2	10	-
Std	24.1	23.1	12.5	9.1	-

3.5.2 Weather Dataset: Lamb Weather Types (LWTs)

The Lamb Weather Types (LWTs) can be obtained from `https://crudata.uea.ac.uk/cru/data/lwt/`. The Jenkinson and Collinson's dataset we download for the year of 2018, this dataset contains 365 days/records, each record has the following :

- Day;
- Month;
- Year;
- PM_1000: Average pressure over the grid points;
- W: Westerly flow;
- S: Southerly flow;
- F: Resultant flow;
- Z: Total shear vorticity;
- G: Gale day;

- Dir: Direction of flow;
- LWT: The daily classification of the weather.

3.6 Summary

In this chapter we introduced the clustering and imputation framework that contains three main stages: Stage 1, where the data pre-processing executes to prepare the data for the clustering algorithm. This stage includes missing value imputation within time series and divides the dataset into training and testing sets.

The second stage includes grouping the stations based on their similarity using a clustering algorithm. In this stage, we follow two approaches to cluster the time series (stations). The first approach is to cluster stations based on individual pollutant, we called this approach the univariate time series clustering. The second approach is to cluster the stations based on all measured pollutants, this is called the multivariate time series clustering (MVTS). The clustering results from this stage are fed in the final stage to impute the missing pollutant time series with imputation models that are based on clustering. The third stage, includes applying all the proposed imputation models, evaluating the imputation results and selecting the best model that gives the most plausible values for each pollutant. In this chapter, we have explained the evaluation methods that we will use to evaluate our proposed imputation models. At the end of this chapter, we included a top level explanation of the datasets that will be used in this research.

Chapter 4

Initial Exploratory Experiment in Clustering Imputation for Air Pollution Data

This chapter covers the initial exploratory experiment in clustering imputation approach. In this experiment, we focus on the ozone (O_3) dataset, which is one of the main pollutants that influence the pollution levels in the UK and the most complicated pollutant. The main goal of this experiment is to select the best imputation method for missing values within the time series in combination with the clustering algorithm and to evaluate how well clustering for the purpose of the pollutant imputation

In this chapter, Section 4.1 is a general introduction to the experiment, Section 4.2 is our designed experiment framework to examine the proposed approach, and Section 4.3 is the results of applying multiple approaches to ozone dataset. Finally, Section 4.4 is a conclusion of this chapter.

4.1 Initial Approach for Clustering Imputation

In this experiment, we examine different approaches to impute the whole time series for a missing pollutant at a monitoring station as well as missing values within a time series.

To develop our approach we combine single and multiple imputation to impute missing observations within TS, nearest neighbour geographical distance method and imputation based TS clustering method through the clustering using two different distance measures (DTW and SBD), these methods are explained in Chapter 3. The ozone (O_3) dataset was selected to validate the proposed methods in this experiment.

For each station that measures ozone, we produce various imputations for this pollutant, then we compare the imputed with the real values using the Root mean squared error (RMSE) (2.38), and its Standard deviation (Std). For missing observations imputation, we use the simple moving average (SMA) as a single imputation method, and MICE as the multiple imputation method. The SMA method replaces each missing value using a weighted moving average. The mean value in this method is calculated from an equal number of observations on either side of a central missing value; the user can identify the length of that window [110]. In our experiment, we set the window length to 30 days (a month), so the missing value is replaced by the monthly moving average before and after the missing value. On the other hand, since MICE is a multiple imputation method, the imputation generates five completed datasets as explained in Section 2.4.2.

For clustering, we start by using k-medoids (PAM) clustering algorithm to cluster the stations based on DTW and SBD.

We select these methods based on the literature, DTW is the most common time series distance measure to analyse TS data, SBD is based on DTW and cross-correlation which would be applicable to show the correlation of pollutant concen-

trations in different locations, and the k-medoids algorithm is selected due to its ability to deal with outliers and noisy data [87].

In the imputation process for the whole TS, we use the cluster medoid to impute the missing pollutants at a station, we called this model **(CM)**, in addition to other proposed imputation models that are: the cluster average **(CA)**, the first nearest neighbour **(1NN)** and the average of the two nearest neighbours **(2NN)**, as explained in Section 3.4. Then we measure the similarity between the imputed and the real values using the RMSE, and its Standard deviation (Std) to select the method that give better imputation.

The ozone dataset, contains 70 stations distributed around the UK. We removed 5 stations that have more than 25% of missing data in this experiment. In total, we included 65 stations in our analysis. In this dataset, we use the daily mean concentrations of O_3 in the clustering and imputation process. Figure 4.1, shows the geographical distribution of all stations that measure O_3 and were included in this experiment.

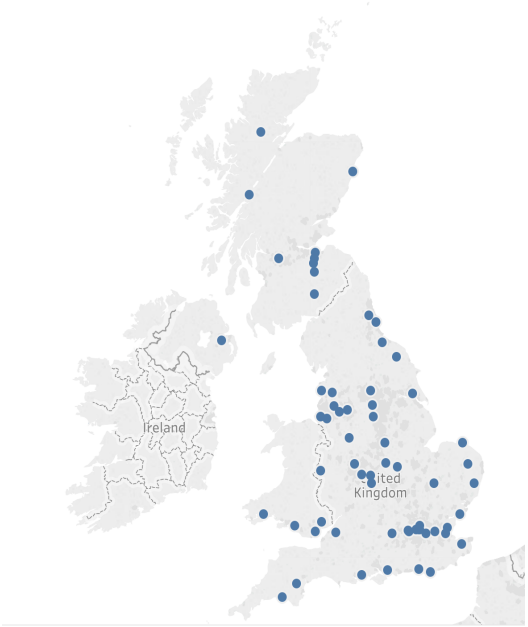


Figure 4.1: Geographical distribution of ozone monitoring stations in the UK used in the experiment.

4.2 Initial Experimental Framework

We divide our experiment into two phases: the first phase includes imputation of missing observations and clustering process to obtain the clustering solution we need for the imputation of the whole TS. This phase is built using the first part of the dataset corresponding to the period (2015-2017) (training set), as shown in Figure 4.2, and the second phase includes our proposed imputation models and their evaluation process, using the second part of the dataset corresponding to the year 2018 (test set), as shown in Figure 4.5. The proposed imputation models and clustering algorithms were implemented in R for this experiments.

4.2.1 Phase 1: Missing observations imputation and clustering process.

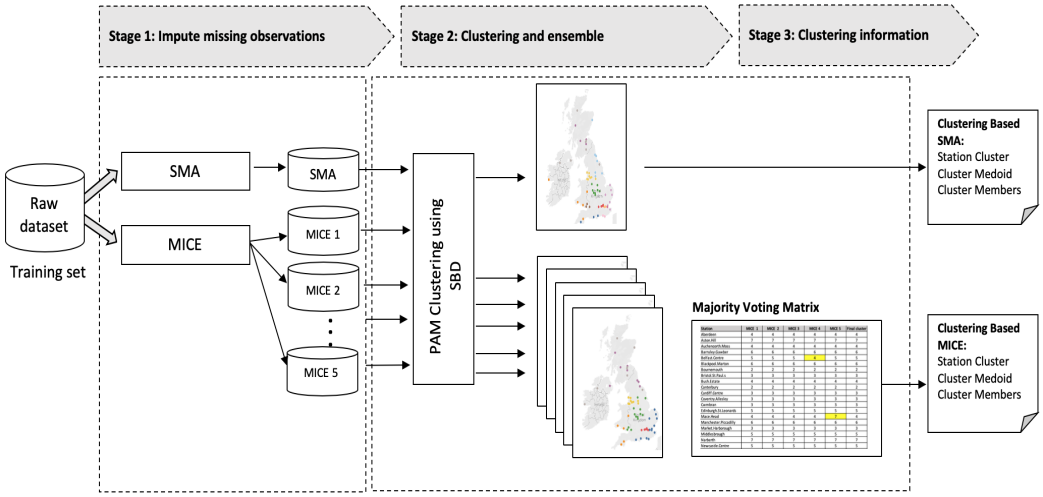


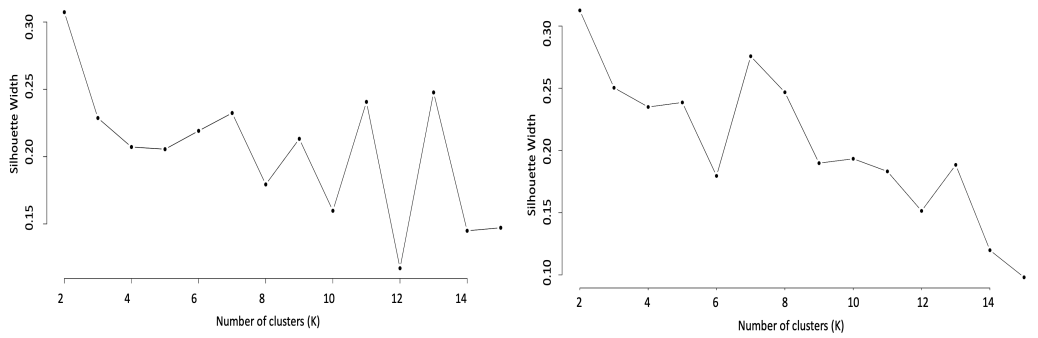
Figure 4.2: Phase 1: Time series missing observations imputation and clustering process.

In this phase, we divide the processes into three stages. In the first stage, we impute the missing observations within each TS using two techniques: SMA and MICE. As part of this process, from the incomplete dataset (raw data) we create six datasets: one dataset created by SMA, and 5 datasets created by MICE.

In the next stage, we cluster these datasets individually using PAM clustering algorithm and two distance metrics, DTW and SBD. For each dataset, we use the Silhouette Width method (Equation 2.9) to decide on the number of the clusters (value of k). In our datasets, the maximum Silhouette index is associated with k=2 in most cases. However, since the breakdown into only 2 clusters gives us little information, we select the next optimal number of k in order to allow the algorithm to create more granular clustering results.

After comparing the clustering results generated using DTW and SBD based on the Average Silhouette Width (ASW) and the distribution of the clusters, we found that using SBD gives better separated clusters than DTW, since the SBD is based on the DTW but also takes into consideration the cross-correlation between TS, we decided to use SBD in our clustering algorithm for the rest of our experiments. This applied to all pollutants we included in this thesis, Appendix B, shows the comparison of the clustering results obtained from PAM algorithm with DTW and SBD for all pollutants.

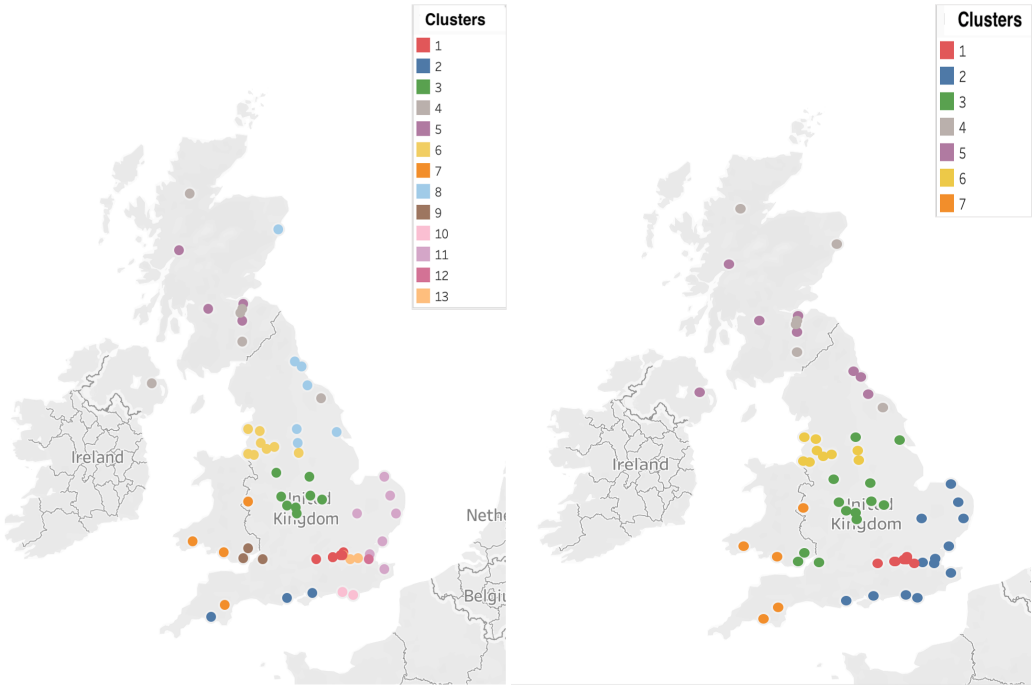
Figure 4.3, shows the optimal number of the clusters for PAM algorithm with SBD using SMA dataset, which is 13 clusters, and similarly the optimal number of the clusters using MICE dataset which is 7 clusters. The optimal number of the clusters for all 5 datasets created by MICE were the same (k=7).



a b

Figure 4.3: Average silhouette widths for 2 to 15 clusters with completed O_3 datasets: (A) dataset generated by single imputation method (i.e SMA) and (B) first datset generated by multiple imputation method (i.e MICE).

The clustering results obtained from each individual dataset created by MICE are slightly different. We merged the clustering results of these datasets into one final result using the majority voting, that reflects the amount of uncertainty in the estimates. We used the majority voting to put a station in one specific cluster and create one clustering solution from these datasets. Majority voting [50] is a simple ensemble technique. Essentially, the ensemble chooses the cluster for a station which is chosen by the majority of the clustering results. The results of this stage, is shown in Figure 4.4. As we can see in (A) we obtained 13 clusters from using the SMA method, and in (B) 7 clusters from using majority voting on the MICE dataset clustering results.



a b

Figure 4.4: Clustering results of training datasets (2015-2017) using PAM clustering algorithm with SBD distance measure on SMA dataset (A) and the combination of MICE datasets clustering (B).

In stage three, for each station we use the information obtained by the clustering such as the cluster number (cluster ID), the cluster medoids, and members of the cluster associated with each station. These results feed into the next phase to impute the whole pollutant.

4.2.2 Phase 2: Pollutant imputation methods.

This phase includes some of our proposed imputation models for TS in the test period, as shown in Figure 4.5. The imputation of missing observations takes place for the test data as it did for the train data, however in the test set we combine the MICE datasets into one by averaging the n imputed values for each individual observation creating one value. Then, based on the clustering results from the first phase, we assign the clustering information for each station such as, the station cluster, the cluster medoids, and the group of other stations that belong to the same cluster. Then the clustering results and nearest neighbours imputation (1NN and 2NN) are used to produce whole imputed TS for each station. The best imputation model is selected using the RMSE, that represents the average of errors between imputed and real values.

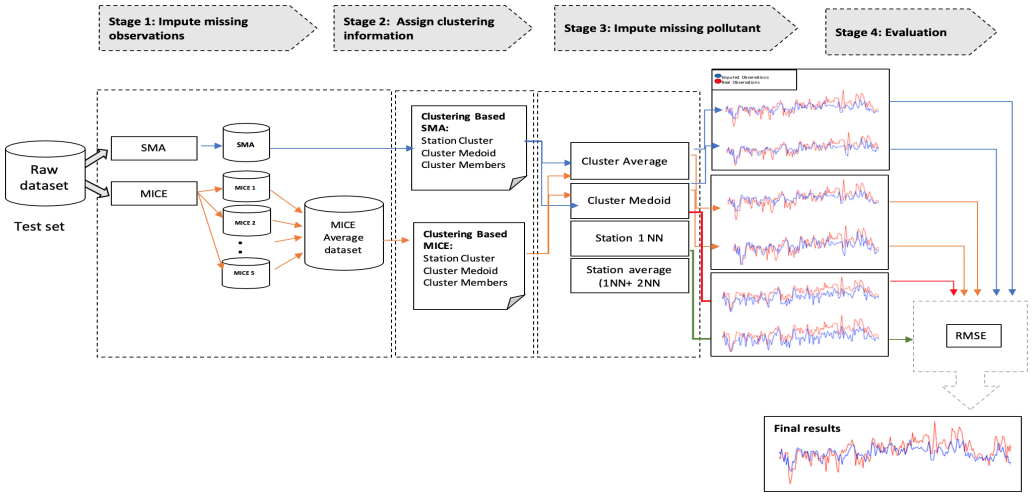


Figure 4.5: Phase 2 : Air pollutant imputation models.

4.3 Results and Discussion.

The clustering results we obtained from the first phase of this experiment, as shown in Figure 4.4, are geographically consistent. We only used the pollutants concentrations as time series to cluster the stations using SBD, which considers the temporal

similarity without any knowledge of the geographical location. Nevertheless, the clustering results show that there is a spatial (geographical) correlation between stations in each cluster.

In the second phase, according to our four described imputation models, using the Cluster Medoid (CM), the Cluster average (CA), the nearest neighbour (1NN) and the average of the 2NN (2NN), we created 8 different imputed TS, 4 for the SMA dataset and 4 for the combined MICE datasets. We evaluated those by calculating the RMSE to measure the difference between the imputed and the real data for each station. For each station, we ranked our imputation models for each dataset based on the value of the RMSE from smallest to largest, hence the best imputation model for SMA will have a rank of 1, etc., and similarly for MICE combined dataset. We then compared these models based on the average ranks to select the best imputation model. Table 4.1 shows the comparison of the average of the RMSE and the average rank for all models from all stations for both the MICE and SMA datasets, respectively, using the Cluster Average (CA) is associated with the minimum average rank (2.25, 2.37), the minimum average error (10.003, 10.315), and the minimum standard deviation (3.901, 4.218). This is followed by (2NN) and then (CM), with (1NN) providing the worst results.

Table 4.1: The average rank and RMSE and Standard Deviation (Std) between observed and imputed TS four imputation models and two dataset MICE dataset (Top table) and SMA dataset (bottom table), including number of stations contributed in each imputation.

Imputation Models	**Average Rank**	**Average Errors (RMSE)**	**Standard deviation (Std)**	**Station contributing**
		MICE Dataset		
Cluster Average (CA)	**2.25**	**10.003**	**3.901**	65
Cluster Medoid (CM)	2.78	11.410	4.544	58
First neighbor (1NN)	2.86	12.116	5.417	65
Average of 2NN	2.41	10.784	5.272	65
		SMA Dataset		
Cluster Average (CA)	2.37	10.315	4.218	64
Cluster Medoid (CM)	2.61	11.692	4.404	52
First neighbor (1NN)	3.05	12.641	5.263	65
Average of 2NN	2.53	11.266	5.121	65

We look at specific example stations (Glasgow Townhead) to compare the four

imputed TS for the period of six months (Jan-Jul of 2018) using SMA (top) and MICE (bottom) datasets. This is shown in Figure 4.6. The visualisation shows that all the imputations reproduce the trend well, though they may generate slightly higher values. Some periods, early in the year appear to show more deviation and this may be due to temperatures having an effect. Using MICE dataset with cluster average imputation model (CA) appears to produce the closest results.

Figure 4.7 shows a different station, "Glazebury". In this figure, the red TS represents the real observations at the stations and we can appreciate there are some missing values in the middle of the TS. The green TS represents imputed missing observations using SMA (Top), and MICE (bottom). On the other hand, the blue TS is the result of imputing the whole pollutant TS using the Cluster Average (CA).

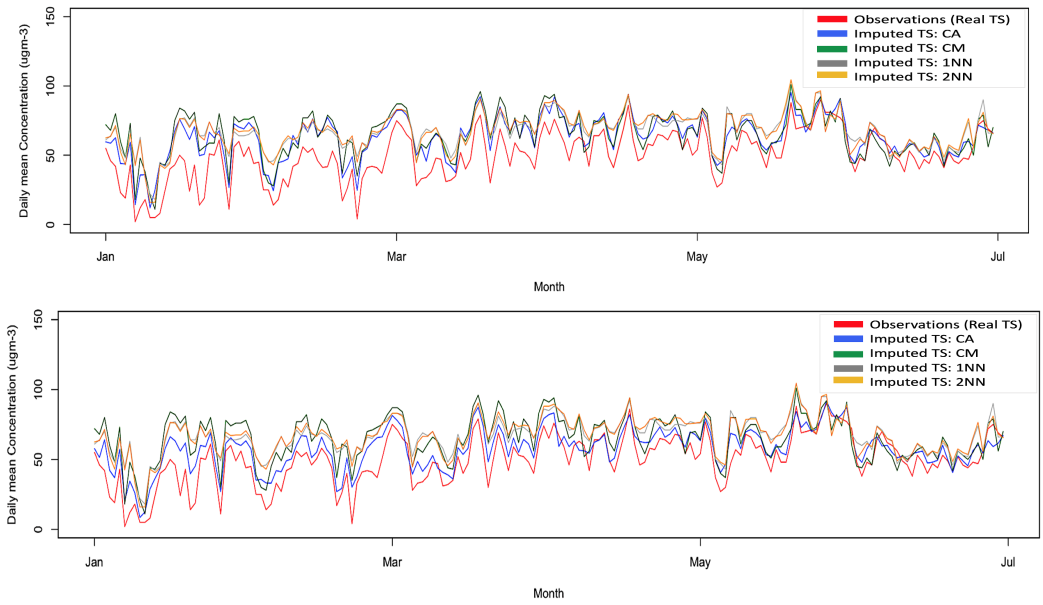


Figure 4.6: Pollutant imputation models using SMA dataset (top) and MICE dataset (bottom) at Glasgow Townhead station.

It is worth noting that using the Cluster Medoid (CM) to impute the missing pollutant is not possible in some cases. If the station we are going to impute is itself the medoid of the cluster, or if the cluster has only one station, then we have no feasible imputation hence we record how many stations the imputation

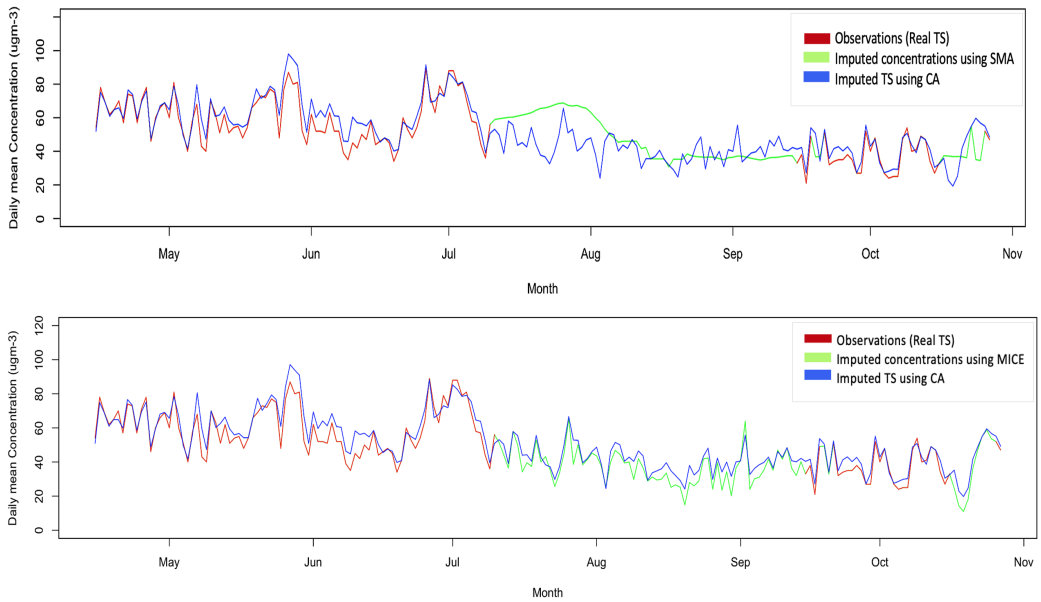


Figure 4.7: Observations and Cluster Average (CA) imputed TS using SMA dataset (top) and MICE dataset (bottom) at Glazebury station.

was possible for as the last column of Table 4.1.

More details about this imputation, can be found in Appendix C). Table C.1 in the Appendix shows that it is not possible to create a cluster average for the SMA dataset at "Rochester Stoke" station, because the cluster has only that station. In this case we cannot use the Cluster Average and the Cluster Medoid in the imputation process.

As a result of this experiment, we found that using the clustering average imputation method to impute the missing pollutants after clustering the stations on the pollutant values over time using SBD and PAM gave better results compared to other techniques. This was true regardless of the method used to impute missing observations (partial imputation). However, the combination of multiple imputation (MICE) for partial missing values and cluster average for pollutant imputation gave the best results. Also, since the cluster average (CA) gives better results than the Cluster Medoid (CM) that was not feasible in some case, we decided to use the k-means algorithm in further experiments as that does not rely on medoids.

4.4 Summary

In this chapter, we have applied and compared a number of techniques to impute ozone values for a station when missing partially or completely. We examined different imputation methods to impute the missing values within the time series, different distance measures to measure the similarity between two stations with PAM clustering algorithm, and different imputation models to impute the whole TS. The main goal of this experiments was to select the method that gave better imputation for the missing values to be able to apply clustering algorithm to cluster stations based on their similarity for the proposed of whole TS imputation. Also, to select the distance metric that gave better clustering results that can represents ozone concentrations around the UK.

As a result of our experiments, we found that using SBD to measure the distance between two pollutants concentrations gave better clustering results than using DTW. Also, using SBD in combination with MICE gave better imputation for whole TS than other single imputation method. In the next chapters, we will apply similar approaches that are based on MICE and the k-means algorithm since the Cluster Average (CA) is more useful than Cluster Medoid (CM) for the purpose of the pollutant imputation. Hence this set of experiments enabled us to select a number of components for our further methods based on their performance in these controlled experimental environment.

Chapter 5

Results of Applying Univariate Time Series Clustering and Imputation

This chapter focuses on univariate TS clustering for pollutant imputation (whole TS), that was introduced in chapter 3.

In this chapter, Section 5.1 starts with the general framework for the experiments. Section 5.2 discusses the results of applying the k-means clustering algorithm to cluster $PM_{2.5}$, PM_{10}, O_3, and NO_3 stations respectively. Section 5.3 includes the clustering evaluation results for each pollutant based on some clustering validity indices (CVIs) to assess cluster compactness and separation, then an evaluation of how good these clusters are for the purposes of pollutant imputation by measuring the RMSE and its standard deviation (Std) between imputed and real pollutant concentrations. Finally, Section 5.4 is a summary of this chapter.

5.1 Air Pollution Imputation Framework Based on Univariate TS Clustering

We followed the main stages of our proposed method as explained in Section 3.1.

In the first stage, as in the previous experiment, we start by imputing the missing observations within the TS using MICE for whole dataset (2015-2018), then we average all the completed datasets created by MICE into one dataset. This dataset is used for the clustering in the next stage. In the second stage, we start by measuring the similarity between stations using SBD; then we apply the basic k-means clustering algorithm to cluster stations. In this approach each pollutant is clustered independently. Then, in the final stage, the clustering results for each individual pollutant is used to impute the pollutant concentrations using the proposed imputation models (Section 3.4.1.1). Figure 5.1 shows the experimental framework for the univariate TS clustering and imputation approach (Approach 1).

For the purpose of this experiment we only used cluster average (CA) and cluster average with station environment types (CA+ENV) imputation models in order to evaluate how good this clustering approach is for pollutant imputation.

5.2 Experimental Results

5.2.1 $PM_{2.5}$ Clustering Results

The total number of stations that measure $PM_{2.5}$ is 77 stations. The result of applying the basic k-means clustering algorithm to this set of stations is shown in Figure 5.2, which represents a geographical map with the stations colour coded according to the clustering results (i.e. (cluster 1, (red), (cluster 2, green), (cluster 3, blue), and (cluster 4, purple)). There are four clusters located in four geographical locations (North, Center, South East, South West). These clusters geographically look compact and well separated even though the clustering is based on pollutant concentration values. This means that there appears to be a geographic pattern to the concentrations of $PM_{2.5}$

To further analyse results, we show the time variation between the clusters' centroids

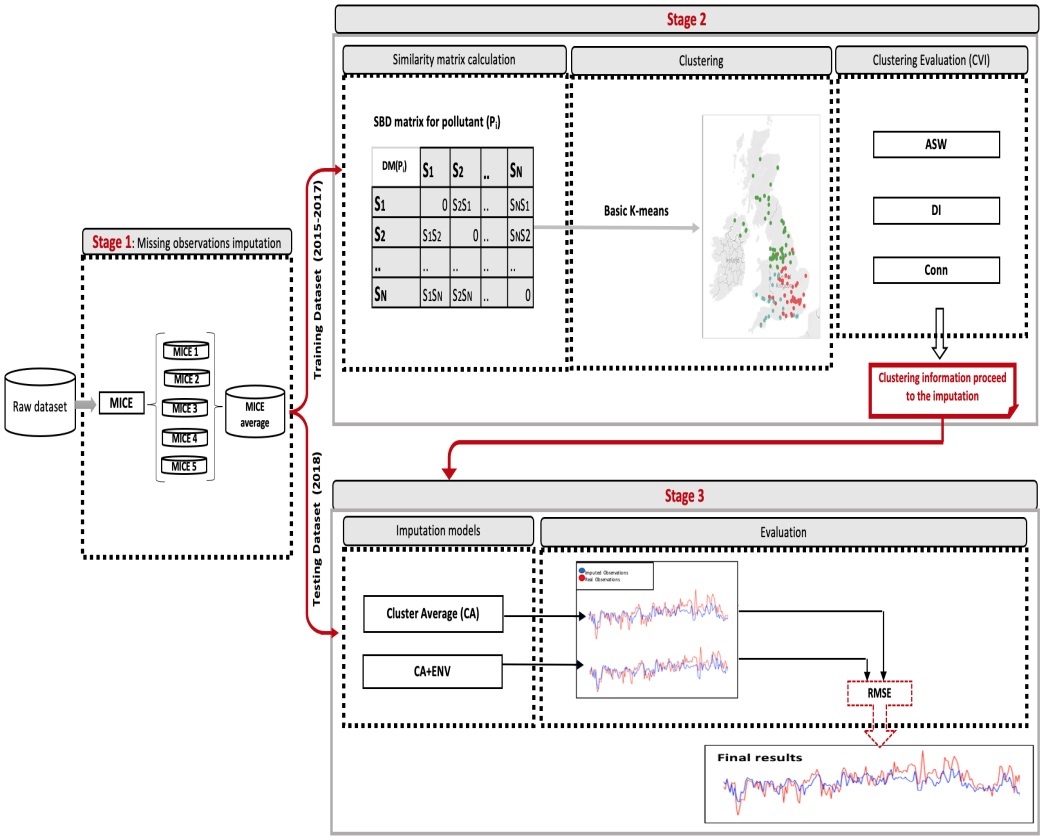


Figure 5.1: Experiment stages of the univariate clustering and imputation (Approach 1).

(and we will do similarly for the other pollutants). This enables us to further understand how $PM_{2.5}$ concentrations are distributed in the UK and if there are specific time effects. The cluster centroids represent the average concentrations in the clusters for a particular pollutant. Figure 5.3, represents the variation in pollutant concentrations of cluster centroids on each day of the week in the top graphs. Then the variation is broken down into hourly, monthly and weekday variations in the bottom graphs. From this figure, we observe the variation of $PM_{2.5}$ concentrations at each cluster centroid, it is noticed that centroid's of cluster 2 (green) is the highest, while concentrations at centroid's of cluster 3 (blue) is the lowest. While the concentrations of remaining two centroids of cluster 1 and 4 (red and purple) have similar behaviours. This variation can also be seen with the monthly, weekly and hourly analysis.

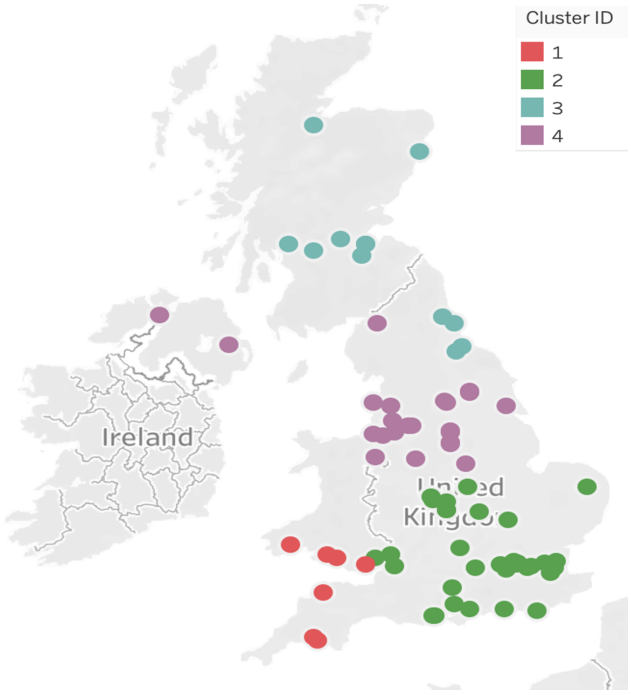


Figure 5.2: Geographical distribution of stations that measure $PM_{2.5}$ with the colour coded clusters obtained using the basic k-means algorithm.

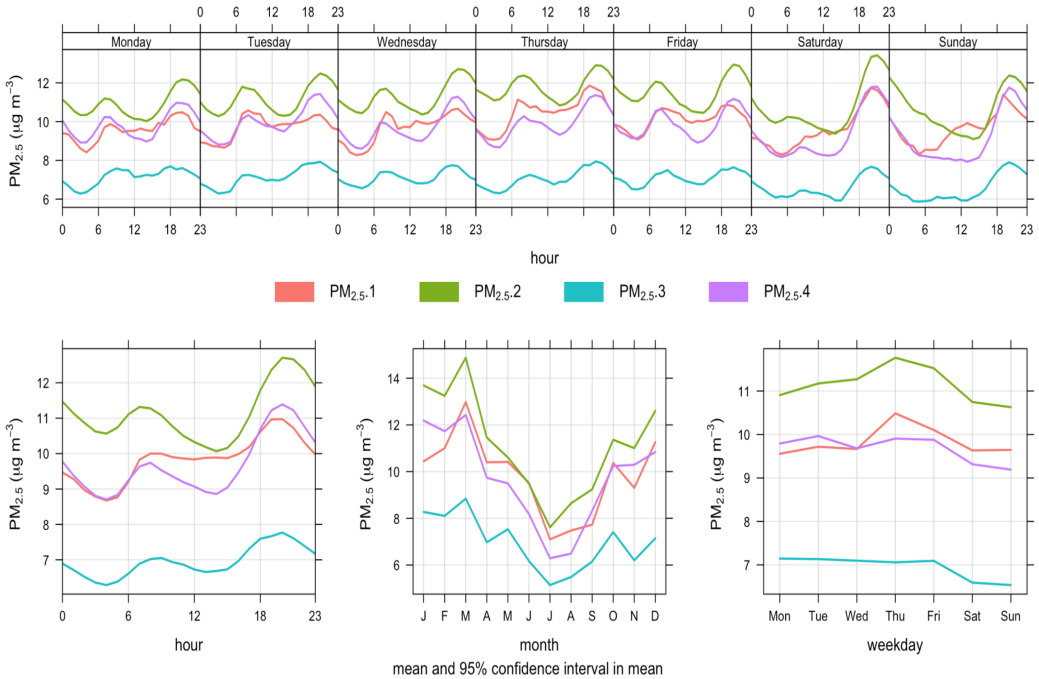


Figure 5.3: Time variations of the four cluster's $PM_{2.5}$ centroids.

Figure 5.4 presents the monthly average concentration for each cluster during three years (2015-2017). This figure shows clearly how the monthly concentrations of

$PM_{2.5}$ follow the same pattern with different magnitude.

Figure 5.5 compares the centroid time patterns side by side based on the daily average concentrations. In this figure, we observe that peaks and troughs of concentration in the TS seem to be related in all clusters but the effect is much higher in the South (green) cluster. Hence concentrations of $PM_{2.5}$ tend to vary similarly in time across the whole country although the magnitude of the variation differs from area to area.

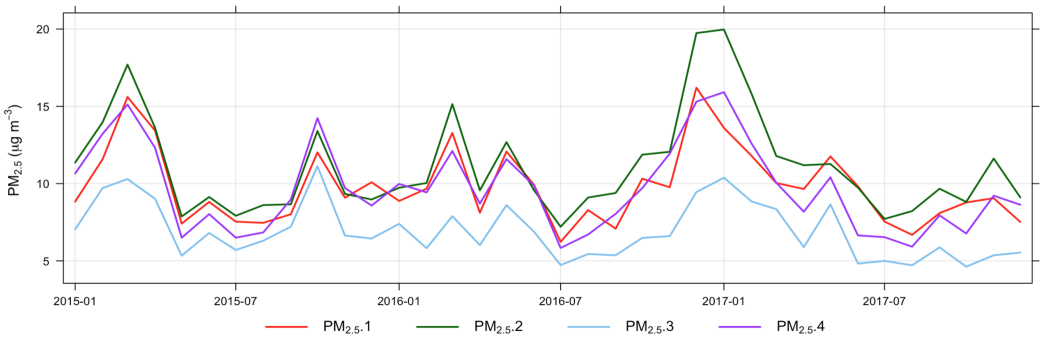


Figure 5.4: Monthly average concentrations of the four cluster's $PM_{2.5}$ centroids.

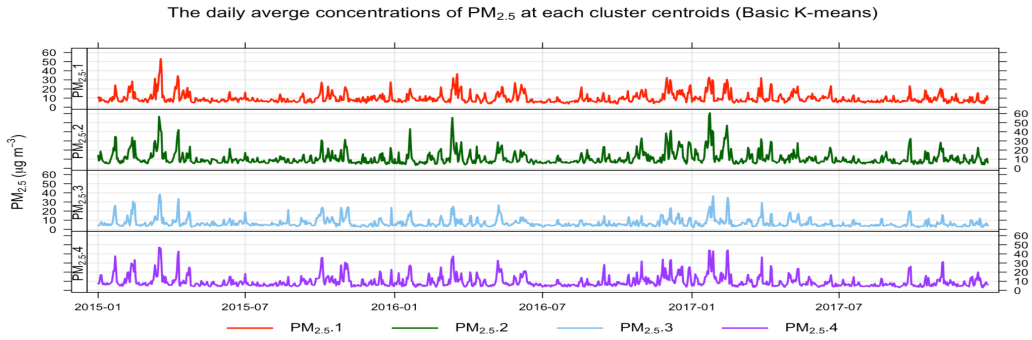


Figure 5.5: Daily average concentrations of the four cluster's $PM_{2.5}$ centroids.

In general, and observable in all the graphs, there is a graduation of the concentrations of $PM_{2.5}$ at the cluster centroids from the South (highest, green) to the North (lowest, blue). The concentrations on clusters in the Centre (purple) and South West (red) are very similar to one another. According to Figure 5.3, concentrations of $PM_{2.5}$ are slightly lower in the weekend for all clusters and appear highest at peak hours, particularly during the evening. As we can see there are low concentrations during the summer (June, July, and August) compared to the

rest of the year. Historically, peak values can be seen in Figure 5.4 in January and particularly during 2017, with the South being markedly higher during those peaks than the North. These geographical variations are consistent with understanding of the sources of $PM_{2.5}$ which tend to be greatest in the south of the UK and the influence of sources in continental Europe [68].

In general, there is a seasonal variation with $PM_{2.5}$ concentrations during these years (as shown in Figure5.4), the concentrations tend to be higher in the winter and lower in the summer. Trend cannot be seen with the monthly concentrations in this plot, it could be noticed with the yearly mean concentrations.

5.2.2 PM_{10} Clustering Results

The total number of stations that measure PM_{10} is 75 stations. The result of clustering this set of stations is shown in Figure 5.6. There are three clusters (i.e. (cluster 1, (red), (cluster 2, green), and (cluster 3, blue)), two large clusters located in the North and the South and a small cluster that contains six stations on the South West. Figure 5.7, shows the time variation analysis of clusters centroids. The centroids of cluster 2 (green), which is located in the North, and cluster 3 (blue), which is located in the South, exhibit very similar behavior. However, cluster 2 in the North has lower concentrations of PM_{10}. In terms of time variation, for the two main clusters, there is still some effect of day of the week with lower values at the weekend (as shown in bottom right plot), and higher peak hourly values although the variations are much less than for $PM_{2.5}$. The summer months (June, July, August) also register the lowest concentrations (as shown in bottom middle plot). The centroid of cluster 1 (red) has average concentrations compared to the other two.

Figure 5.8 shows the monthly average concentrations of each cluster's centroid. We see similar patterns and trends in cluster 2 and 3 (green and blue), while cluster 1 has a higher peak during 2016. At other times, the South (blue) cluster is generally

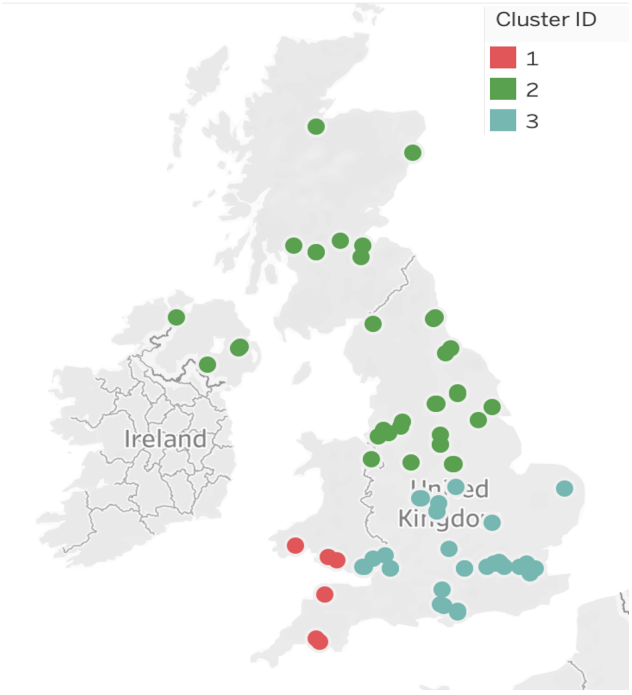


Figure 5.6: Geographical distribution of stations that measure PM_{10} with the colour coded clusters obtained using the basic k-means algorithm.

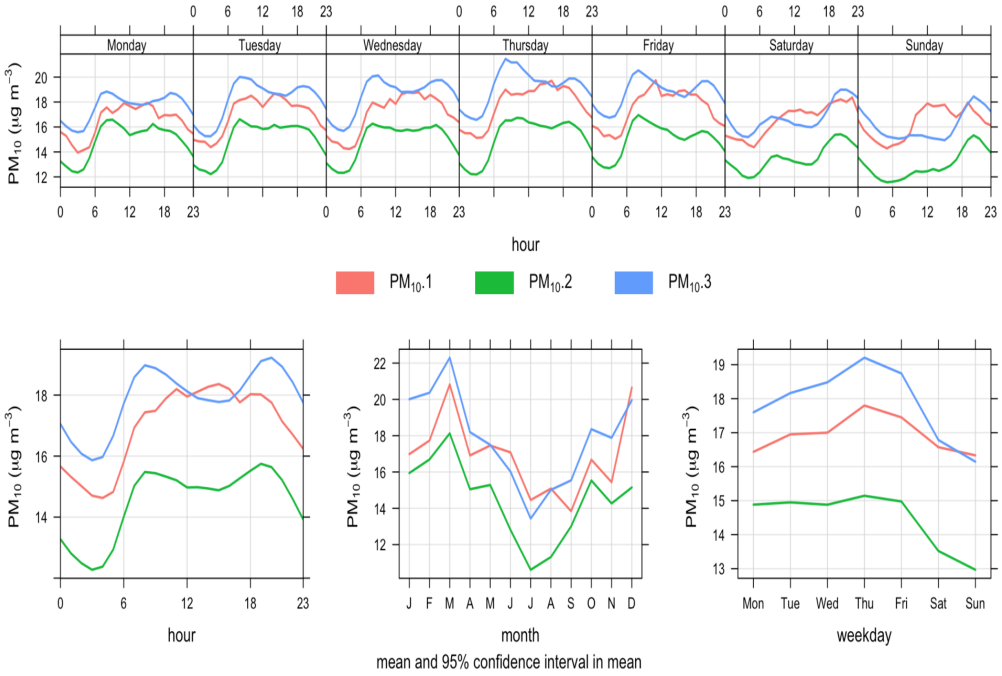


Figure 5.7: Time variations of the three cluster's PM_{10} centroids.

higher. The seasonal variation for PM_{10} is very similar to $PM_{2.5}$. Similarly, peaks and troughs of the daily average concentration in the TS of cluster 2 and 3 (green

and blue) seem to be related more than cluster 1 (red), as shown in Figure 5.9.

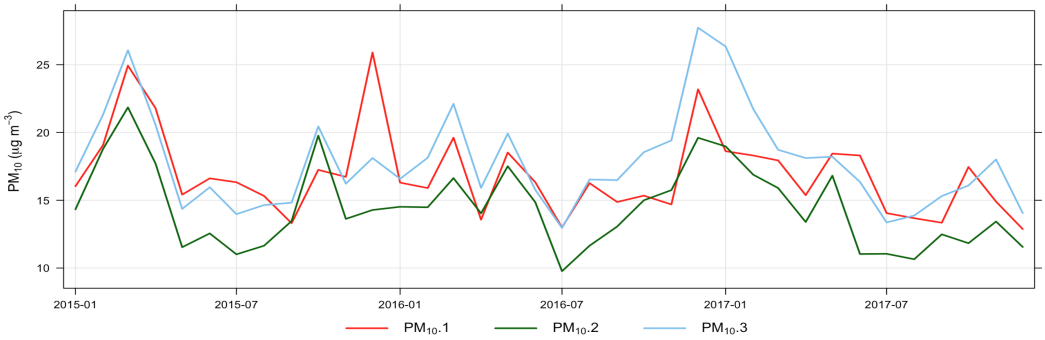


Figure 5.8: Monthly average concentrations of the three cluster's PM_{10} centroids.

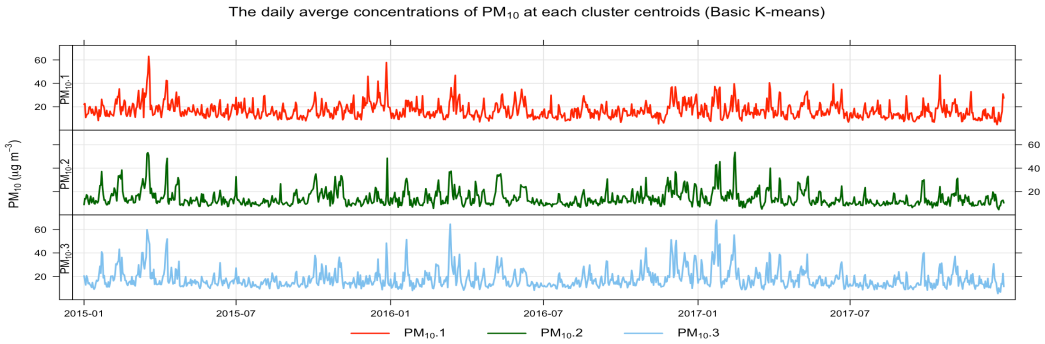


Figure 5.9: Daily average concentrations of the three cluster's PM_{10} centroids.

5.2.3 O_3 Clustering Results

The total number of stations that measure O_3 is 70 stations. The clusters obtained for these stations are shown in Figure 5.10. In this map, there are three clusters (i.e. (cluster 1, (red), (cluster 2, green), and (cluster 3, blue)) located roughly across North/West, Center, South/East though for O_3 the geographical separation of the clusters is less clear. The clusters are separated geographically except for some stations that are blue (cluster 3, mostly covering the North) which are mixed within the green points (cluster 2, mostly covering the Center).

Figure 5.11 shows the time variations of the cluster centroids. In general, based on the clustering results there are two levels of O_3 concentrations, high in the North/West and lower in the Center and the South/East (as shown in the top plot).

We can see that the centroid of cluster 3 (blue) that is located in the North/West has the highest concentrations among all other clusters centroids. However, cluster 1 and 2 are almost identical and have low concentrations compared to cluster 3. Concentrations appear higher at the end of the week (Friday-Sunday) and during the early afternoon (as shown in the bottom right plot).

We can also see there are higher concentrations during March, April, and May compared to the rest of the year, hence O_3 shows dissimilar behaviour to the particulate matter (as shown the bottom middle plot). Figure 5.12 shows the monthly average concentrations for each cluster centroid. According to this figure there is some seasonality with peaks occurring in late Spring. Figure 5.13 shows the 3 clusters display similar trends, though the concentrations are higher for the blue (North/West) cluster.

These spatial and temporal distributions are consistent with the UK being a net sink of surface O_3 due to emissions of NO_x and dry deposition to the surface [69]. Tropospheric background O_3 peaks in the spring due to photochemical production and exchange with stratosphere. This is mainly imported into the UK in the prevailing westerly air flow. Greater NO emissions in the south east and during week days and rush hours reduce the surface concentrations of O_3.

5.2.4 NO_2 Clustering Results

The total number of stations that measure NO_2 is 157 stations, so this is the most measured pollutant. The map in Figure 5.14, shows the clustering of these stations. There are three clusters located roughly around the North (cluster 2, Green), Center (cluster 4, Purple) and South (cluster 3, blue) and the fourth cluster (cluster 1, red) that is spread all over the other clusters hence NO_2 does not show the same neat geographical division as other pollutants. The red cluster has the highest concentrations among all other centroids as shown in Figure 5.15. This cluster includes 95% traffic urban stations, that are located near to traffic (roads, motorways, high-

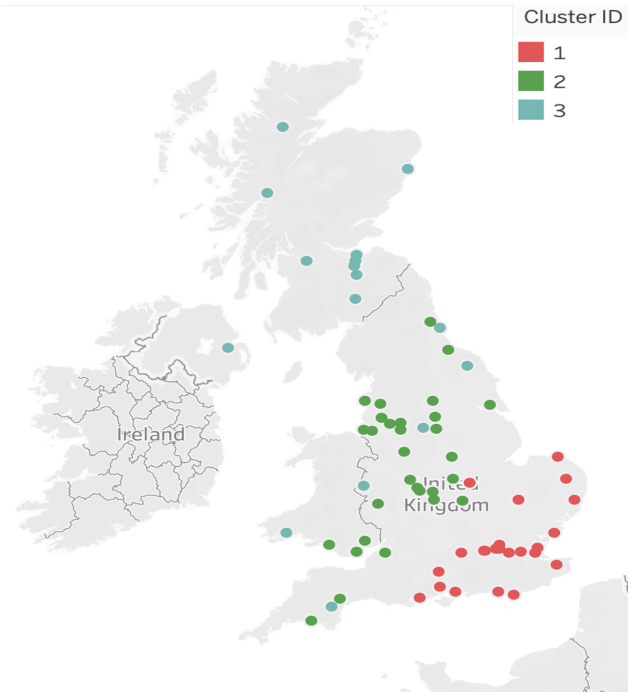


Figure 5.10: Geographical distribution of clustering the stations that measure O_3 using the Basic k-means algorithm.

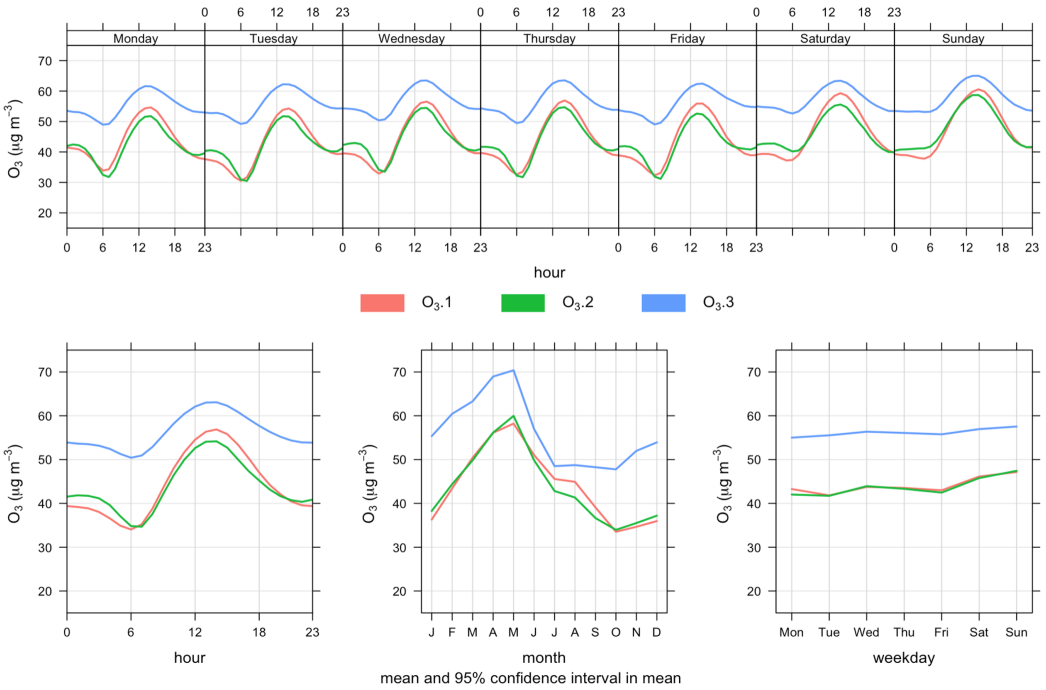


Figure 5.11: Time variations of the three cluster's O_3 centroids.

ways), and the pollution level at these stations is determined predominantly by the emissions from nearby traffic, and 5% background urban stations that are loc-

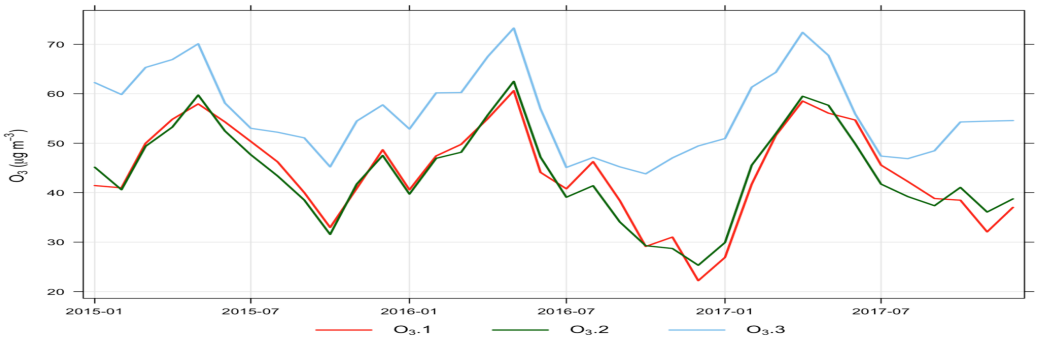


Figure 5.12: Monthly average concentrations of the three cluster's O_3 centroids.

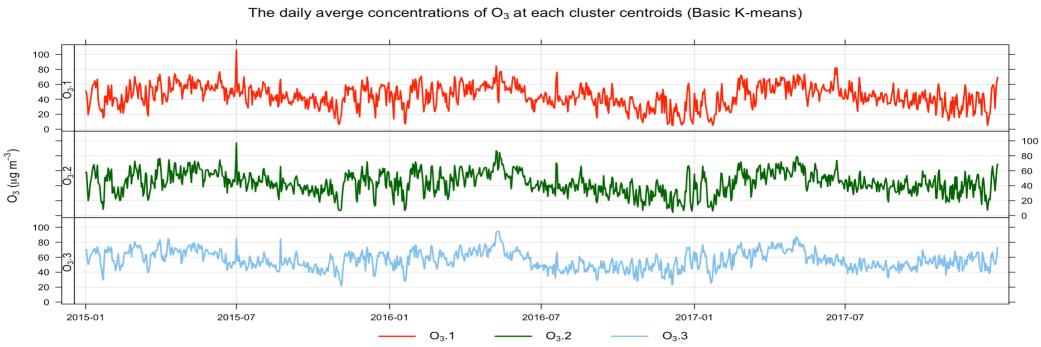


Figure 5.13: Daily average concentrations of the three cluster's O_3 centroids.

ated in the big and crowded cites such as Greater London, Nottingham, etc. Since NO_2 is the main traffic related air pollutant and it has a lifetime of just minutes to hours, it is not unsurprising that these sites form a cluster and also that the centroid concentrations are higher than those of the other clusters.

The centroids of the other 3 clusters are however very similar. Cluster 3 (blue) located in the South has slightly higher NO_2 concentrations followed by cluster 4 (purple) at the Center, then the cluster at the North (green) has the lowest concentrations. There are lower concentrations for all clusters on weekend days and there are peaks during the rush-hours (around 7 am and 6 pm) as shown on the left bottom plot. There are also lower concentrations for all clusters during the summer months (June, July, August) as shown on the middle bottom plot. Figure 5.16 and Figure 5.17 shows the monthly and the daily average concentrations of each cluster centroid again exemplifying how the red cluster is very different to the others although it shows similar seasonality.

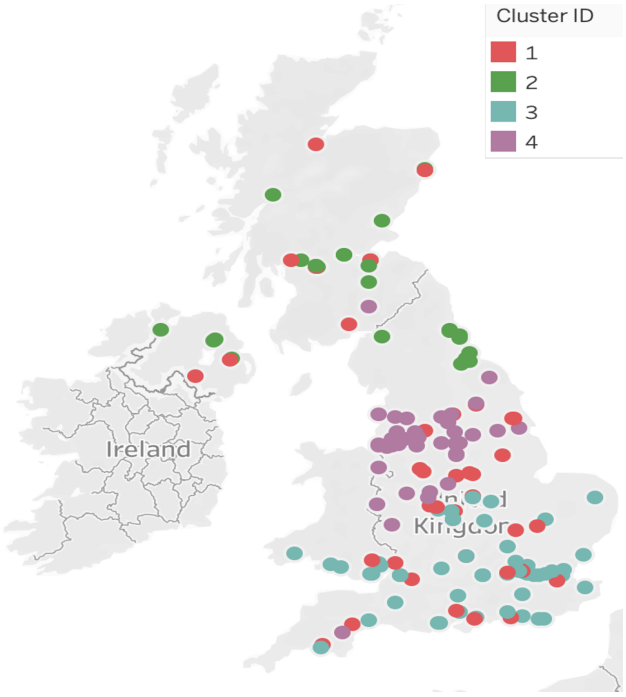


Figure 5.14: Geographical distribution of clustering the stations that measure NO_2 using the Basic k-means algorithm.

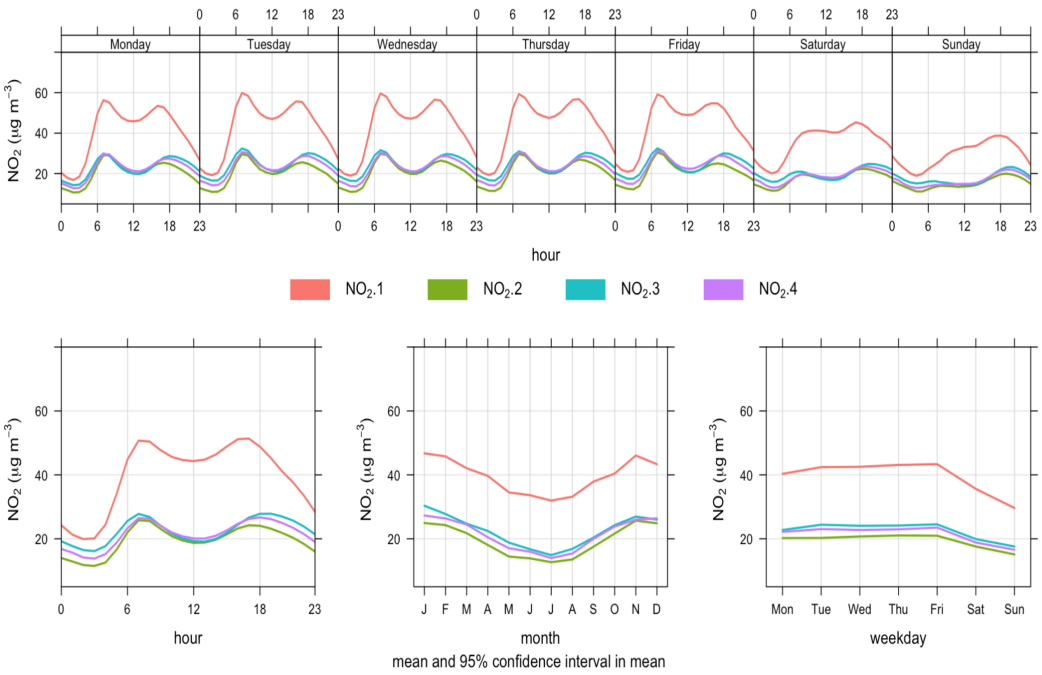


Figure 5.15: Time variations of the four cluster's NO_2 centroids.

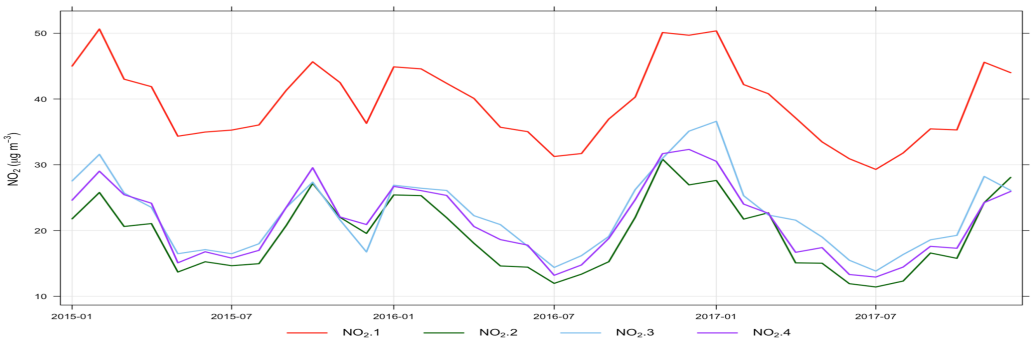


Figure 5.16: Monthly average concentrations of the four cluster's NO_2 centroids.

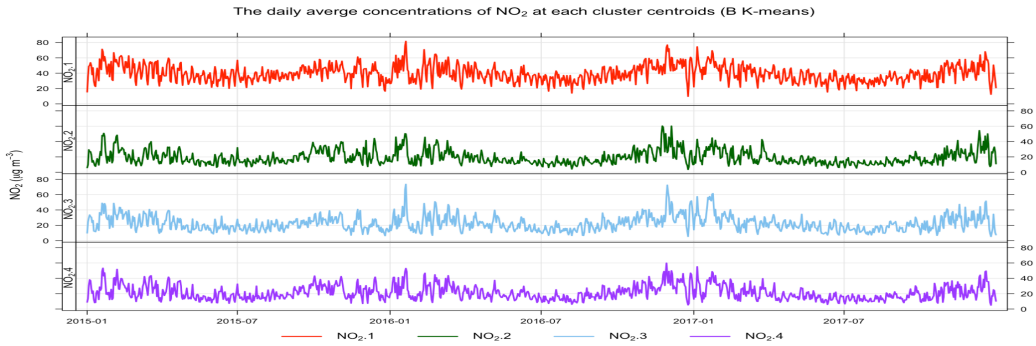


Figure 5.17: Daily average concentrations of the four cluster's NO_2 centroids.

5.3 Evaluation

In this section, we evaluate the clustering solutions produced for each pollutant based on CVIs, these measures are explained in details in chapter 2. Then we evaluate how good these clusters are for the purposes of imputing the pollutants with our proposed imputation models.

Table 5.1 shows the comparison of these clusters using three CVIs indices. In this table, we use the Dunn Index (DI) to measure cluster compactness and separation, Average Silhouette Width (ASW) to measures how close each point in one cluster is to points in the neighboring clusters, and the connectivity measure (Conn) to reflect how connected objects are within the clusters. We will use these metrics to compare the univariate TS clustering solutions to others generated from the second approach using the multivariate TS clustering in the next chapter This will help us to select the clustering approach that gives well separated and compact clusters

to use with pollutant imputation.

Table 5.1: Comparing the k-means clusters for each pollutant using the Cluster Validity Indices (CVI) in experiment 1.

Measure	**Criteria**	**$PM_{2.5}$**	**PM_{10}**	**O_3**	**NO_2**
Optimal number of cluster (k)		4	3	3	4
Average Silhouette Width (ASW)	Maximised	**0.235**	0.183	0.227	0.129
Dunn index (DI)	Maximised	0.911	**0.990**	0.761	0.871
Connectivity ($Conn$)	Minimised	36.033	**27.812**	28.962	102.088

From this table, PM_{10} produces the best clustering solution based on DI and the best connectivity compared to the other pollutants, so the stations clustered together are similar to one another yet dissimilar from stations in other clusters. However, the clustering solution obtained for $PM_{2.5}$ has the maximum ASW because the number of the cluster is higher. It has similar DI value for PM_{10}.

On the other hand, to evaluate how good these clusters are for imputing the pollutants, we compare the RMSE of our imputation methods using the clustering solutions for individual pollutants. In Table 5.2, the method that gives the lowest RMSE is (CA+ENV) for NO_2 and O_3, and (CA) for $PM_{2.5}$ and PM_{10}. This indicates that NO_2 and O_3 concentrations change from a location to another based on the environmental type, for example the stations that are located at the roadside have higher concentrations of NO_2 than those at the rural background. However, $PM_{2.5}$ and PM_{10} have more regional patterns, and wider distribution. If we compared these values, we can see that the lowest error average is associated with the imputation of the PM_{10} and $PM_{2.5}$, this supports our previous evaluation of the clustering quality. Detailed results for each pollutant imputation based on this approach are in Appendix D.

Table 5.2: The average RMSE and its standard deviation (Std) between observed and imputed TS using the basic k-means clustering algorithm with SBD (univariate TS clustering) from experiments 1.

Imputation Method	**NO_2**		**O_3**		**$PM_{2.5}$**		**PM_{10}**	
	RMSE	**Std**	**RMSE**	**Std**	**RMSE**	**Std**	**RMSE**	**Std**
CA	15.037	7.260	14.877	4.169	**5.728**	1.524	**8.312**	2.750
CA+ENV	**14.095**	7.051	**14.500**	3.885	6.147	1.779	8.367	2.739

5.4 Summary

In this chapter, we have applied the first approach which is using univariate TS clustering to cluster and impute missing pollutants. In the experiment, we applied the k-means clustering algorithm based of the SBD to cluster stations for an individual pollutant, then we imputed each pollutant based on its own clustering solution. We evaluated how good these clusters are for the purposes of imputing the pollutants with different imputation models.

From the analysis, we found that the k-means clustering algorithm with the SBD measures is able to generate good clustering solutions that are compact and well separated and geographically correlated to the pollutant behaviours. Our analysis of the clusters' centroid, show how plausible the clusters are in terms of the pollutant concentrations around the UK.

We found a clear graduation of the concentrations of $PM_{2.5}$ across the UK from the South to the North. Similarly, for PM_{10} clusters showed clear geographical groups that have different concentrations represented by the cluster centroids. PM_{10} clusters, compared to others pollutants, achieved better quality in term of compactness and connectivity based on the CVIs measures.

On the other hand, O_3 clusters have less clear geographical separation than PM_{10} and $PM_{2.5}$ but better than NO_2. Clusters with NO_2 were the most difficult to separate, they do not show the same neat geographical division as other pollutants. The main reason for that, is the strong spatial pattern of NO_2 concentrations, as this pollutant is concentrated near to the sources (roads) more than in other areas. At the end of this chapter, we evaluated these clustering solutions based on the clustering validity indices (CVIs) to see how good the clusters are in term of compactness and separation.

In the next chapter, we will apply the second approach which is based on our proposed multivariate TS clustering method and compare the results derived from

both univariate and multi-variate clustering in terms of which clustering solutions produce better imputation. This will enable us to evaluate the quality of our proposed MVTS clustering method in the context of this particular application.

Chapter 6

Results of Applying Multivariate Time Series (MVTS) Clustering and Imputation

This chapter focus on the second approach, which is based on the multivariate TS clustering including all pollutants (O_3, PM_{10}, $PM_{2.5}$, and NO_2). This approach is introduced in chapter 3.

In this chapter, Section 6.1 starts with the general framework for the experiments. Section 6.2 presents the results of applying the proposed multivariate TS clustering (MVTS) based on the k-means clustering algorithm with the fused distance (FDM). Section 6.3 presents the clustering results evaluation based on the clustering validity indices (CVIs) first, then evaluates how good the clustering solutions are for pollutant imputation. This section also includes analysis of the best clustering solution based on the time variations between cluster centroids and a comparison between the imputed and the real TS using the RMSE and its standard deviation (Std). Section 6.4 compares and discusses the derived clustering and imputation results from the MVTS clustering approach against the univariate TS clustering approach. Finally, Section 6.5 is a summary of this chapter.

6.1 Air Pollution Imputation Based on Multivariate Clustering Framework

This framework includes the main stages of our proposed solution as explained in Section 3.1. The main stage under this approach is stage 2, that includes time series clustering and evaluation process. In this stage we calculate the fused distance between MVTS, which represents the similarity between the stations in our dataset based on the measured pollutants, then applying the proposed clustering algorithms (i.e. included in Section 3.3.2) that are based on the k-means clustering algorithms and select the best clustering solutions.

Using intermediary fusion in this approach as explained previously in Section 3.3.2 enables us to measure the distances between stations based on four pollutants. That is, the similarity of each single pollutant (TS) is computed based on SBD distances, then the overall similarity of the MVTS is obtained by fusing the single similarity matrices into one matrix that represents the similarity distance of all the pollutants, as calculated in Section 3.3.2.1. This is to investigate if clustering taking into account all of the pollutants measured at a particular station and their similarity may give a better understanding of the patterns of concentration and better imputation results. In our experiments, we use different versions of the k-means algorithm to cluster stations based on the calculated fused distance matrix. Then we evaluate the clustering solutions using the clustering quality measures (CVIs), to select the best clustering solutions that is compact and well separated to use in the imputation stage. Figure 6.1 shows the main stages of our experiments.

We conduct two experiments under this approach: In the first experiment, we use MVTS clustering based on the fused similarity with the k-means clustering algorithm including all stations in the clustering process, however in the second experiment, we exclude stations that measure only one pollutant, which is always NO_2. We find that these stations are difficult to cluster using the fused similarity,

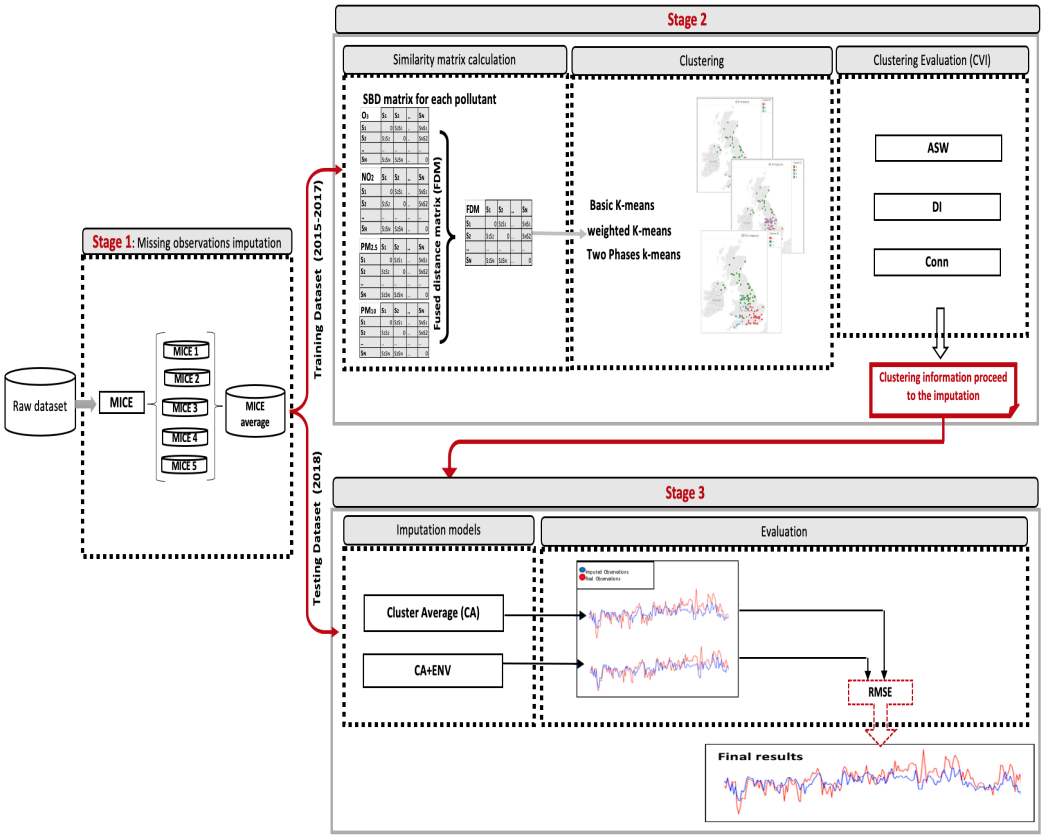


Figure 6.1: Experiment stages of MVTS clustering and imputation (Approach 2).

because the fused distance of these stations is the distance of the only measured pollutant. As an alternative solution, we allocate these stations to the clusters based on their similarity to the cluster centroids.

In the first experiment, we run the k-means algorithm using different criteria to calculate the cluster centroids. The optimal number of cluster k is selected based on Silhouette Width (ASW) function, as explained in Section 2.2.3.2. The first algorithm we run is the basic k-means, as described in Section 3.3.3. This algorithm clusters the stations based on the fused distance matrix without considering any uncertainty. In this case, we give all the objects/stations the same weight to contribute to the centroids.

The second algorithm is the weighted k-means algorithm. This algorithm uses the average uncertainty of the objects/stations as their weights to the cluster centroid

calculation as described in Section 3.3.4. Since the total average uncertainty from all the stations in the dataset is equal to 0.340, we use this as a threshold to identify certain and uncertain stations. In this clustering algorithm, the certain objects, which are the objects with average uncertainty less than (0.340), will have higher weights, than the uncertain objects. To transfer the uncertainty to weight, we re-scaled the average of the uncertainty, then reverse the values. This reverses the weight by subtracting 1 (which is the max value after re-scaling the average uncertainty) from each value.

The third algorithm is the two-phases k-means algorithm, this algorithm will run the k-means algorithm using two phases. In the first phase, we cluster the certain objects only and give all the objects the same weight, so it works as a basic k-means. In the second phase, we assign the uncertain objects to the cluster centroids from the first phase, but we used the uncertainty average of these objects as their weights to the cluster centroid, this clustering algorithm is explained in Section 3.3.5.

In the second experiment, we exclude 43 stations from the clustering process as these stations measure NO_2 only and including them in the FDM may negatively affect the clustering process. What we do this time is to allocate these stations to clusters after clusters are constructed. The allocation is based on their partial similarity of NO_2 to the cluster centroids.

We evaluate all the clustering solutions derived from these experiments based on CVIs and then select the best clustering solution for pollutant imputation (stage 3).

6.2 Experimental Results

6.2.1 Experiment 1:

In this experiment, we use three versions of the k-means algorithm to cluster the fused distance matrix we calculated in Section 3.3.2.1. We compare the clustering

results based on their geography first, then based on the clustering validity indices to evaluate the compactness and separation of the clusters.

6.2.1.1 MVTS clusters using the basic k-means clustering

Figure 6.2 shows the geographical distribution of clustering the stations based on the basic k-means. From this figure, there are three clusters located in three geographical areas (cluster 2/North, cluster 1/South East, and cluster 3/South West). These clusters are well separated, however there are some stations from cluster 1 (red), that appear within other cluster's geographical areas.

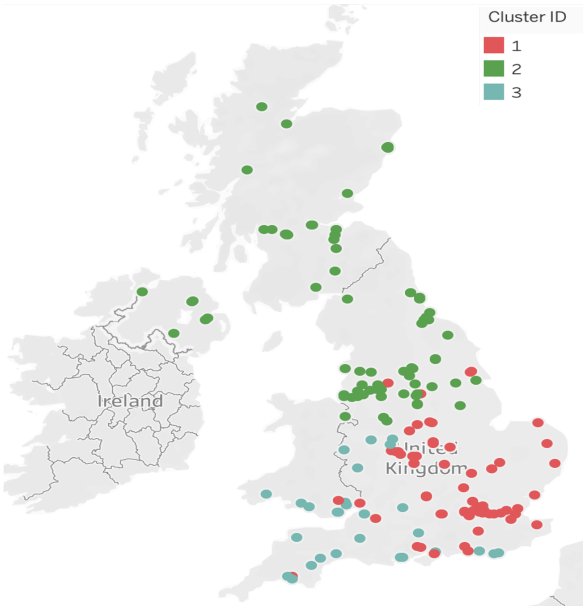


Figure 6.2: Geographical distribution of clustering stations using the basic k-means algorithm experiment 1.

6.2.1.2 MVTS clusters using the weighted k-means clustering

Figure 6.3, shows the result of clustering the stations using the weighted K-means algorithm. As shown on the map, there are four clusters located on four geographical areas (North, Center, South East, and South West). We can observe that the clustering follows good geographical boundaries although again a few stations cross those boundaries.

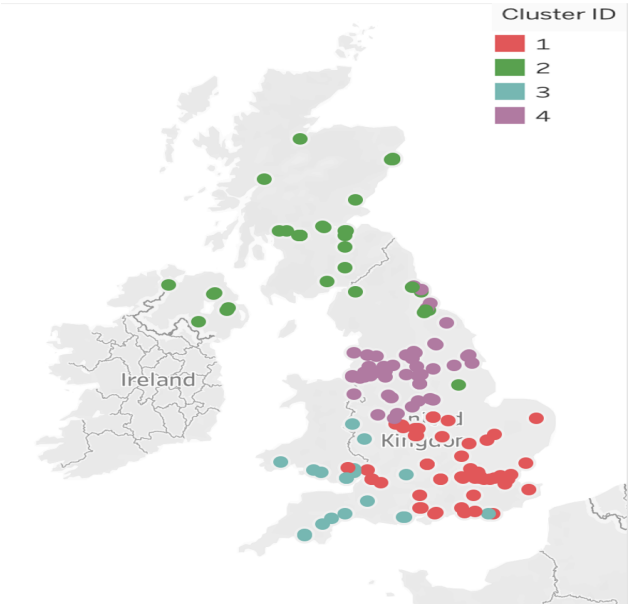


Figure 6.3: Geographical distribution of clustering stations using the weighted k-means algorithm experiment 1.

6.2.1.3 MVTS clusters using two-phases k-means clustering

Figure 6.4 shows the result of clustering the stations using the two phases k-means algorithm. As shown on the map, there are three clusters, with distribution similar to the first clustering solution in Figure 6.2 except with some stations on the West (between Liverpool and Preston) that changed from cluster 2 to cluster 3.

6.2.2 Experiment 2:

Figure 6.5 shows the geographical distribution of stations, noting that this time we have four clusters. As we can see there is good geographical distribution with the four clusters located in the North, central, South East, and South West. These clusters are well separated; in fact better than those in experiment 1, because stations that only measure NO_2 disrupt the geographical connectivity of the clusters. This is because the sites that only measure NO_2 are usually sited in areas where there are concerns about compliance with NO_2 air quality standards, which is normally close to sources such as roads.

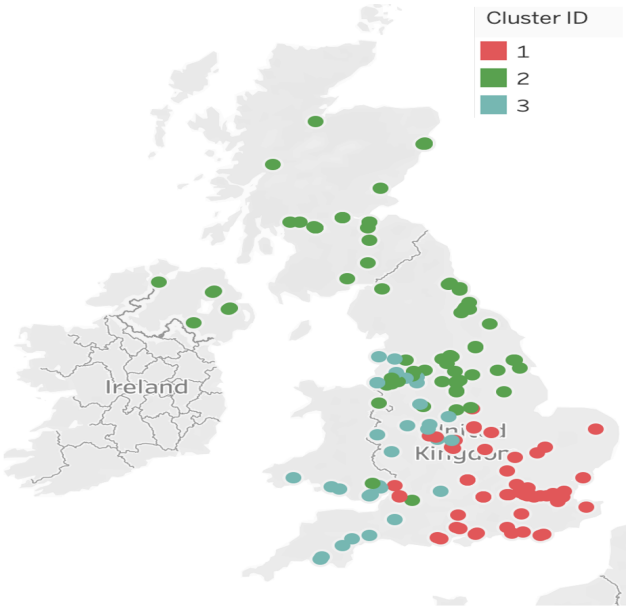


Figure 6.4: Geographical distribution of clustering stations using the two-phases k-means algorithm experiment 1.

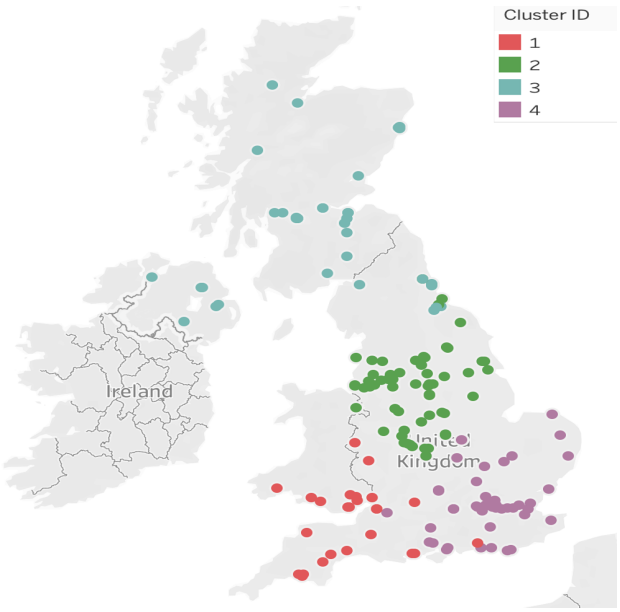


Figure 6.5: Geographical distribution of clustering stations using the basic k-means algorithm experiment 2.

6.3 Clustering Evaluation and Analysis

We evaluate the quality of these clustering solutions based on the internal CVIs to select the best clustering solution, i.e. one that is more compact and well separated.

Table 6.1 shows the comparison of these indices. Then, we impute the missing pollutants using our imputation methods and compare the imputed with the real TS using RMSE and its standard deviation (Std) as shown in Table 6.2.

Table 6.1 shows a comparison between all clustering solutions obtained from the first and the second experiment. In the first experiment, based on these results, we can say that neither k-means version provided particularly improved results so we compare only the results of the basic k-means algorithm with the fused distances with results from the second experiment.

We focus on comparing the basic K-means clustering solution in the second experiment (column Basic k-means (Exp. 2)) followed by the basic k-means clustering solution in the first experiment (Basic k-means (Exp. 1)). As shown in Table 6.1, the differences between these indices of the basic k-means clustering solutions are very small. That may indicate the stability of these clusters even though the stations are clustered under different criteria. The clustering solution from the second experiment is slightly better than the one from the first experiment with a number of these indices including: WCSS measuring the variability within each cluster, BCSS measuring variability between the centroids of the clusters; and it also achieved the highest value with the ASW measuring how close each point in one cluster is to points in the neighbouring clusters. While the clustering solution from the fist experiment is better with the compactness and connectivity indicated by DI and Conn measures.

Table 6.1: Cluster Validity Indices (CVIs) for clustering solutions from experiment 1 and 2 (highlighted cells represent better results in the comparison between the two experiments).

Measure	**Criteria**	**Basic k-means (Exp. 1)**	**Basic k-means (Exp. 2)**	**Weighted K-means (Exp. 1)**	**Two Phases K-means (Exp. 1)**
Optimal number of cluster (k)		3	4	4	3
Average Silhouette Width (ASW)	Maximised	0.124	**0.135**	0.128	0.104
Dunn index (DI)	Maximised	**0.129**	0.104	0.087	0.083
Connectivity ($Conn$)	Minimised	**69.197**	91.809	103.351	102.796
Within Clusters Sum of Squares ($WCSS$)	Minimised	0.341	**0.325**	0.035	0.035
Between Clusters Sum of Squares ($BCSS$)	Maximised	0.425	**0.426**	0.345	0.372

On the other hand, we evaluate these clustering solutions in terms of missing pollutants imputation so we compare different imputation models (CA and CA+ENV)

using the RMSE, as shown in Table 6.2. We compare the clustering imputation methods for the basic k-means clustering algorithm derived from both experiments. Looking at the first experiment only, at the top of Table 6.2, we find that using CA+ENV gives the lowest RMSE for NO_2 and O_3 with (13.947, 14.717, respectively). However, using CA method gives the lowest RMSE (5.234, 8.247) for $PM_{2.5}$ and PM_{10} respectively. This is consistent with the single pollutant imputation results in Table 5.2 in the previous chapter.

For the second experiment, at the bottom of Table 6.2, the result for the best imputation methods for each pollutant agreed with experiment 1. Importantly, using this clustering solution to impute the pollutants gave the lowest average error for all the pollutants except NO_2. This indicates that it is a good clustering solution and helpful for imputation. Detailed results for each station based on this experiment are in Appendix E.

Table 6.2: The average RMSE and its standard deviation (Std) between observed and imputed TS using the basic k-means clustering algorithm with fused distance (MVTS clustering) from experiments 1 and 2.

Imputation models	**NO_2**		**O_3**		**$PM_{2.5}$**		**PM_{10}**	
	RMSE	**Std**	**RMSE**	**Std**	**RMSE**	**Std**	**RMSE**	**Std**
			experiment 1					
CA	15.805	8.107	15.223	3.575	5.234	1.253	8.247	2.720
CA+ENV	**13.947**	7.299	14.717	3.528	5.427	1.226	8.283	2.650
			experiment 2					
CA	16.024	8.443	14.701	3.968	**4.986**	1.155	**7.943**	2.775
CA+ENV	13.965	7.355	**14.125**	3.846	5.332	1.197	8.284	2.751

6.3.1 Analysis of pollutant concentrations in each cluster

Since the second experiment gives the best clustering solution in terms of the clustering quality, we analyse the time variation for all the pollutants in this solution in order to compared the cluster centroids. Figures. 6.6, 6.7, 6.8, and 6.9 show the time variation for $PM_{2.5}$, PM_{10}, NO_2 and O_3 respectively.

Also, we use the monthly and the daily average concentrations that show these differences clearly in Figure 6.10. From these figures, we can see the differences of

the pollutant concentrations between these centroids.

In terms of analysing time variations, there are 4 cluster centroids to compare for this experiment. The centroid of cluster 1 (red), located in the South West, has high concentrations of PM_{10}, $PM_{2.5}$ and the highest concentrations of O_3 among all other cluster centroids, but has the lowest concentrations of NO_2. The centroid of cluster 2 (green), located in the center, has an average concentration of all the pollutants compared to other centroids. The centroid of cluster 3 (light blue), located in the North has the lowest concentrations of PM_{10}, $PM_{2.5}$, low NO_2 but high concentrations of O_3. The centroid of cluster 4 (purple), located in the South West, has high concentrations of PM_{10}, $PM_{2.5}$ and NO_2, but the lowest concentrations of O_3.

From these figures, if we compare the time variation of pollutants concentration based on the location of the clusters from the MVTS clustering (i.e. experiment 2) with individual pollutants clustering (Section. 5.2), we can see that the pollutants concentration in these locations similar to one another. For example, the UK North region has the lowest concentrations of PM_{10}, $PM_{2.5}$, NO_2 but highest concentrations of O_3, while the opposite is true for South regions. Which confirmed the ability of the MVTS clustering to reflect and understand multi pollutants behaviour.

6.3.2 Pollutants imputation examples

We show an example of our imputed TS compared to the real TS for each pollutant using the selected imputation methods. For this comparison, we select two stations that associated with the highest and the lowest RMSE for each pollutant. We compare the daily mean of the imputed and the real TS for the period of six months (Jan-Jun) of the year of 2018.

Figure 6.11, shows a comparison between the imputed PM_{10} using Cluster Average (CA) method and the real TS at 'London N Kensington' station as that associates

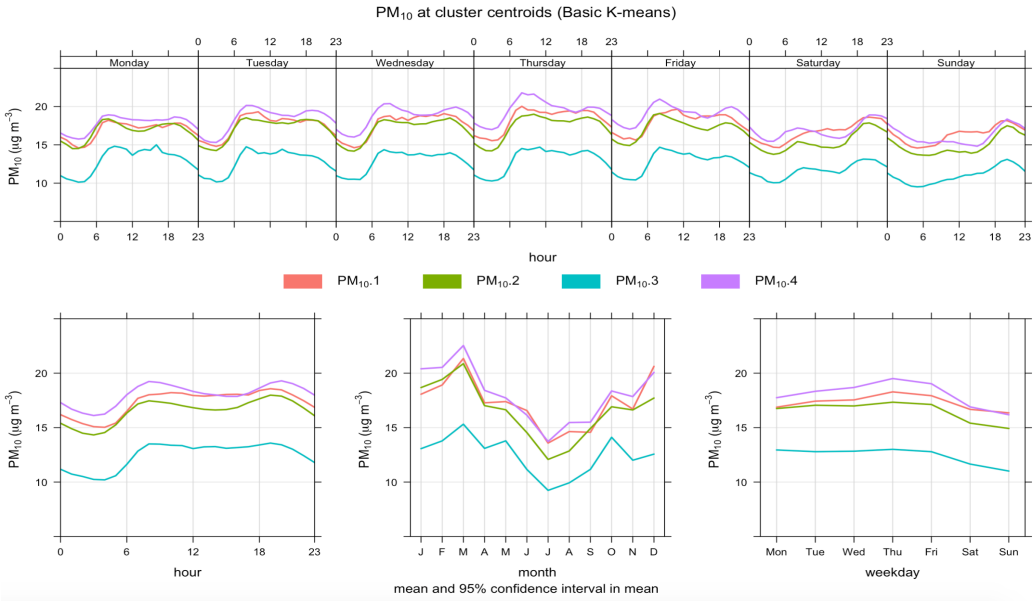


Figure 6.6: Time variation of the basic k-means cluster centroids of PM_{10} concentrations.

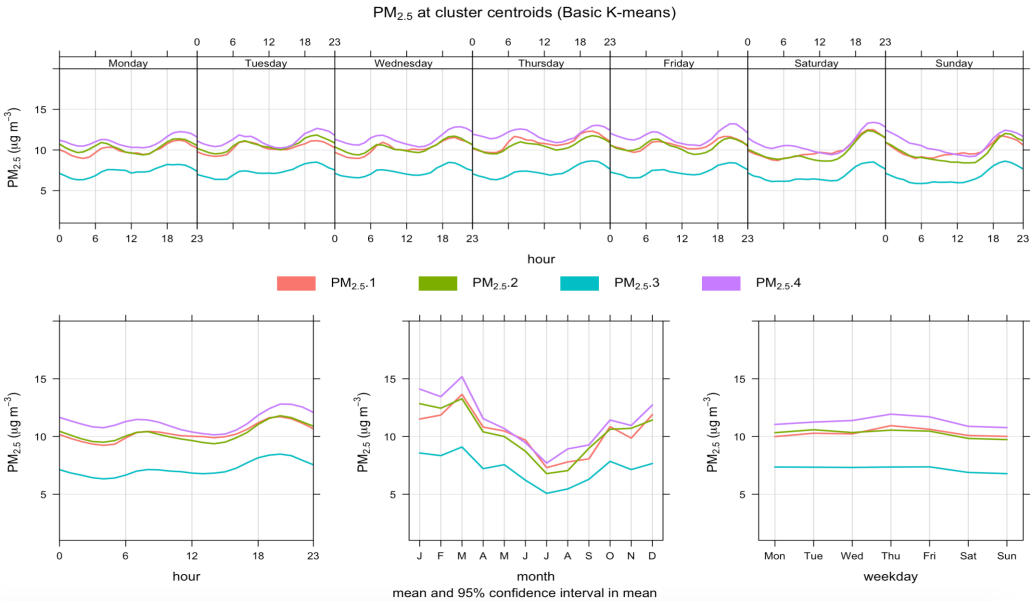


Figure 6.7: Time variation of the basic k-means cluster centroids of $PM_{2.5}$ concentrations.

with the lowest RMSE (which is 5.18) among all the stations that measure PM_{10}, and at 'Cardiff Centre' station which associates with the highest RMSE (which is 13.17). The red TS represents the real TS and the blue is the imputed TS. The imputed TS is very similar and reproduces the same peaks in the real TS, although

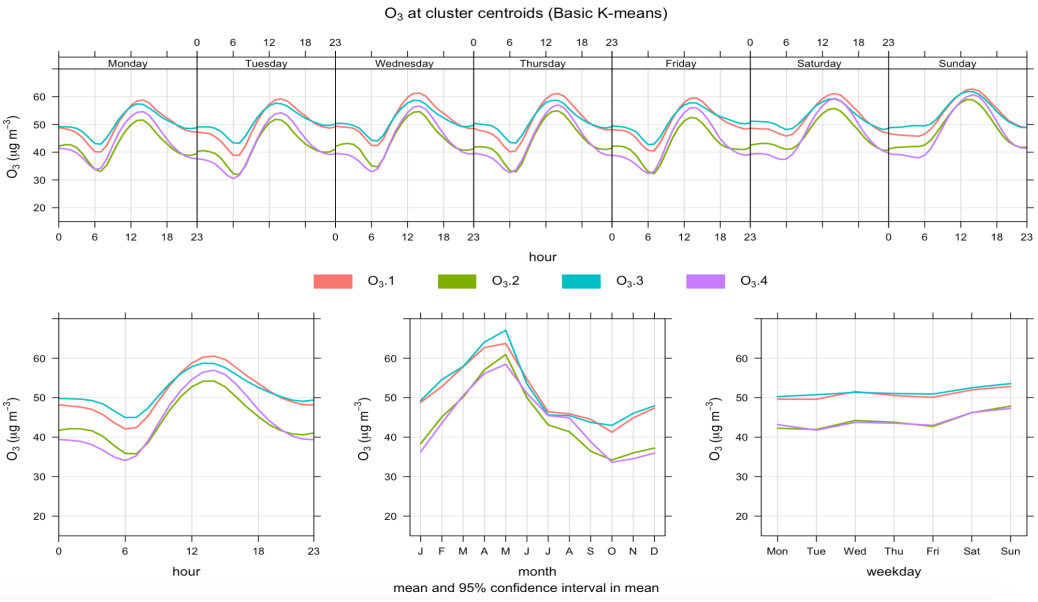


Figure 6.8: Time variation of the basic k-means cluster centroids of O_3 concentrations.

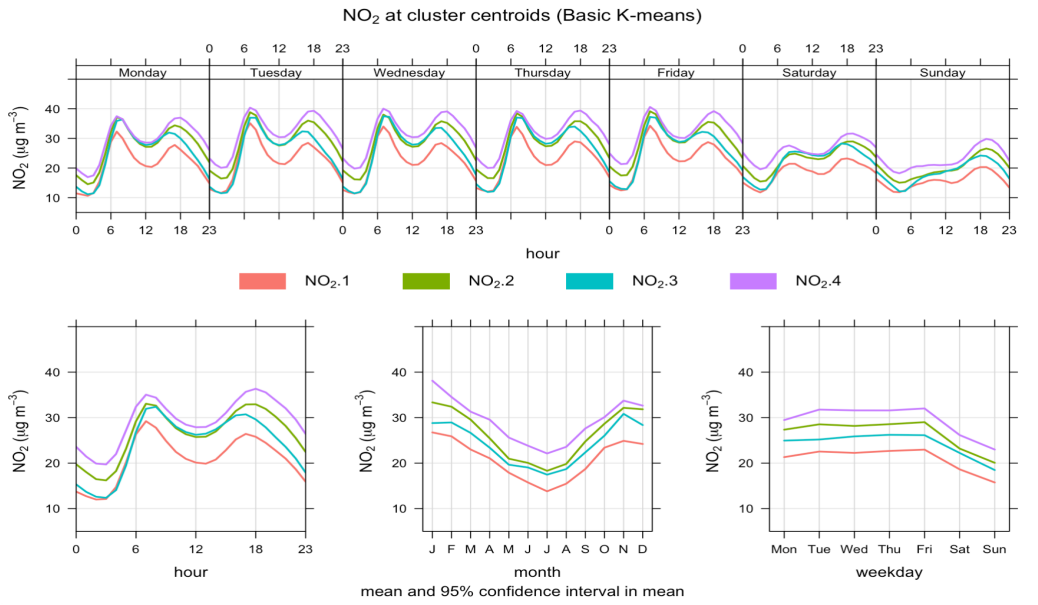


Figure 6.9: Time variation of the basic k-means cluster centroids of NO_2 concentrations.

the imputed TS has slightly higher values. In the second plot, even though, these TSs have the highest RMSE among all the stations that measure PM_{10}, the imputed TS represent a good imputation too and follows the same trend lines.

Figure 6.12 is an example of the $PM_{2.5}$ imputed using the CA method and real

Figure 6.10: Monthly average concentrations of the basic k-means cluster centroids obtained from experiment 2 for PM_{10}, $PM_{2.5}$, O_3, and NO_2 (Top to Bottom).

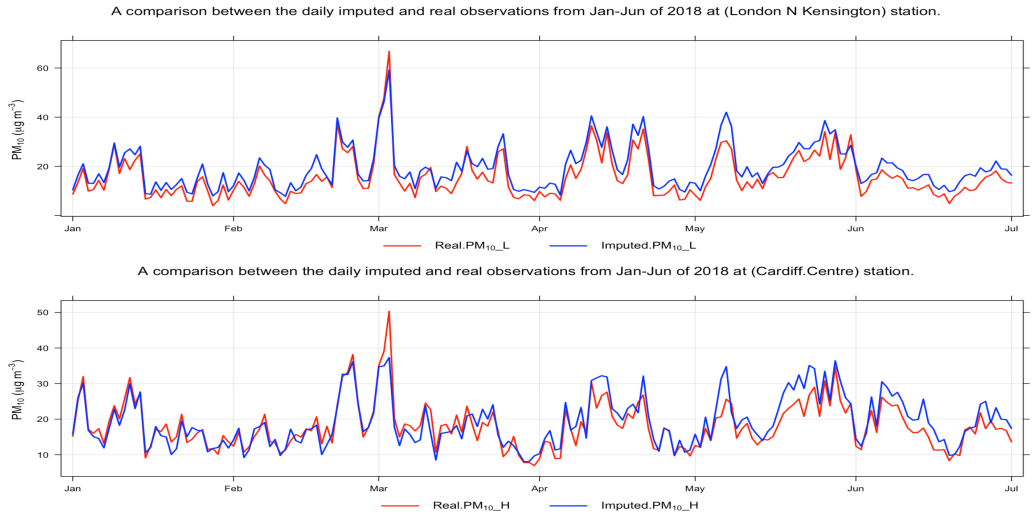


Figure 6.11: Imputed and real TS comparison for PM_{10} with the lowest RMSE (top) and the highest RMSE (bottom) using CA imputation model.

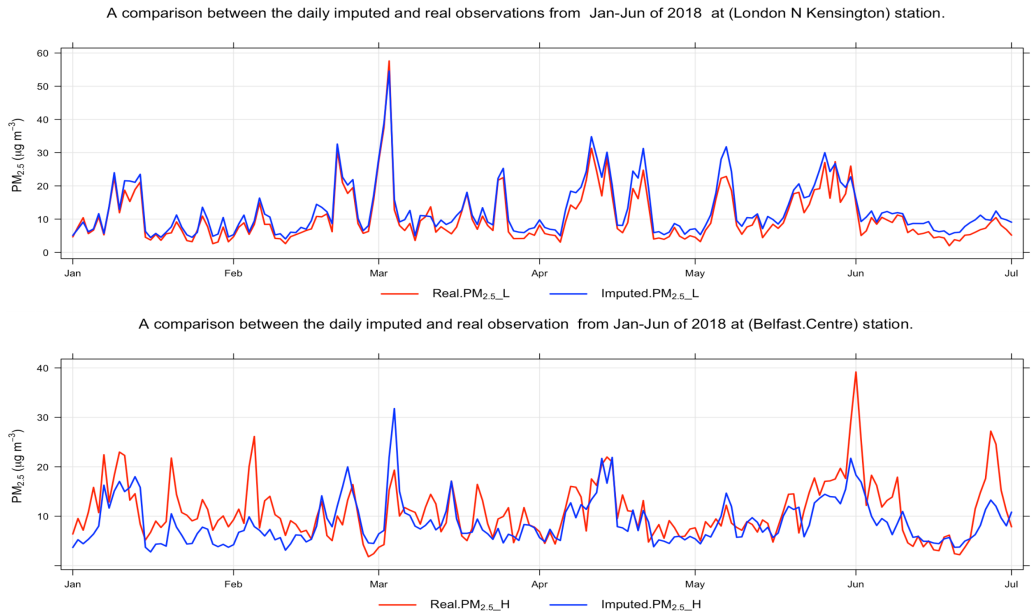


Figure 6.12: Imputed and real TS comparison for $PM_{2.5}$ with the lowest RMSE (top) and the highest RMSE (bottom) using CA imputation model.

TS at 'London N Kensington' (lowest RMSE, 3.17) and 'Belfast Centre' stations (highest RMSE, 6.78). The imputed TS for London N. Kensington has slightly higher values than the real TS, while the opposite is true for Belfast Center, where the imputed TS is slightly lower than the real TS. Again, the trends are very similar and they represent valid imputations.

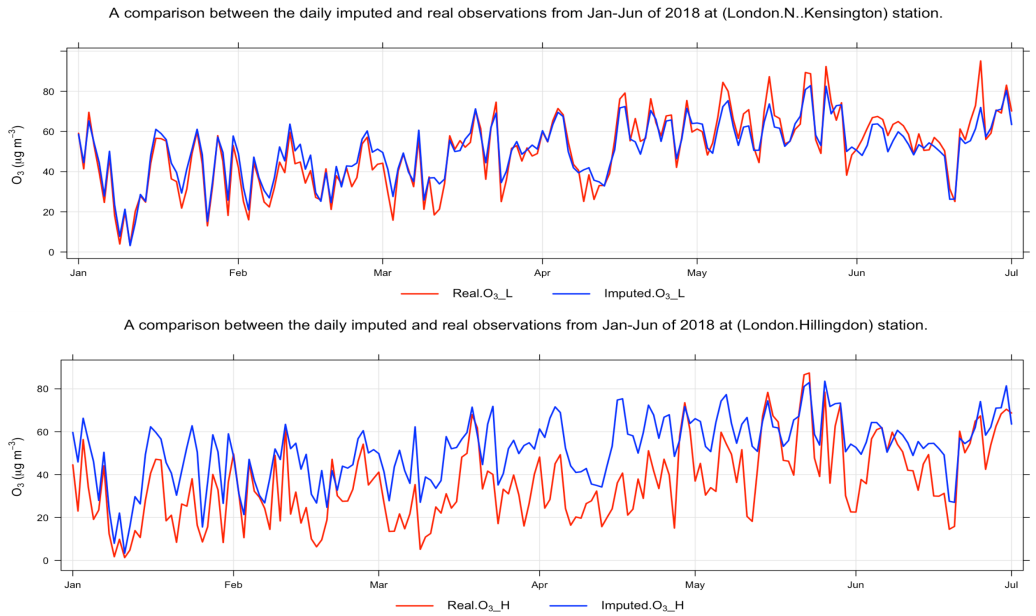


Figure 6.13: Imputed and real TS comparison for O_3 with lowest RMSE (top) and the highest RMSE (bottom) using CA+ENV imputation model.

Figure 6.13 shows a comparison between the imputed O_3 TS using the CA+ENV method and real TS at 'London N Kensington' in the top (lowest RMSE, 9.68), and 'London Hillington' in the bottom (highest RMSE, 21.38). We can see that in the second plot for 'London Hillington' site the variation between the imputed and the real TS is slightly higher compared to the previous example although again the trend is good.

Figure 6.14, is an example of the NO_2 imputed using the CA+ENV method and real TS at 'Narberth' in the top figure (lowest RMSE, 2.85) and 'London Hillington' in the bottom (highest RMSE, 35.50). The first plot in the top, show a good imputation, compared to the second example where very large differences can be observed. That may be affected by the type of the station (London Hillington) or its influence by external factors that cause higher NO_2 concentrations, which make it difficult to impute. This station is also associated with the highest RMSE with O_3 imputation, we can see that the O_3 concentrations in this station is lower than the imputed TS, which suggest that compared to many of the other sites in this cluster there are greater local emissions of NO that are converted to NO_2 via

reaction with O_3.

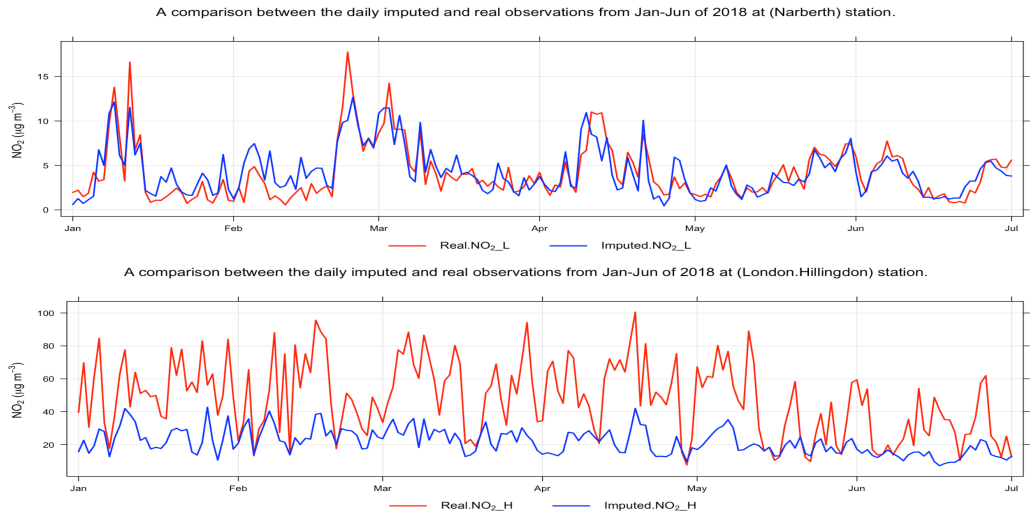


Figure 6.14: Imputed and real TS comparison for NO_2 with lowest RMSE (top) and the highest RMSE (bottom) using CA+ENV imputation model.

6.4 Discussion

To fully evaluate how well our MVTS approach works for our imputation problem, we compare the selected MVTS clustering solution obtained from the second experiment (6.2.2) with the univariate TS clustering solution as explained Chapter 5. This comparison includes the clustering evaluation and imputation analysis.

Our clustering analysis showed that the basic k-means algorithm with fused distances results in geographical patterns that are consistent with our understanding of sources and lifetimes of these pollutants. We found that using the basic k-means with the MVTS clustering and fused similarity in the first and second experiments gave a clear geographical correlation between the stations. Our results of analysing the centroids of the clusters identify similar pollutant concentrations levels and geographical distribution to the results in one of the most recent reports from Centreforcities [4]. This report focuses on analysing the concentrations level across the UK for NO_2 and $PM_{2.5}$, and explores the fact that NO_2 and O_3 have an anti-

correlation [100, 62], hence their concentrations in a region are at opposite ends of the scale. This is corroborated in our clustering results.

In terms of imputation, using the basic k-means with the defined imputation models helped to impute/estimate plausible concentrations of multiple pollutants at a station. Although the best imputation method with lowest error average may be different from one pollutant to another, all experiments agreed that using CA+ENV to impute NO_2 and O_3 gave the lowest error average (RMSE), and using CA is better for the imputation of $PM_{2.5}$, and PM_{10} concentrations due to the behaviour of each pollutant.

We also observe that univariate and MVTS clustering analysis lead to different clustering results. Comparing the error average of these methods from the univariate TS clustering (Chapter 5, Table 5.2) to the MVTS clustering (experiment 1 and 2 Table 6.2) showed that the error average using CA+ENV for NO_2 imputation decreased by (0.15, 0.13) in the first and second experiments using MVTS clustering compared to using the univariate TS clustering. Even though, the error averages increased for pollutant imputation in the first experiment for O_3, they decreased in the second experiment by 0.3, 0.7, 0.4 for O_3, $PM_{2.5}$, and PM_{10} respectively. This indicates that using NO_2 data from NO_2 only sites has a detrimental effect in the imputation of O_3 and PM.

As a conclusion, using the basic k-means with the fused distance performs better than other clustering algorithms for imputation and gives very compact geographical clustering. This indicates that using the fused distance to measure the similarity between the pollutants helped us to solve some of the uncertainty problems associated with missing pollutant values and enabled us to discover multiple patterns of pollutant behavior that are manifested in different areas around the UK. This knowledge can then be used to understand the behaviour of the pollutants that indicate the air pollution level.

Furthermore, MVTS clustering enables imputation even when no measurement is

available for a given pollutant since the station can be allocated to a cluster based on the value of the other pollutants measured as we demonstrated with the stations that measure only NO_2 in the second experiment 6.2.2. This is a real advantage of our MVTS method. Next, we will use clustering results obtained from the second experiment, i.e. the one based on the k-means algorithm with the fused distance, to impute the pollutant concentrations for all pollutants using all proposed imputation models described in Section 3.4.1 and select which model is able to produce the most plausible concentrations.

6.5 Summary

In this chapter, we conducted two experiments to evaluate our proposed MVTS clustering approach. We introduced the full framework to cluster the stations based on their fused similarity with the k-means clustering algorithm, then use the produced clustering solution in the imputation process.

In these experiments, we run three versions of the k-means algorithms, then we evaluated their clustering solutions based on CVIs to select the best clustering solution. We also used all produced clustering solutions to impute the whole TS and compare the clustering solutions in terms of their imputation using the RMSE between imputed and real TS.

As a result, in the first experiment, we found that the basic k-means algorithm with the fused distance achieved better clustering results than other algorithms' versions. Based on some CVIs (i.e DI and Conn), that indicate that this clustering solution has more compactness and connectivity than other versions in the same experiment. However, the second experiment improved the clustering solution based on other CVIs (i.e ASW, WCSS, and BCSS) that indicate better separation and correlation between clusters, adding to that the clustering shows a clear geographical correlation between the stations (Figure 6.5). On the other hand, after imputed the pollutants concentrations based on these clustering solutions,

our results show that using the clustering solution from the second experiment improved the imputation by reducing the error average using the RMSE by 0.6 for O_3 based on CA+ENV and by 0.2, 0.3 for $PM_{2.5}$, and PM_{10} respectively based on CA imputation model.

At the end of this chapter, we further compared the imputation results from the best clustering solution using MVTS clustering with the univariate TS clustering imputation from the previous chapter. Based on this comparison, we found that using the MVTS algorithm helps to understand pollutants' behaviours and their relation, also it helps to give plausible imputation, even when a pollutant is not measured at a particular station.

In the next chapter, we will use experiment 2 clustering results, and apply all the proposed imputation models to evaluate and analyse the imputed concentrations and understand model behaviours for each pollutant.

Chapter 7

Evaluation of the Imputation Models

In the previous Chapter 6, we selected the best clustering technique that gave well compact and separated clusters in terms of the clustering quality and also the one that helped to derived good imputation for missing pollutants (whole TS). In this chapter, we first analyse the proposed pollutant imputation models using some statistical and graphical air pollution modeling evaluation functions. Then, we evaluate the imputation models performance based on the comparison between the observed and imputed DAQI. Section 7.1 reviews the model evaluation techniques we are going to use in the evaluation process; Section 7.2 includes model evaluation results based on some statistical and graphical air pollution modeling evaluation functions; Section 7.3 includes models performance under different weather conditions; and Section 7.4 includes the best model performance based on the comparison between the imputed and observed DAQI. Finally, Section 7.5 is a summary of this chapter.

7.1 Models Evaluation Techniques

In the previous Chapter 6, based on the comparison between univariate and multivariate time series clustering, we found that using the basic k-means with the fused distance gave the best clustering and imputation results. Hence, we will continue using this clustering approach with all imputation models that require clustering information. We apply all imputation models (as described in Section 3.4.1) to generate different imputed TSs for each observed TS, then we evaluate how plausible the imputation is using different models by comparing truth values to imputed values. The model evaluations are based on the test dataset, which is the 2018 data. As earlier mentioned, we do this by taking each existing TS for which we have values, one at a time, and consider them missing. Then, we impute the whole TS by various models and compare that to the ground truth. We are evaluating our models against the real concentrations which contain missing values, hence, we ignore all the missing values in this evaluation. For each model, we average the imputation models' behaviour from all the stations to compare and establish the one that provides imputed values closest to the real values.

We compare the real values to the imputed values using different statistical and graphical model evaluation functions, as previously included in Section 3.4.2 and Section 2.4.3. The model that gives the lowest error on average, the highest correlation and the highest degree of agreement between imputed and observed concentrations for all stations (i.e. imputed TS) will be considered the best model. Note that, the selected model may change from one pollutant to another.

To fully evaluate our models and understand how the models perform under different weather conditions, we evaluate the imputation models performance under different weather types using Lamb Weather Types (LWTs).

After selecting the best model for each pollutant, we impute the measured pollutants in all the stations. For our evaluation, we calculate DAQI from the imputed

data (i.e. the imputed DAQI), as explained in Section 3.4.2. Then, we use the observed DAQI (i.e. calculated based on observations) as a performance tool to evaluate our imputation model on its ability to reproduce the daily air quality index.

7.2 Air Pollution Imputation Modeling Evaluation.

In this section, we first analyse the proposed pollutant imputation models using some statistical and then graphical air pollution modeling evaluation functions.

7.2.1 Model Evaluation Based on Statistical Analysis.

Table 7.1, shows the statistical analysis results for all proposed imputation models. In this table, N is the number of stations that measure each pollutant. The table also shows the Fraction of predictions within the factor of two (FAC2), Mean Bias (MB), Normalised Mean Bias (NMB), Root Mean Squared Error (RMSE), Coefficient of correlation (R), and Index of Agreement (IOA), these metrics as previously described in Section 2.4.3.

In general, model 9 (Median.ENV), which is the model that uses the ensemble technique of other models considering a station type with geographical neighbors, gives the lowest error average (RMSE), the highest Pearson correlation coefficient (R), and the highest agreement between imputed and observed concentrations based on (IOA) for NO_2 and O_3. However, model 8 (Median) achieving slightly higher performance with $PM_{2.5}$ and PM_{10}.

Since, NO_2 and O_3 concentrations influence by station's location and type, station environment type is important for their imputation. NO_2 shows local patterns, as it is concentrated where it is emitted in urban areas and near to the roadside. NO_2 is shorter lived and shows greater spatial variability, with concentrations being strongly influenced by the environment type (e.g. roadside, urban background,

rural), that change concentrations from one location to another based on the environmental type [34]. That is why station environment type is important for NO_2 imputation, we can see that using model 5 (1NN.ENV), which considers nearest neighbour from same environment type to impute NO_2 in a station, reduces the error average (RMSE) by 2.74 compared to model 4 (1NN), which considers nearest neighbour in general. However, the best RMSE is given by imputation using model 9 (Median.ENV).

A similar pattern emerges for O_3 as that has strong spatial correlation and anti-correlation with NO_2. Using model 5 (1NN.ENV) reduces the RMSE by 0.51 compared to model 4 (1NN), but the best is still model 9 (Median.ENV).

On the other hand, $PM_{2.5}$ and PM_{10} have wider spatial correlation and less variations between sites, so using the nearest neighbours in model 4 (1NN) and model 6 (2NN) gives better imputation than considering neighbours within the same environment type (model 6 and 7 respectively), shown by increases of the error average (RMSE). That is also shown as model 8 (Median) gives better imputation than model 9 (Median.ENV).

All the selected models performed well, with 71-89 % of their imputations falling within a factor of two of the observed concentrations as shown by the FAC2 values in Table 7.1. According to [49], an air quality model minimum requirement is that the FAC2 value is higher than 0.50 and NMB values should be in the range between -0.2 and +0.2. Both are met by our models. NMB measures if the models under or over predict, as it estimates the difference between the mean observed and imputed concentrations. Negative NMB means that the models under-predict and vice versa. All the models have very small biases.

7.2.2 Model Evaluation Based on Taylor's Diagram Analysis.

We use Taylor's diagram to analyse three main statistics: correlation coefficient (R), the standard deviation (sigma) and the root-mean-square error (centred), as ex-

Table 7.1: Performance of the hourly pollutant concentrations imputation models based on statistical measures. Best values in bold for FAC2, RMSE, R and IOA

Imputation models	N	FAC2	MB	NMB	RMSE	R	IOA
			NO_2				
model 1 (CA)	157	0.628	0.010	0.000	18.325	0.514	0.599
model 2 (CA+ENV)	157	0.708	0.248	0.010	15.985	0.665	0.661
model 3 (CA+REG)	157	0.630	0.171	0.007	18.360	0.527	0.600
model 4 (1NN)	157	0.605	2.278	0.095	22.587	0.465	0.533
model 5 (1NN.ENV)	157	0.639	1.573	0.066	19.842	0.579	0.593
model 6 (2NN)	157	0.618	2.774	0.116	20.456	0.494	0.558
model 7 (2NN.ENV)	157	0.679	0.742	0.031	17.277	0.639	0.637
model 8 (Median)	157	0.675	0.109	0.005	16.810	0.616	0.642
model 9 (Median.ENV)	157	**0.714**	-0.675	-0.028	**15.668**	**0.680**	**0.674**
			O_3				
model 1 (CA)	70	0.867	0.012	0.000	15.284	0.794	0.711
model 2 (CA+ENV)	70	0.876	1.070	0.022	14.663	0.814	0.728
model 3 (CA+REG)	70	0.872	0.025	0.001	15.082	0.805	0.722
model 4 (1NN)	70	0.831	-1.068	-0.022	17.547	0.756	0.681
model 5 (1NN.ENV)	70	0.837	-0.923	-0.019	17.037	0.768	0.691
model 6 (2NN)	70	0.870	-0.735	-0.015	15.200	0.807	0.720
model 7 (2NN.ENV)	70	0.866	-0.733	-0.015	15.493	0.799	0.717
model 8 (Median)	70	**0.887**	-0.314	-0.006	13.834	0.835	0.745
model 9 (Median.ENV)	70	**0.887**	0.496	0.010	**13.741**	**0.838**	**0.747**
			$PM_{2.5}$				
model 1 (CA)	77	0.828	-0.063	-0.006	5.275	0.788	0.712
model 2 (CA+ENV)	77	0.805	-0.017	-0.002	5.610	0.760	0.694
model 3 (CA+REG)	77	0.831	-0.015	-0.001	5.064	0.809	0.724
model 4 (1NN)	77	0.778	0.066	0.007	5.545	0.790	0.698
model 5 (1NN.ENV)	77	0.756	0.098	0.010	6.156	0.743	0.665
model 6 (2NN)	77	0.813	0.058	0.006	4.961	0.823	0.725
model 7 (2NN.ENV)	77	0.799	0.179	0.018	5.478	0.782	0.698
model 8 (Median)	77	**0.846**	-0.093	-0.009	**4.756**	**0.832**	**0.742**
model 9 (Median.ENV)	77	0.837	-0.052	-0.005	4.985	0.814	0.729
			PM_{10}				
model 1 (CA)	75	0.857	-0.140	-0.009	8.750	0.669	0.667
model 2 (CA+ENV)	75	0.848	-0.132	-0.008	9.033	0.651	0.662
model 3 (CA+REG)	75	0.859	-0.023	-0.001	8.798	0.674	0.670
model 4 (1NN)	75	0.813	0.113	0.007	10.359	0.610	0.627
model 5 (1NN.ENV)	75	0.811	-0.002	0.000	10.775	0.577	0.620
model 6 (2NN)	75	0.856	0.121	0.007	9.229	0.662	0.668
model 7 (2NN.ENV)	75	0.843	0.036	0.002	9.422	0.642	0.655
model 8 (Median)	75	**0.880**	-0.196	-0.012	**8.225**	**0.715**	**0.697**
model 9 (Median.ENV)	75	0.873	-0.263	-0.016	8.450	0.698	0.689

plained in Section 3.4.2. The standard deviation represents the variability between modelled and observed concentrations. The observed variability is plotted on the x-axis. The magnitude of the variability is measured as the radial distance from the plot's origin. The black dashed line shows this for the observed value. The grey lines are isopleths for the correlation coefficient (R) as indicated by the arc shaped axis; the correlation increases along the arc towards the x-axis. The centred root-mean square error (RMS) is represented by the concentric brown dashed lines.

The furthest the points/models are from the observed value the worst performance they have [33]. Figures. 7.1 and 7.2 show Taylor's Diagram plots for all imputation models with all pollutants.

In almost all cases the models exhibit less variability than the observed, indicated by the points being closer to the origin than the black dashed line. In general, Model 4 (1NN) followed by Model 5 (1NN.ENV) show variability that is most similar to the observations as indicated by their relative closeness to the black dashed line. However, these models tend to have the lowest correlation coefficients, indicated by the grey lines, and the greatest RMSE, indicated by the brown dashed lines.

Models 4, 5, 6, and 7 use the concentrations from a single site or two (i.e. the nearest stations) in the imputation, where as the other models use a cluster average (CA, CA+REG, CA+ENV) or a model ensemble average (Median and Medina.ENV), so it is reasonable for model 4, 5, 6 and model 7 to have fairly similar variability to the observed concentrations. All the other models display less variability than the observed concentrations (as indicated by their points being further from the black dashed line) which may be consistent with their derivation methods which may smooth out some of the variability.

Model 8 (Median) and model 9 (Median.ENV), regardless of their ability to capture variability, are confirmed as having the highest correlation coefficient, and the lowest centred root means squared. As confirmed by the statistical analysis model 8 performs better with PM, while model 9 does better with NO_2 and O_3. Hence, with Taylor's Diagram analysis we come to the same conclusion derived by the statistical analysis that shows how model 4 (1NN) and model 5 (1NN.ENV) are different in their correlation and centred RMSE in NO_2 and O_3 imputation comparing to PM imputation, even though they are equivalent in their variability/standard deviation.

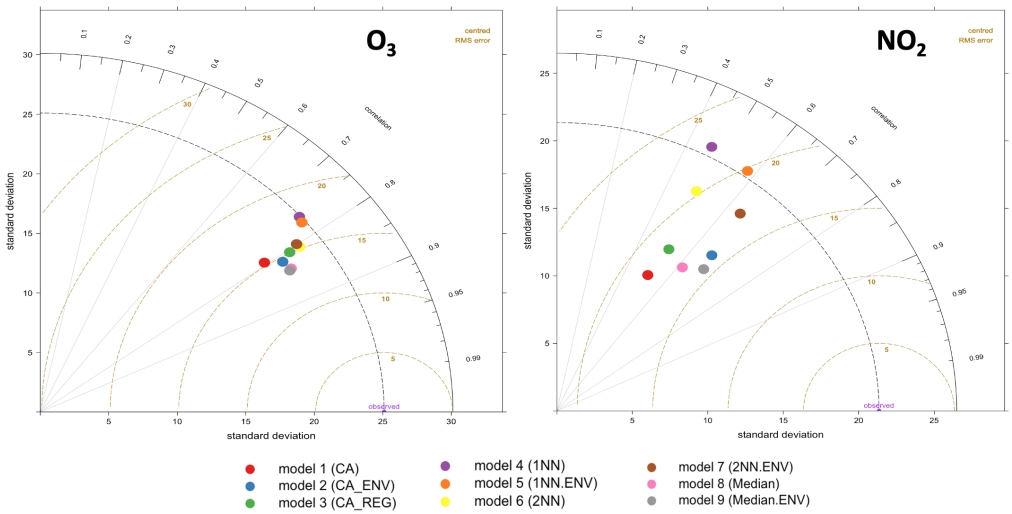


Figure 7.1: Taylor diagrams comparing observed and modelled concentrations from nine imputation models for O_3 (left plot) and NO_2 (right plot).

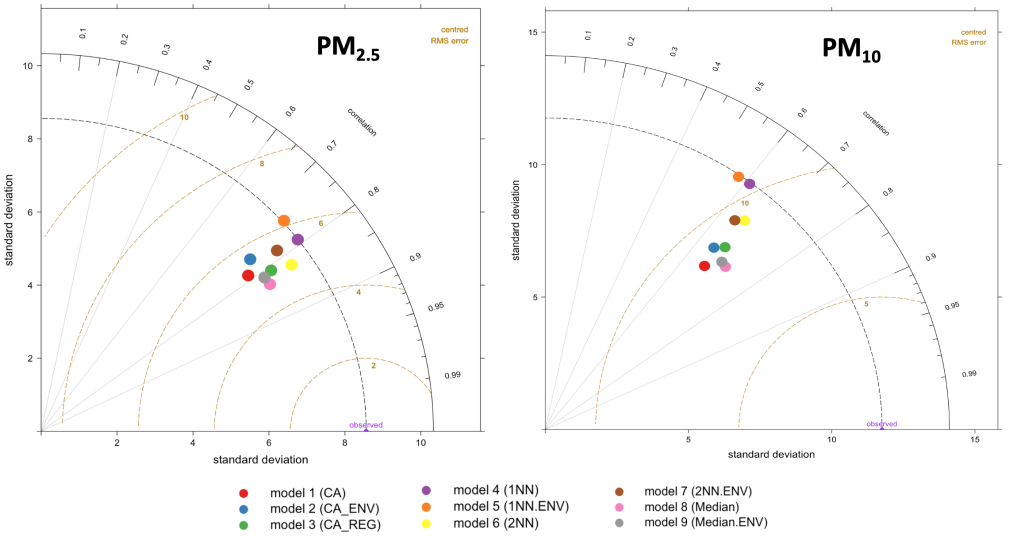


Figure 7.2: Taylor diagrams comparing observed and modelled concentrations from nine imputation models for $PM_{2.5}$ (left plot) and PM_{10} (right plot).

7.2.3 Model Evaluation Based on Conditional Quantile Analysis.

We analyse the spread of the modelled and observed pollutant concentrations using conditional quantile plots. Figures. 7.3 and 7.4 show the conditional quantile plots for the nine imputation models (panels A to I). This visualisation splits the concentrations into bins according to values of the modelled concentrations. The

median line of these values and the 25/75th and 10/90th quantile values are plotted together with a blue line showing a "perfect" model. Also shown are histograms of modelled concentrations (shaded grey bars) and histograms of observed concentrations (blue outline bars).

These plots show how the modelled concentrations compare with the observed concentrations and how the models capture the variability in the concentrations. The spread of the modelled concentrations around the perfect model line (blue line) are shown by the shaded portions/quantile intervals. If narrow, it indicates high agreement/precision between the modelled and observed concentrations. The quantile intervals also represent the uncertainty bands. In some cases these intervals do not extend along with the median line due to insufficient concentrations to calculate them. A good model is obtained when the median (red line) coincides with the perfect model (blue line) and when the spread in the percentile is as narrow as possible.

From these plots, in general, the histograms indicate that model 4 (1NN) (Panel D) and model 5 (1NN.ENV) (Panel E) have better estimation of the variability between the observed and modelled concentrations, as observed before, even though the median line does not match the perfect model. These models are positively biased at high concentrations, as shown by the departure of the median line below the blue line for all pollutants. This result supports our analysis from the Taylor's diagram that model 4 (1NN) and model 5 (1NN.ENV) have the lowest variability between modelled and observed concentrations, but with lower correlation coefficient and the highest-centred root means squared for all pollutants.

In Figure 7.3, (right plot), NO_2 models show different behaviours from this analysis. Even though the statistical analysis shows that model 9 (Median.ENV) (Panel I) gives the best performance, it is clear that in this model, the modelled concentrations tend to be lower than observations for most concentration levels (the medians are under the blue line) and the width of the 10/75th and 10/90th percentiles is quite broad. The only advantage of using this model is its ability to capture a wide

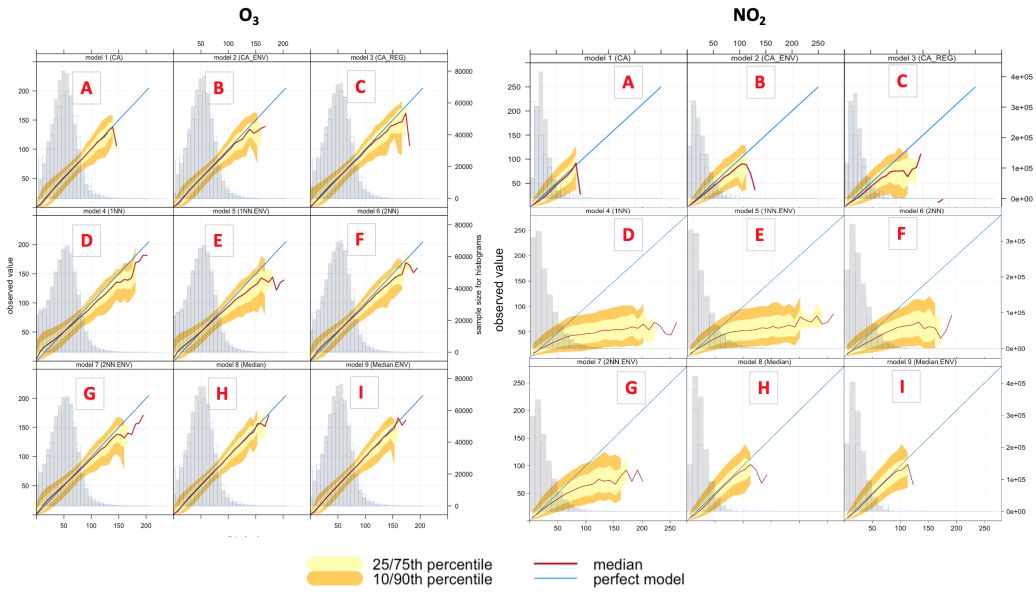


Figure 7.3: Conditional quantile plot of modelled and observed pollutants concentrations of O_3 (left plot) and NO_2 (right) for proposed imputation models; (A) model 1 (CA), (B) model 2 (CA+ENV), (C) model 3 (CA+REG), (D) model 4 (1NN), (E) model 5 (1NN.ENV), (F) model 6 (2NN), (G) model 7 (2NN.ENV), (H) model 8 (Median), and (I) model 9 (Median.ENV).

range of concentrations. Model 4 (1NN) (Panel D) compared to other models can reproduce the higher concentrations (higher than 125 μg m^{-3}) as it does not take an average approach. However, this model is positively biased (NMB = 0.095), which is shown by the departure of the median line from the blue one.

In the same figure (left plot), O_3 models show that most modelled concentrations match the observations well for a wide range of values. The histograms indicate underestimation in general at the extreme low and high concentrations. In general, the cluster and median imputation methods (i.e. that use averaging) will tend to struggle to reproduce the lowest and highest concentrations since they take an average approach. Moreover, the highest concentrations are typically limited to relative few data points. The cases of the high ozone concentrations typically occur during specific meteorological conditions and are episodic in nature, and there may be small differences in timings of the peak concentrations at different sites. The very low ozone concentrations are likely to occur at specific sites (near to roads where emissions of nitric oxide are large) and so may not be reproduced

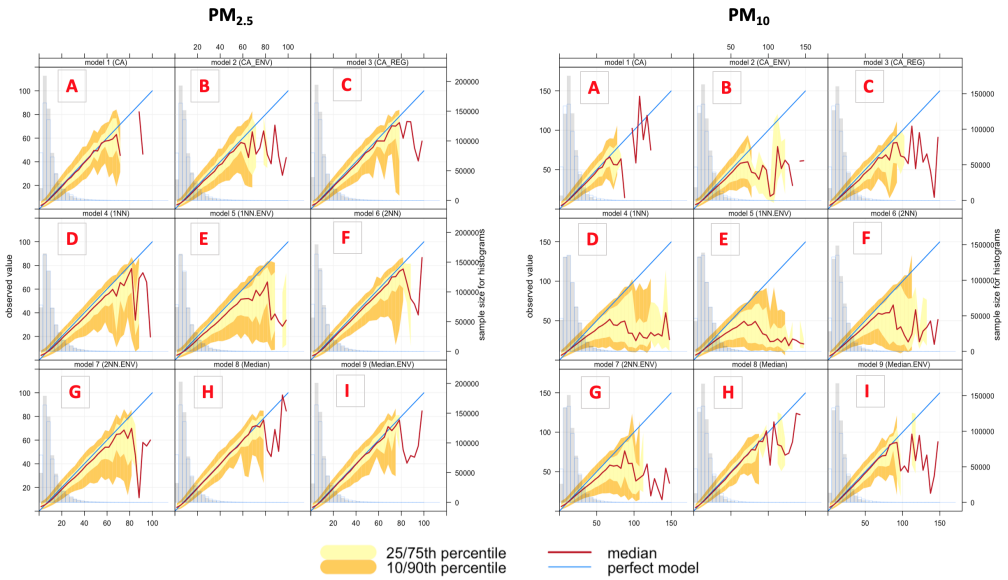


Figure 7.4: Conditional quantile plot of modelled and observed pollutants concentrations of $PM_{2.5}$ (left plot) and PM_{10} (right plot) for proposed imputation models; (A) model 1 (CA), (B) model 2 (CA+ENV), (C) model 3 (CA+REG), (D) model 4 (1NN), (E) model 5 (1NN.ENV), (F) model 6 (2NN), (G) model 7 (2NN.ENV), (H) model 8 (Median), and (I) model 9 (Median.ENV).

in the models which take a cluster average or where a nearest neighbour site is not a similar type of site. An example of high ozone episodes is what happened in Summer 2018 in the UK, where two serious wildfires to the moors in the North West of England coincided with a national heatwave caused a very high pollution level specially with ozone and particulate [3].

Model 9 (Median.ENV) (Panel I) has the best performance indicated by an overlapping median line with the blue line. This model has the lowest RMSE and the highest degree of agreement indicated by the narrow spread of the modelled concentration quantile intervals.

The variation between $PM_{2.5}$ models in Figure 7.4 (left plot) show similar performance for the different models. The quantile intervals are wider within the area of high concentrations $\leq 60 \mu g m^{-3}$, and all models underestimate the high concentrations $\leq 80 \mu g m^{-3}$. Note that these concentrations are very low frequency events.

Model 8 (Median) (Panel H) gives better performance indicated by the narrow spread of the modelled concentration quantile intervals and minimal bias, indicated by the overlaps between the red and blue lines compared to other models. Models for PM_{10} (right plot) show similar performance to $PM_{2.5}$.

From the histograms in Figures. 7.3 and 7.4, we note that the overall distribution of the observed concentrations of NO_2, $PM_{2.5}$, and PM_{10} are skewed to the lower values, while O_3 has a more normal distribution.

We also notice, from these histograms, that the distributions of the modelled O_3 concentrations are shifted to lower values, while other pollutants modelled concentrations are shifted to higher values. Hence the model is not always able to reproduce the edges of the distribution correctly. The skewness in modelled values is mostly associated with model 1 (CA). As a consequence, this shows the greatest difference in skewness between the distributions of the observed and modelled values. However, model 1 (CA) in combination with others, as part of model 8 (Median) and model 9 (Median.ENV), reduce the skewness in modelled values and generates better imputation, resulting in the lowest RMSE.

7.2.4 Model Evaluation Based on Conditional Quantile Analysis for Station's Environmental Types.

In this analysis, we focus on the performance of model 8 (Median) and model 9 (Median.ENV), as those performed best for the different pollutants in the previous section, but now we break down the analysis for the six environmental types (background rural, background urban, background suburban, and industrial urban, industrial suburban, and traffic urban) to which stations belong. Notice that a pollutant may/may not be measured in all stations and the number of stations under each type is different as shown in Table 7.2. We also use conditional quantiles to analyse our model's performance within each environmental type.

First, we show the monthly average concentrations for each pollutant under each

Table 7.2: Performance of the hourly pollutant concentrations imputation models using model 9 (Median.ENV) for NO_2 and O_3, and Model 8 (Median) for $PM_{2.5}$, and PM_{10} based on statistical measures for all station environment types for all pollutants.

Imputation models	**Environment Type**	**N**	**MB**	**NMB**	**RMSE**
		NO_2			
model 9 (Median.ENV)	Background Rural	15	2.486	0.352	7.919
model 9 (Median.ENV)	Background Suburban	5	5.302	0.379	12.236
model 9 (Median.ENV)	Background Urban	58	2.038	0.103	11.836
model 9 (Median.ENV)	Industrial Suburban	4	2.742	0.126	11.720
model 9 (Median.ENV)	Industrial Urban	11	1.655	0.092	9.705
model 9 (Median.ENV)	Traffic Urban	65	-4.578	-0.140	20.253
		O_3			
model 9 (Median.ENV)	Background Rural	19	-3.161	-0.054	13.947
model 9 (Median.ENV)	Background Suburban	3	1.618	0.032	14.143
model 9 (Median.ENV)	Background Urban	40	0.847	0.018	13.042
model 9 (Median.ENV)	Industrial Suburban	1	-1.392	-0.030	11.311
model 9 (Median.ENV)	Industrial Urban	4	2.554	0.055	12.547
model 9 (Median.ENV)	Traffic Urban	3	15.463	0.473	21.286
		$PM_{2.5}$			
model 8 (Median)	Background Rural	5	2.206	0.301	5.014
model 8 (Median)	Background Suburban	2	-0.124	-0.011	3.447
model 8 (Median)	Background Urban	41	-0.022	-0.002	4.695
model 8 (Median)	Industrial Urban	6	0.156	0.017	4.013
model 8 (Median)	Traffic Urban	23	-0.732	-0.069	5.106
		PM_{10}			
model 8 (Median)	Background Rural	5	4.244	0.376	8.053
model 8 (Median)	Background Urban	26	1.251	0.084	7.098
model 8 (Median)	Industrial Urban	7	-0.029	-0.002	10.023
model 8 (Median)	Traffic Urban	37	-1.921	-0.105	8.587

environment type in our test dataset (i.e. data for year 2018), to help understanding the normal variation of the pollutant concentrations in different environment types. Then, Figures. 7.6, 7.8, 7.10, and 7.12 show conditional quantile plots by the environmental types for the selected models. Table 7.2 shows the statistical measures of performance also broken down by environment type.

As known, that the most common sources for NO_2 are roads, however NO_2 concentrations are influenced by traffic density, road locations, and meteorological conditions, which cause variation from one roadside location to another. Figure 7.5 shows that high NO_2 concentrations are found at traffic urban followed by industrial suburban, then background urban sites, while the background rural sites

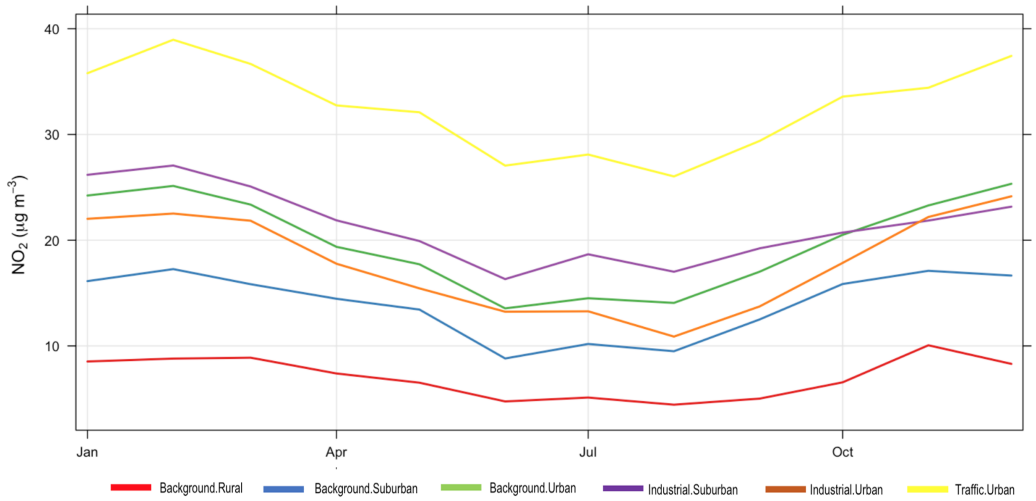


Figure 7.5: Monthly average concentrations of observed NO_2 for each environmental types for year 2018.

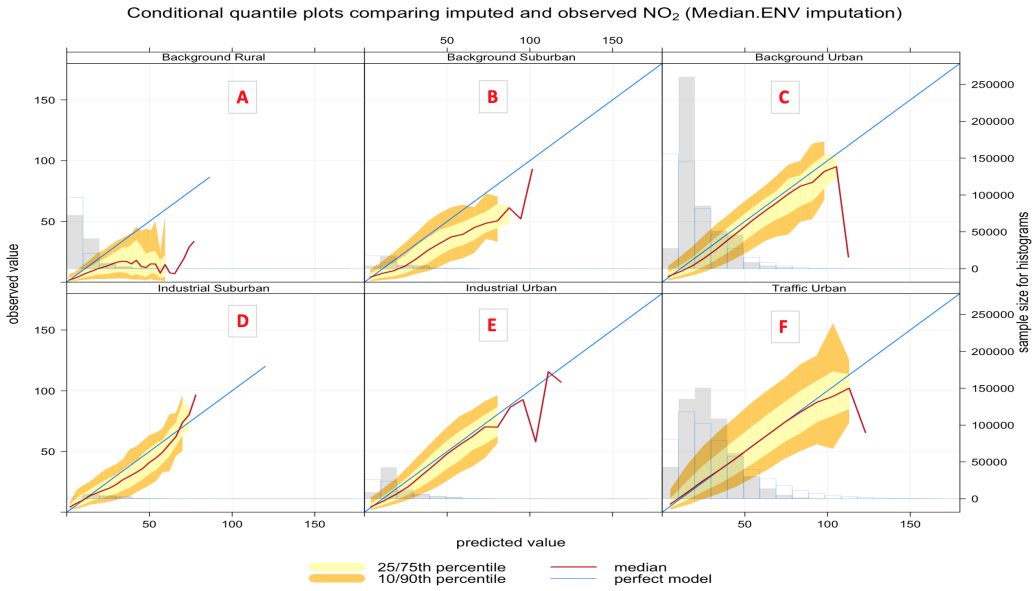


Figure 7.6: Conditional quantile plot of modelled and observed pollutants concentrations of NO_2 based on model 9 (Median.ENV) for all station environmental types, (A) background rural, (B) background suburban, (C) background urban, (D) industrial suburban, (E) industrial urban, and (F) traffic urban stations.

have the lowest NO_2 concentrations.

Figure 7.6 shows the conditional quantile plots by station type for NO_2 imputation using model 9 (Median.ENV). Here, we see that modelled concentrations are higher than observed concentrations with all environmental types excepts with

traffic urban sites. This is confirmed by all the statistical model quality measures presented in Table 7.2 where we can observe positive mean bias for NO_2 with all station types and a negative mean bias with traffic urban sites.

As NO_2 distributions in general are skewed to the lower values, and our selected model model 9 (Median.ENV) is based on the average concentrations, the model performs better with lower concentrations.

From Table 7.2 based on model RMSE, the model's best performance is associated with background rural stations, while the worst performance is shown for traffic urban stations. Contrasting this with quantile plots, Figure 7.6 (Panel A) shows that for background rural stations the histogram and the median line show better performance with lower concentrations (less than 30 μg m^{-3}). On the other hand, for traffic urban stations (Panel F), the quantile intervals are wider within the area of high concentrations (higher than 25 μg m^{-3}), and the modelled concentrations tend to be lower than observed concentrations.

As, we mentioned earlier that NO_2 is short lived so it has large differences between sites near sources (roadside) and those further away. Based on the RMSE Model 9 (Median.ENV) performs better with lower NO_2 concentrations than high values, and since these high NO_2 values exist near to traffic, the model performs the worst with traffic urban stations as shown in Figure 7.6 (Panel F). In contrast, the model best performance is associated with background rural stations that have the lowest NO_2 concentrations.

Figure 7.7 shows the monthly average concentrations of observed O_3 concentrations at each environment type. From that we can see that ozone in all environment types follow a similar trend. However, background rural stations have the highest concentrations and traffic urban have the lowest. Looking at model 9 (Median.ENV) performance in Table 7.2 based on the RMSE, the best model performance is associated with industrial suburban site, where there is only one site in this group. Hence, the next best RMSE associated with industrial urban sites and its average

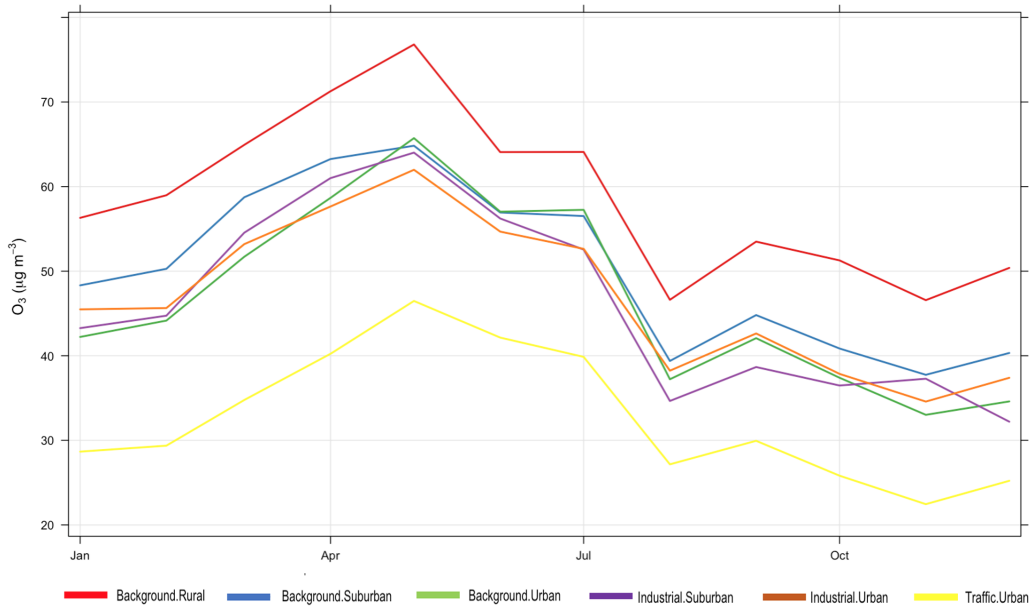


Figure 7.7: Monthly average concentrations of observed O_3 for each environmental types for year 2018.

performance is associated with background rural stations (those with higher concentrations in Figure 7.7), while its worst performance is associated with traffic urban stations (those with lower concentrations in Figure 7.7).

Conditional quantile analysis in Figure 7.8, shows the performance of model 9 (Median.ENV) for imputing O_3 for the six environmental types (Panels A to F). The model shows similar performance for industrial suburban (Panel D) and background rural stations (Panel A). For both types, the model is negatively biased (see also Table 7.2), meaning that the modelled concentrations tend to be lower than observed concentrations (the median lines are above the blue lines). The worst performance based on the RMSE is associated with traffic urban stations (Panel F), which are the stations located at the roadsides. With those stations, the modelled concentrations are higher than observed concentrations, i.e. shifted to the right. This is indicated by the model positive bias (0.473). The best model performance is associated with industrial urban stations (Panel E) according to the RSME, even though background urban stations (Panel C) appear to have the best performance by looking at the conditional quantile plots. The histogram of Panel

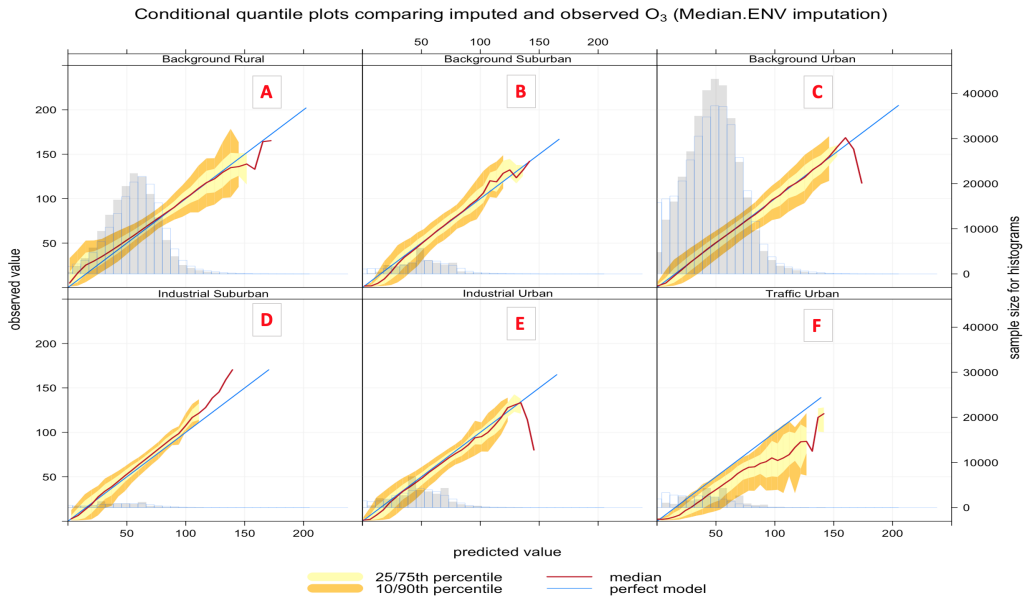


Figure 7.8: Conditional quantile plot of modelled and observed pollutants concentrations of O_3 based on model 9 (Median.ENV) for station environmental types, (A) background rural, (B) background suburban, (C) background urban, (D) industrial suburban, (E) industrial urban, and (F) traffic urban stations.

C indicates that the distribution of the observed and modelled concentrations tend to be closer to each other for higher concentrations. However, the model overestimates the average concentrations at these stations (between 25 to 70 μg m^{-3}) and underestimates the very low concentrations.

Model 9 (Median.ENV) performance with O_3 imputation changes from one environmental type to another due to the ozone's behaviour at these locations. As we know, ozone is not directly emitted into the air, but it is formed as a secondary pollutant by chemistry involving nitrogen oxides (NO_x), the sum of NO_2, nitric oxide (NO) and volatile organic compounds (VOC) in the presence of sunlight [48]. This chemistry is non-linear and newly emitted NO can react with O_3 leading to reductions in O_3 concentrations close to sources of NO (e.g. in urban areas and in particular, close to roads). Consequently, ozone concentrations in urban areas are often lower than those at rural areas [71], as shown in Figure 7.7.

Figure 7.8 (Panel A), shows see that the model produces a distribution shifted to the left toward lower values, not capturing the ozone for rural areas that are

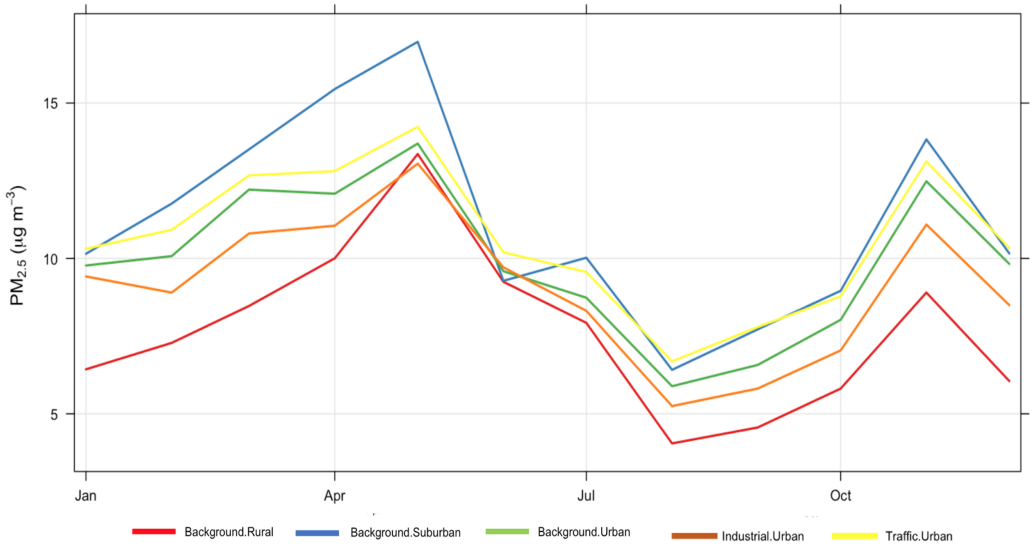


Figure 7.9: Monthly average concentrations of observed $PM_{2.5}$ for each environmental types for year 2018.

associated with higher concentrations of O_3. Similarly, industrial suburban stations (Panel D) have a higher frequency of high concentrations (higher than 25 μg m^{-3}), as shown in the histogram (Panel D). Note that majority of stations measuring O_3 are background rural or background urban, with few stations in other categories. While with traffic urban (Panel F), where the model performs the worst, some modelled concentrations are much higher than observed measurements. This lack of fit may be explained because ozone is suppressed by new emissions of NO close to sources (traffic) which reduce the amount of O_3 in those station types.

From the same figure (Panel C), as shown in the histogram for background urban stations, there is a high frequency of low concentrations (less than 10 μg m^{-3}) at these stations that the model does not capture. This is consistent with the reaction of newly emitted NO from urban roadside that reduces the concentrations of ozone at urban areas. Based on the RMSE and NMB, the model is a middle performing model. As shown in Table 7.2, the majority of stations measuring O_3 belong to this type.

Figure 7.9, shows $PM_{2.5}$ concentrations at rural areas are lower that those at suburban, urban background and traffic urban areas. That is consistent with the model

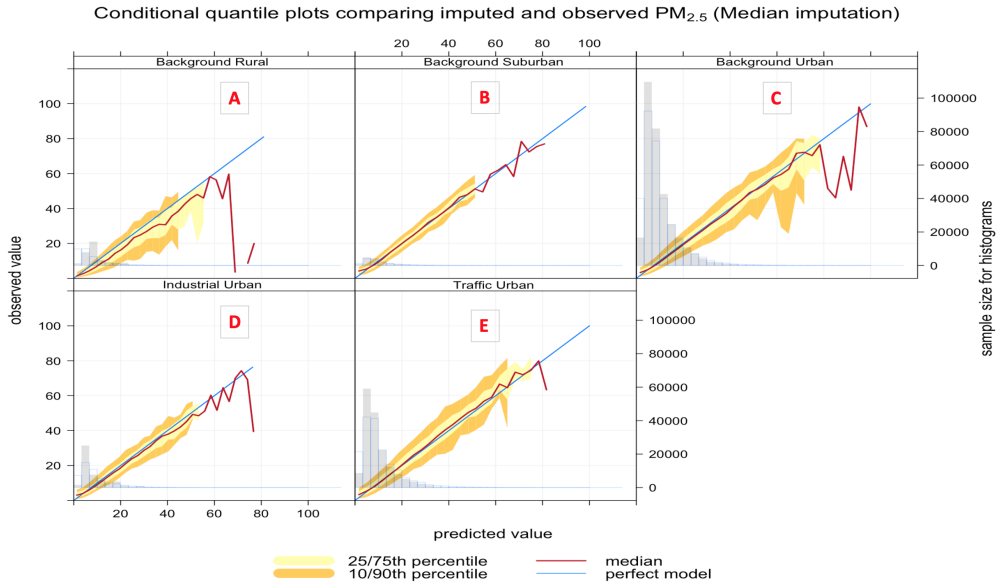


Figure 7.10: Conditional quantile plot of modelled and observed pollutants concentrations of $PM_{2.5}$ based on model 8 (Median) for station environmental types, (A) background rural, (B) background suburban, (C) background urban, (D) industrial suburban, and (E) traffic urban stations.

performance at these sites. Figure 7.10 shows corresponding conditional quantile plots by station types. Model 8 (Median) for imputing $PM_{2.5}$ concentrations has similar performance among different station types. In general, the model underestimates the concentrations of $PM_{2.5}$ especially for high concentration levels. Table 7.2, shows that the model underestimates high concentrations at suburban, urban background and traffic urban areas, indicated by the model negative biases, while it overestimates the concentrations at industrial urban and background rural sites. The model shows worst performance for traffic urban (panel E), and this is also indicated by the highest RMSE (5.106) shown in Table 7.2. The model underestimates the concentrations at these stations, which is confirmed by the model bias (-0.069) in Table 7.2. On the other hand, the model's best performance is associated with background suburban sites (Figure 7.10 (panel B)), even though it underestimates $PM_{2.5}$ concentrations with a mean bias of (-0.011).

Finally, PM_{10} levels at background rural and urban areas are lower that those at industrial and traffic urban areas as shown in Figure 7.11. For PM_{10}, imputation

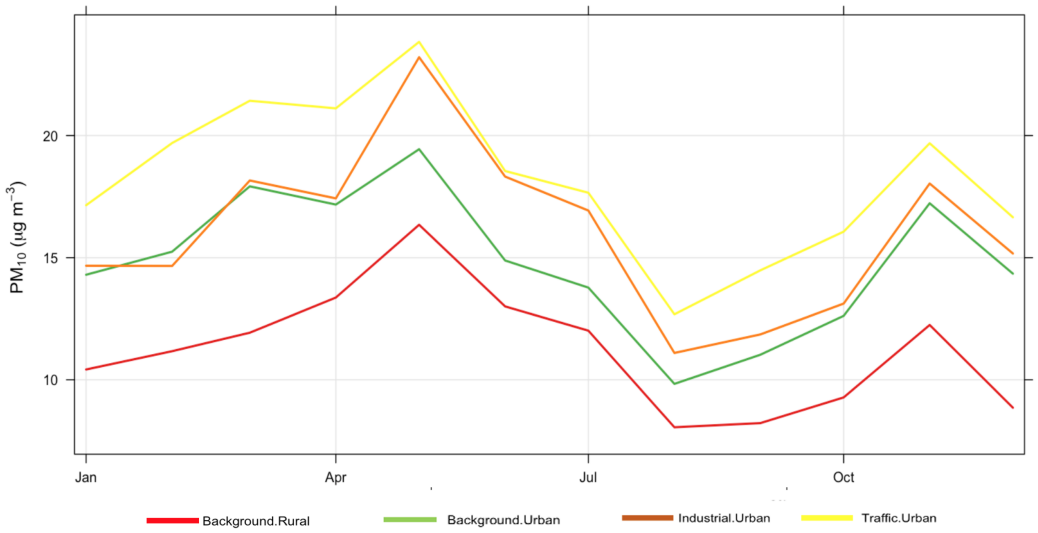


Figure 7.11: Monthly average concentrations of observed PM_{10} for each environmental types for year 2018.

performance shown in Figure 7.12 is similar for background urban and background rural sites (panels A and B). The model overestimates the concentrations of PM_{10} that are $\leq 10\mu gm^{-3}$, while it underestimates the high concentrations of PM_{10} at industrial urban (slightly) and traffic urban sites (panels C and D). That is confirmed by the model mean bias at these sites (-0.002, -0.105) as shown on Table 7.2.

$PM_{2.5}$ and PM_{10} have many varied sources so in roads and industrial sites it can be associated with local sources, for example widespread primary sources (direct emissions) and diffused secondary sources (i.e. produced in the atmosphere following emissions of precursor gases). At the roadside PM concentrations are often greater than other locations [2]. The particles can last for several days in the atmosphere, which make the PM concentrations distributed widely. However large particles subject to greater loss via sedimentation than others, so $PM_{2.5}$ is more evenly distributed than PM_{10} [8]. This behaviour can also be observed with Model 6 (Median) performance, where there are less variation with the model performance under different environment types compared to the variation of NO_2 and O_3, as shown in Table 7.2.

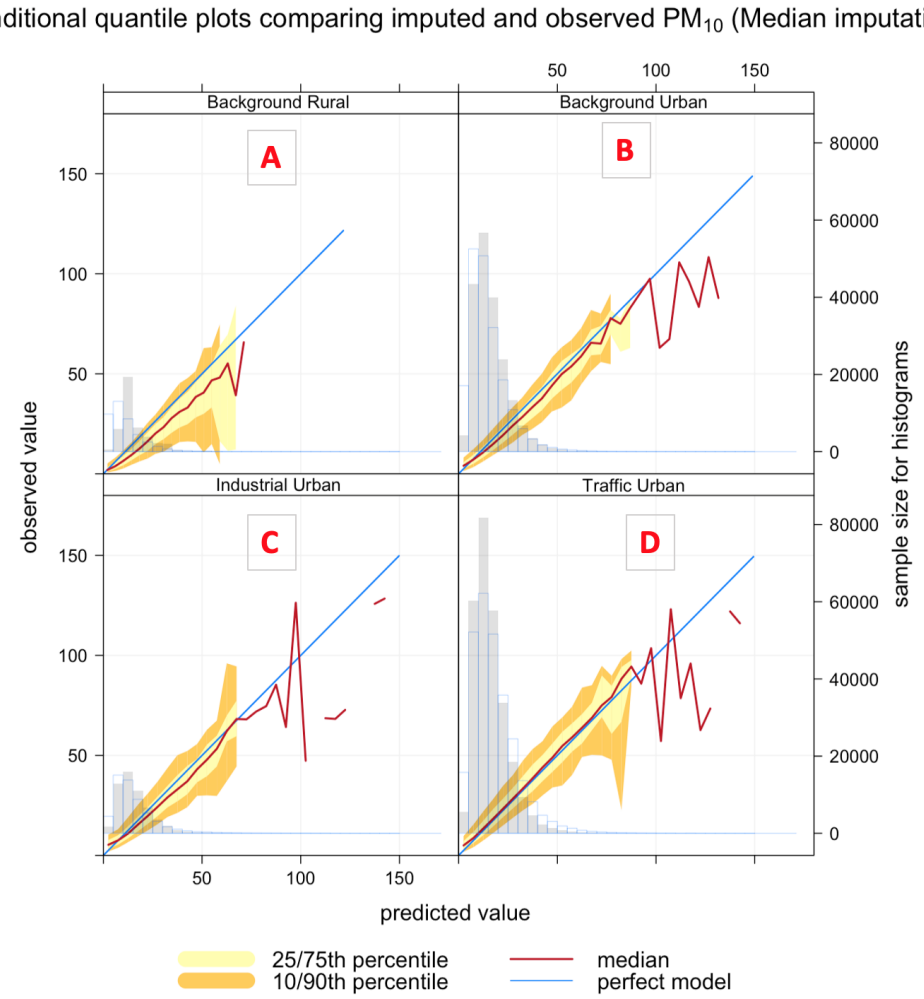


Figure 7.12: Conditional quantile plot of modelled and observed pollutants concentrations of PM_{10} based on model 8 (Median) for station environmental types, (A) background rural, (B) background urban, (C) industrial urban, and (D) traffic urban stations.

We also observed that the distributions of NO_2, $PM_{2.5}$ and PM_{10} are skewed to lower concentrations which impact model performance at higher concentrations. All models perform worse for high concentrations with NO_2, $PM_{2.5}$ and PM_{10} than O_3, indicated by the width of the quantiles at high values as shown in Figures 7.3 and 7.4. Similarly, for lower concentrations, these models tend to perform better for NO_2, $PM_{2.5}$, and PM_{10} than for O_3. However, our selected models (model 8 (Median) and model 9 (Median.ENV)) are able to overcome this impact slightly. Conditional quantile plots of modelled and observed pollutants concentrations for

each station and for each pollutant, can be found in Appendix F.

7.3 Model Evaluation Based on LWTs

We analyse the performance of the selected models under different weather types using Lamb Weather Types (LWTs), as explained in Section 3.4.2. LWTs dataset (Section 3.5.2) are a synoptic classification of daily weather patterns across the UK [91]. Since the LWTs dataset change on a daily basis, we use the daily modelled concentrations in this evaluation.

In this analysis, we use Taylar's diagram that compares how well the models reproduce the observations and how models perform relative to each other to see whether the models preform differently in different LWTs. Figures 7.13, 7.15, 7.17, and 7.19, show all nine imputation models performance under LWTs including 11 weather types for NO_2, O_3, $PM_{2.5}$, and PM_{10}. From these plots we can see that the model performance patterns are very similar for all LWTs for each pollutants, with the exception of UC, where there is very little difference in performance between models for the different LWTs. All models (with the exception of model 4 and model 5) show lower standard deviation than the observed indicating the model imputation data set to be less variable than observed, that is also confirmed by the previous analysis. Also, we can see that the variation between models performance is much higher for NO_2 compared to other pollutants.

Similarly, Figures 7.14, 7.16, 7.18, and 7.20, show model performance in the main three classification of wind sectors. Clearly we can see that models work equally well for all weather types. From these plots, we can say that the selected models (model 8 and model 9) work equally well for all LWTs for all pollutants.

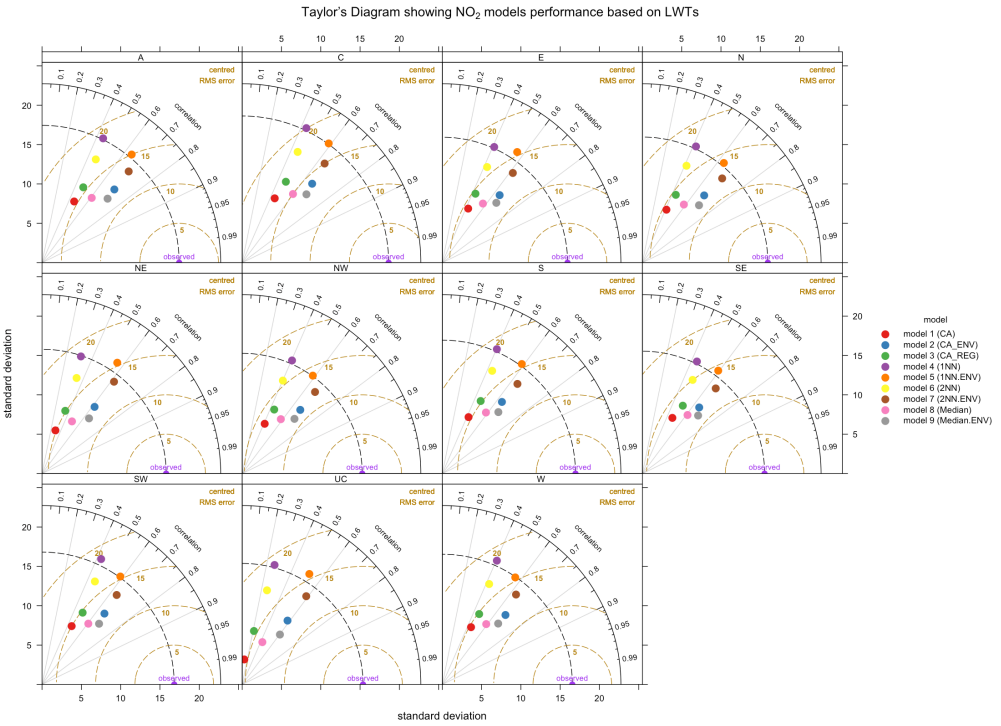


Figure 7.13: Taylor's Diagram to compare the performance of the nine imputation models to impute NO_2 under LWTs including 11 weather types.

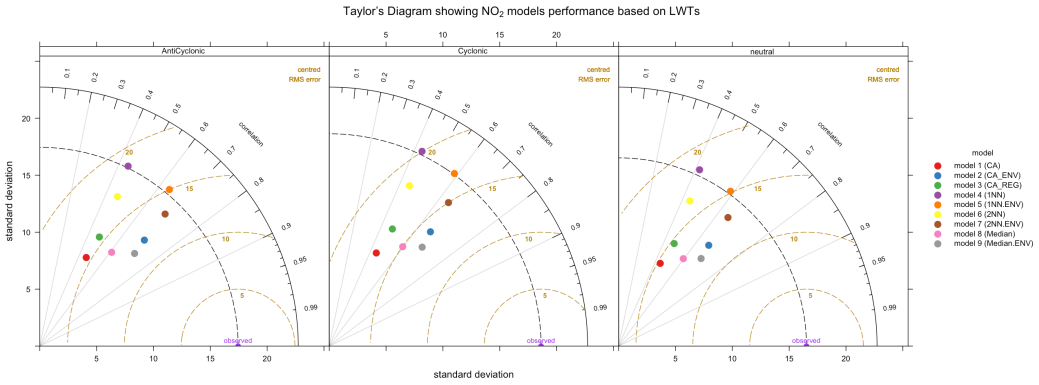


Figure 7.14: Taylor's Diagram to compare the performance of the nine imputation models to impute NO_2 based on the main three classification of wind sectors.

7.4 Evaluate Imputed Concentrations Based on The Daily Air Quality Index (DAQI).

After selecting the best imputation models for each pollutant based on the previous analysis, we impute the measured pollutants in all the stations then we calculate

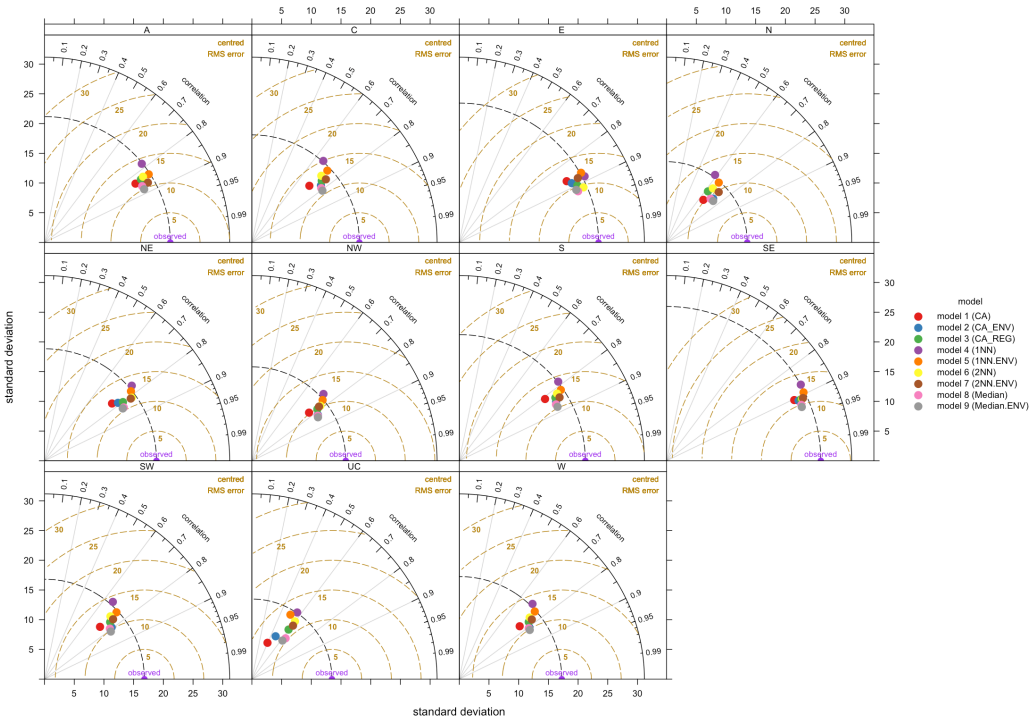


Figure 7.15: Taylor's Diagram to compare the performance of the nine imputation models to impute O_3 under LWTs including 11 weather types.

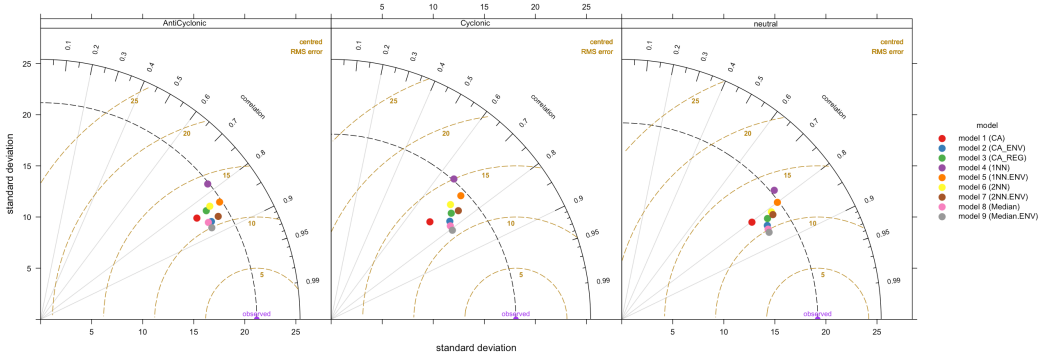


Figure 7.16: Taylor's Diagram to compare the performance of the nine imputation models to impute O_3 based on the main three classification of wind sectors.

DAQI from the imputed data. The selected models are model 9 (Median.ENV) for O_3 and NO_2, and model 8 (Median) for $PM_{2.5}$ and PM_{10}.

We use DAQI here as an evaluation function, we compare the imputed DAQI with the observed DAQI based on the RMSE, and the number of days where there are agreements and disagreements. The total number of days in our data set is 60,955 days (167 stations * 365 days), there are 2,212 days with missing observed DAQI

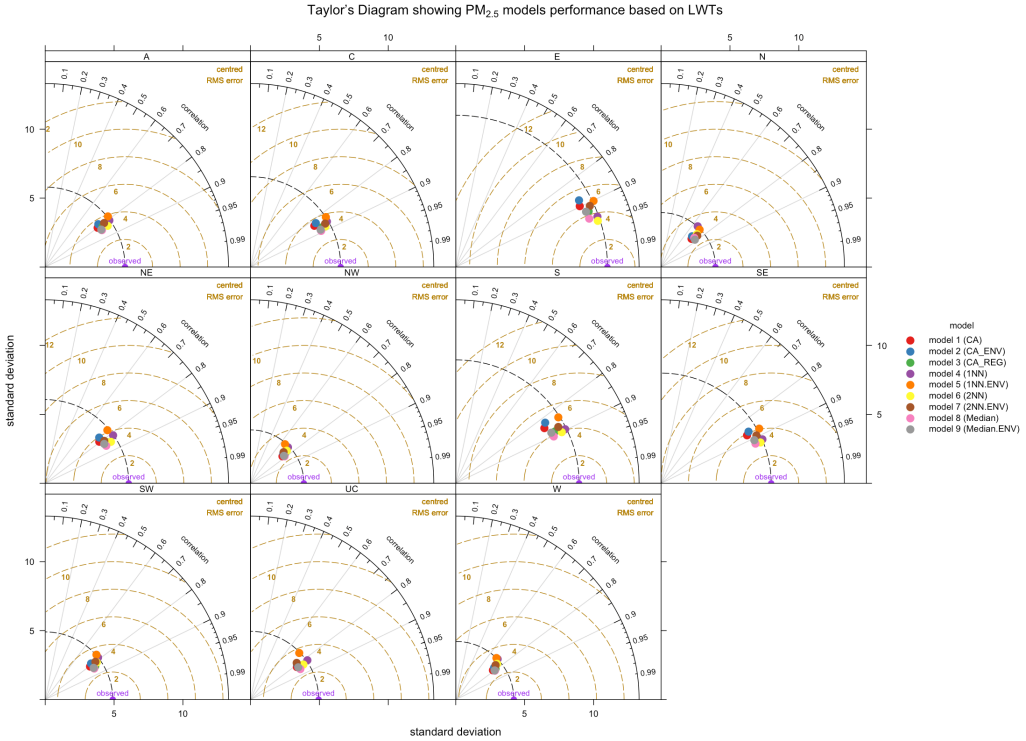


Figure 7.17: Taylor's Diagram to compare the performance of the nine imputation models to impute $PM_{2.5}$ under LWTs including 11 weather types.

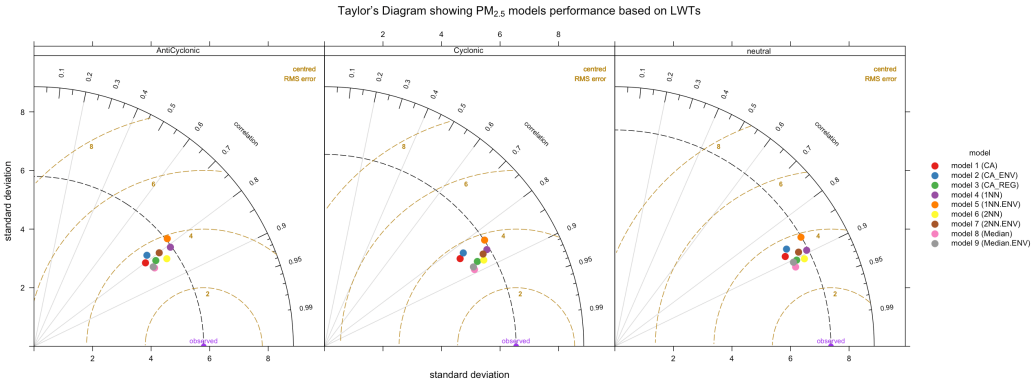


Figure 7.18: Taylor's Diagram to compare the performance of the nine imputation models to impute $PM_{2.5}$ based on the main three classification of wind sectors.

(DAQI = 0) that have resulted from missing observations on those days. The total number of days to compare is 58,743 days.

In general, the total average of RMSE from all days in all stations is (0.54). As the station type and the region may affect our imputation, Figure 7.21 shows the average RMSE based on air quality regions in the UK (Panel A), and station

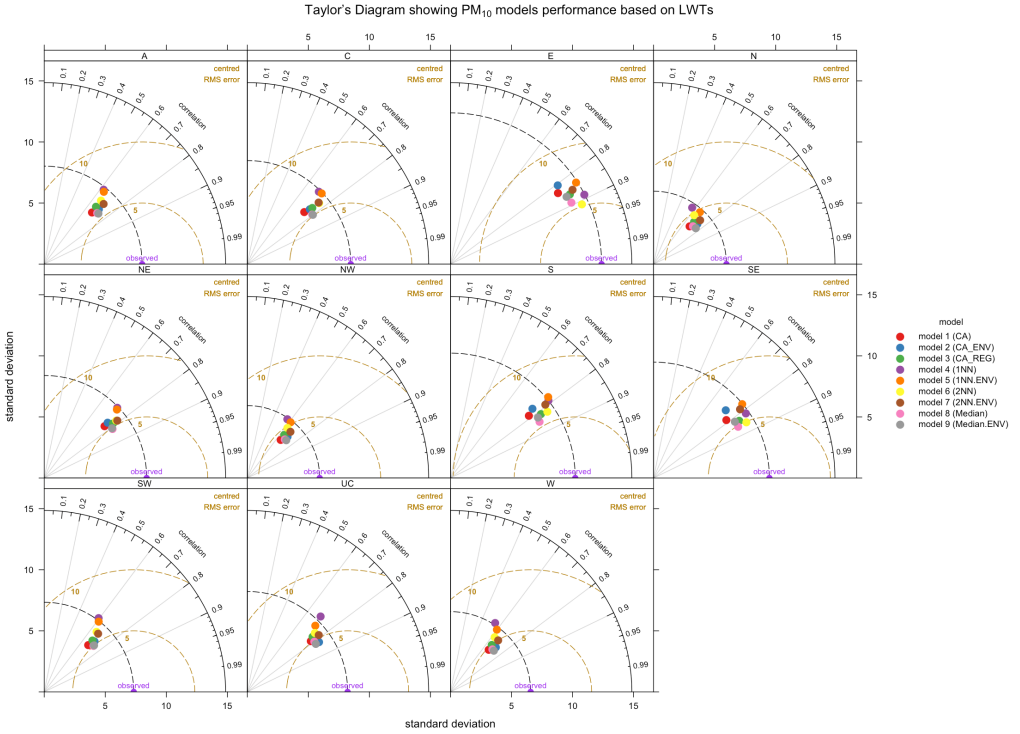


Figure 7.19: Taylor's Diagram to compare the performance of the nine imputation models to impute PM_{10} under LWTs including 11 weather types.

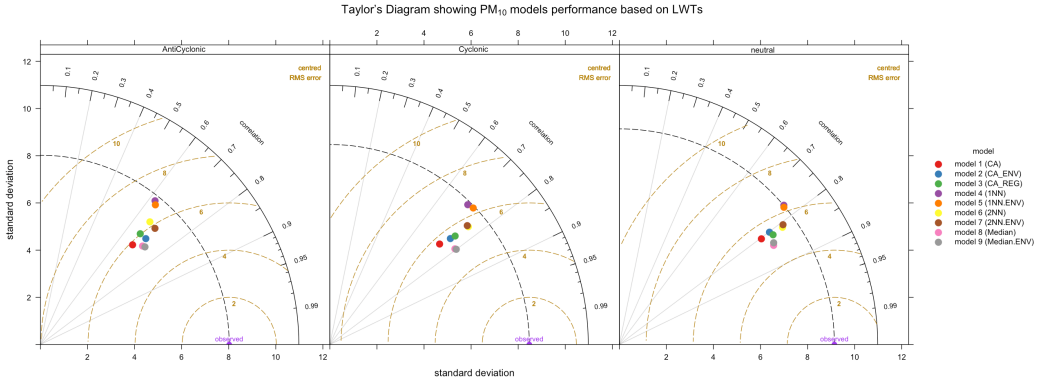


Figure 7.20: Taylor's Diagram to compare the performance of the nine imputation models to impute PM_{10} based on the main three classification of wind sectors.

environmental types (Panel B); the size of the circles are the number of stations at each type. Panel A shows that stations classed as Traffic Urban are associated with the highest RMSE (0.62), while Industrial Suburban stations have the lowest RMSE (0.37). Panel B shows that the Central Scotland region is associated with the lowest RMSE (0.44), while North Wales has the highest RMSE (0.71) between

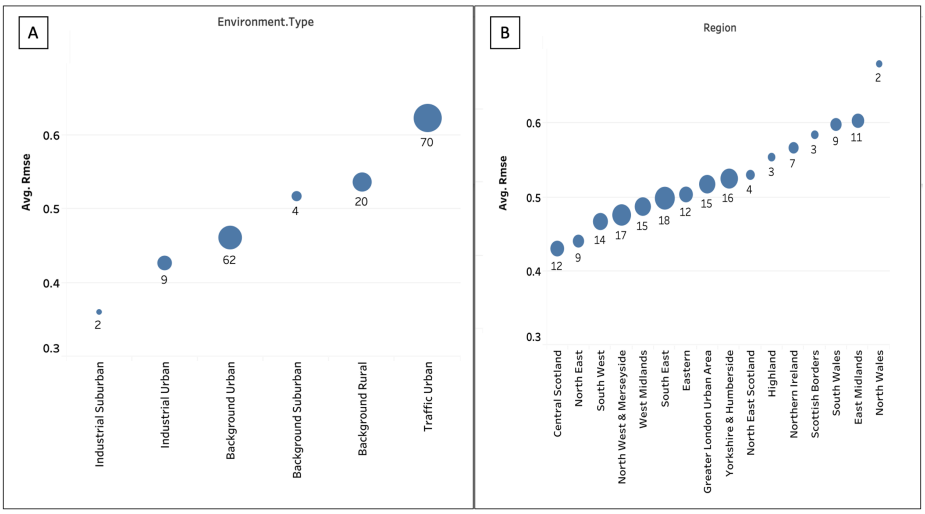


Figure 7.21: The model performance based on DAQI RMSE: (A) the average of the RMSE based on station environmental types, (B) the average of the RMSE based on air quality regions.

imputed and observed DAQI.

We also study the correlation between number of measured pollutants in a station and the agreement between modelled and observed DAQI to see if number of measured pollutants impacts our model's performance.

First, we classify stations based on number of measured pollutants to: stations that measured one, two, three and all four pollutants, as shown in Table 7.3. Each row in this table represents one group. The second column is the total number of days with associated DAQI from all stations in each group. The RMSE and index of agreement (IOA) are the average of errors and the degree of agreement between observed and modelled DAQI from all stations in each group, then the percentage of each pollutant in each group. Based on this table, we find that stations that measure four pollutants have the lowest RMSE (0.50) and the highest (IOA) (0.81), while stations that measured one pollutant have the worst performance. The majority of stations with one pollutant are stations that measure NO_2, there are 43 station out of 50 stations in this group.

We also compare the imputed and the observed DAQI based on the number of days

Table 7.3: Comparing observed and modelled DAQI based on number of measured pollutants in stations.

Number of measured pollutants	**Number of days in all stations**	**Number of stations**	**RMSE**	**IOA**	**O_3** (%)	**NO_2** (%)	**$PM_{2.5}$** (%)	**PM_{10}** (%)
1	15684	50	0.542	0.769	6.1	87.8	2.0	4.1
2	17581	48	0.583	0.756	24.5	48.0	8.2	19.4
3	15443	43	0.516	0.814	14.0	31.8	32.6	21.7
4	9398	26	0.496	0.814	25.0	25.0	25.0	25.0

where the imputed DAQI agrees and disagrees with the observed DAQI. Table 7.4, shows those results and the percentage of time that these situations occurred, meaning when agreement/disagreement is found for each DAQI. The total number of days where the imputed DAQI agrees with the observed DAQI is 44,118 day (75%), while there is 14,625 (25%) days of disagreement. We classify the disagreement into two types: the imputed DAQI is higher or lower than the observed DAQI. We find that there are 10,730 days (73% of total disagreement cases), where the imputed DAQI is lower than the observed DAQI, and 3,895 days, where the imputed DAQI is higher than the observed DAQI. In most cases, the imputed DAQI is lower than the observed DAQI, in accordance with our analysis of the imputation models that showed underestimation of the pollutant concentrations. From this table, we can see that the highest percentage of disagreement occurs in 10.4% of total number of disagreements (14,625), when the observed DAQI is 2 and imputed DAQI is 1, followed by 5.5% of disagreements when observed DAQI is 3 and imputed DAQI is 2. We can see that there is a very small percentage when there is a large difference between imputed and observed DAQI e.g. observed DAQI 1, modelled DAQI 8, and in another case when observed DAQI 8 and modelled DAQI 2.

In this table, there are 13 cases (i.e. rows marked with (*) and represents 0.1% of disagreement), where the difference between the imputed and the observed DAQI is more than three index values. We analyse the pollutant concentrations in these cases to find an explanation for these situations. We include the analysis of two cases/days in this chapter, and the analysis of other cases (11 days), can be found in Appendix G.

Our analysis of these cases found that when the observed DAQI is higher than

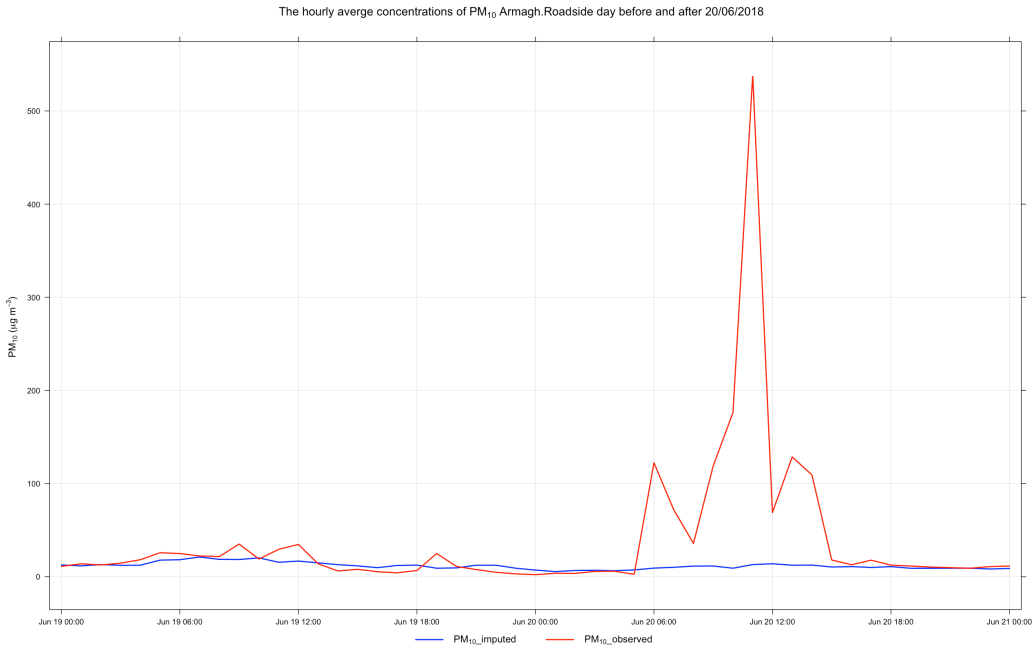


Figure 7.22: Hourly concentrations of PM_{10} at Armagh Roadside on 26-27/06/2018, showing the difference between imputed (blue) and the observed concentrations (red).

the imputed DAQI, there is a sudden change in the pollutant concentrations for a few hours. An example is the 'Armagh Roadside' site on 26/06/2018, when the observed DAQI is 7 based on PM_{10} level at this day, while the imputed DAQI is 2.

Figure 7.22 shows the difference between imputed and observed hourly PM_{10} concentrations on 26-27/06/2018. It is clear that this peak is due to some sudden change that gives higher PM_{10} for one hour.

On the other hand, when the imputed DAQI is higher than the observed DAQI, we found that may be due to missing observations at this period. As an example at 'London Eltham' site on 03/03/2018, the imputed DAQI is 8 based on $PM_{2.5}$, while the observed DAQI is 1 calculated based on NO_2 and O_3. Our imputation shows that there is a high level of $PM_{2.5}$ on this day while the observations are missing. Our imputation for $PM_{2.5}$ at this site at other days are good and represent the trend very well, as shown in the representation of the daily average concentrations for imputed and observed $PM_{2.5}$ for the year 2018 at this site in Figure 7.23.

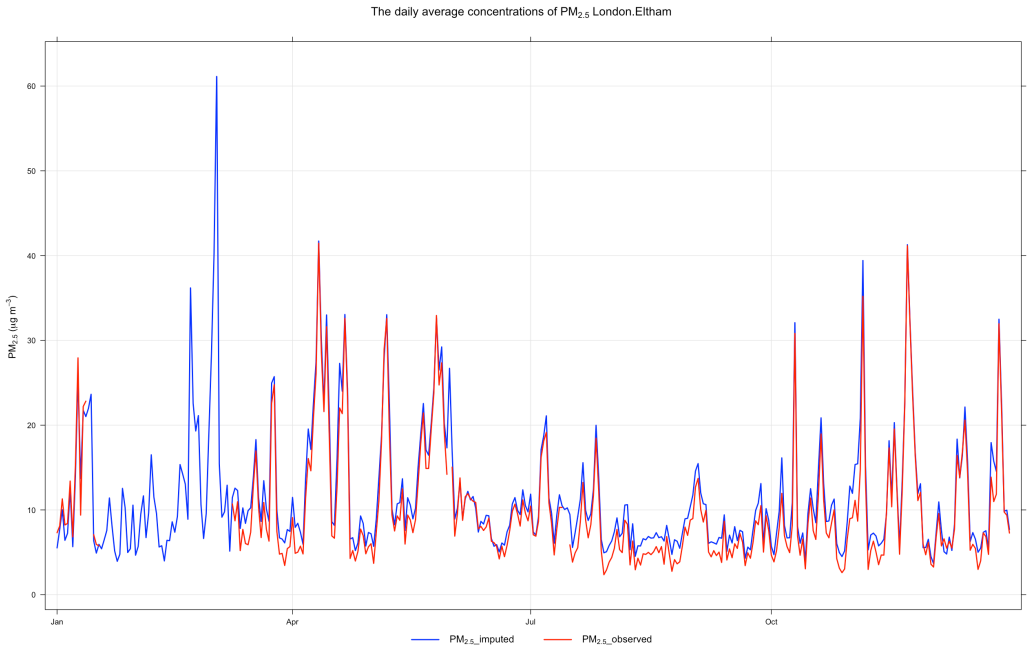


Figure 7.23: Hourly concentrations of $PM_{2.5}$ at London Eltham for year 2018, showing the difference between imputed (blue) and observed concentrations (red) with a period of missing observations on March.

7.5 Summary

In this chapter, we evaluated our proposed models to impute missing pollutants in a station based on statistical and graphical model evaluation functions (Taylor's diagrams and Conditional quantile plots), that are designed to evaluate air pollution modelling. We found that the best imputation models based on statistical analysis are model 8 (Median) for PM_{10}, and $PM_{2.5}$ and Model 9 (Median.ENV) for NO_2 and O_3 imputation. The advantage of these models are that they aggregate the spatial and temporal imputation. The spatial imputation is obtained from the nearest stations and the temporal imputation is obtained by MVTS clustering that clusters the stations based on similarity in time.

The graphical model evaluation functions showed selected models' performance based on the distribution of the concentrations and the degree of agreement between imputed/modelled and observed concentrations. These functions help us to understand the relationship between the distributions of the observations and the models'

performance, and analyse models' performance under each environment type.

Model 8 (Median) and model 9 (Median.ENV) are based on the median concentrations from stations with temporal and spatial similarity, so these models' expected performance is to underestimate the highest values and overestimate the lowest values with a normal distributed dataset. As confirmed in our analysis, the performance of these models is very good with a slight underestimation with PMs (PM_{10}, and $PM_{2.5}$) imputation using model 8 (Median), and with O_3 imputation using model 9 (Median.ENV), especially with high concentrations. At the opposite end, model 9 (Median.ENV) slightly overestimates the NO_2 concentrations, due to the regional behaviour of this pollutant. We found that these models' performance can vary based on the environmental type and the nature of the pollutant, as shown in the models' performance and the DAQI analysis.

Through our analysis, we also found that the variation of the model's performance with different environmental types is due to the pollutant behaviour and its emitted sources. However, these models work equally well under different weather types, as shown in the analysis of models' performance and Lamb Weather Types (LWTs).

Table 7.4: Number of days where imputed DAQI agrees/disagrees with observed DAQI, the asterisks represent cases where the disagrees difference between the imputed and the observed DAQI is more than 3 index values (13 cases).

Index Agreement							
Observed DAQI (band (index))	**Imputed DAQI (band (index))**	**Number of days**	**Percentage of Days**	**Observed DAQI (band (index))**	**Imputed DAQI (band (index))**	**Number of days**	**Percentage of Days**
Low (1)	Low (1)	18933	32.230	Moderate (5)	Moderate (5)	89	0.152
Low (2)	Low (2)	16632	28.313	Moderate (6)	Moderate (6)	17	0.029
Low (3)	Low (3)	7951	13.535	High (7)	High (7)	4	0.007
Moderate (4)	Moderate (4)	486	0.827	High (8)	High (8)	6	0.010
Total agreement							44,118
Total Percentage							75%
Index Disagreement							
Observed DAQI	**Imputed DAQI**	**Number of days**	**Percentage of Days**	**Observed DAQI**	**Imputed DAQI**	**Number of days**	**Percentage of Days**
*Low (1)	High (8)	1	0.00	Moderate (5)	Low (3)	68	0.12
Low (1)	Low (2)	1920	3.27	Moderate (5)	Moderate (4)	213	0.36
Low (1)	Low (3)	141	0.24	Moderate (5)	Moderate (6)	10	0.02
Low (1)	Moderate (4)	10	0.02	Moderate (6)	High (7)	3	0.00
*Low (1)	Moderate (5)	1	0.00	*Moderate (6)	Low (2)	2	0.00
*Low (1)	Moderate (6)	1	0.00	Moderate (6)	Low (3)	8	0.01
*Low (2)	High (8)	1	0.00	Moderate (6)	Moderate (4)	34	0.06
Low (2)	Low (1)	6082	10.35	Moderate (6)	Moderate (5)	42	0.07
Low (2)	Low (3)	1489	2.53	High (7)	High (8)	1	0.00
Low (2)	Moderate (4)	12	0.02	*High (7)	Low (2)	1	0.00
Low (2)	Moderate (5)	2	0.00	*High (7)	Low (3)	1	0.00
*Low (2)	Moderate (6)	1	0.00	High (7)	Moderate (4)	3	0.01
Low (3)	Low (1)	318	0.54	High (7)	Moderate (5)	11	0.02
Low (3)	Low (2)	3220	5.48	High (7)	Moderate (6)	10	0.02
Low (3)	Moderate (4)	232	0.39	High (8)	High (7)	3	0.00
Low (3)	Moderate (5)	11	0.02	High (8)	High (9)	1	0.00
Low (3)	Moderate (6)	3	0.01	*High (8)	Low (2)	1	0.00
Moderate (4)	Low (1)	8	0.01	*High (8)	Low (3)	1	0.00
Moderate (4)	Low (2)	46	0.08	High (8)	Moderate (5)	1	0.00
Moderate (4)	Low (3)	644	1.09	High (8)	Moderate (6)	1	0.00
Moderate (4)	Moderate (5)	52	0.09	High (9)	High (7)	1	0.00
Moderate (4)	Moderate (6)	2	0.00	High (9)	High (8)	1	0.00
Moderate (5)	High (7)	1	0.00	Very High (10)	High (7)	1	0.00
*Moderate (5)	Low (1)	1	0.00	Very High (10)	High (8)	1	0.00
Moderate (5)	Low (2)	6	0.01	*Very High (10)	Moderate (6)	1	0.00
Total disagreement							14,625
Total Percentage							25%

Chapter 8

Model Application

In the previous chapter 7, two imputation models were selected as the best imputation models among other proposed models based on modeling evaluation results: model 8 (Median) is the best imputation model for PM_{10} and $PM_{2.5}$ and model 9 (Median.ENV) is the best for NO_2 and O_3. The performance of these models was evaluated using various evaluation statistical and graphical functions.

As mentioned earlier, if we are going to impute pollutants that are not measured, we would lack 'ground truth' to measure how good the imputation is. For this reason, the model evaluation was based on the models' ability to reproduce/impute pollutants that are already measured at stations. In this chapter, we are going to apply the selected imputation models to impute the missing pollutants that are not measured at all.

We will attempt to measure how effective these imputations are for the assessment of air quality at stations with imputed (unmeasured) pollutants and provide a DAQI that is more realistic. As DAQI calculated from observed data only may give a false representation of the air quality, for example, if there were high concentrations of an air pollutant that was not being measured, the air quality may be worse than indicated by the DAQI. This analysis will be based on the comparison between DAQI calculated from observed data (i.e. requiring no imputation) and DAQI after imputing the missing (unmeasured) pollutants. The analysis will assess how

DAQIs calculated with and without the imputation of missing pollutants compare and how often and by how much do the DAQIs calculated without the imputation underestimate the severity of the air pollution. In this chapter, Section 8.1 includes model application results and analysis based on DAQI comparison; Section 8.2 includes further analyse of cases where there is a disagreement between the imputed and observed DAQI; and Section 8.3 includes a discussion of the obtained results. Finally, Section 8.4 is a summary of this chapter.

8.1 Model Application to Calculate DAQI Values

After imputing all missing (not measured) pollutants in all stations in our test dataset for the year 2018, we calculate the DAQI based on the observed pollutants and another DAQI based on observed+imputed pollutants in all stations. The model application includes 141 out of 167 stations that have one or more imputed pollutants. These stations together have 49,344 days are compared. As shown in Table 2.2, in our dataset, 26 stations measured all the four pollutants; hence as they do not include any imputed pollutants, they are excluded in this comparison.

Table 8.1 shows the agreement and disagreement between the observed DAQI, calculated based on the observations and imputed DAQI calculated based on observed in addition to imputed pollutants. The comparison will help us to understand the contribution that further measurements could make to air quality assessment, assuming the imputation is valid and reflective of potential true values.

The top table shows the number of days where the imputed DAQI agrees with observed DAQI, and the percentage of agreement at each index. In contrast, the bottom table represents the cases of disagreement, where the imputed and observed DAQI are different. From this table, we can see that the imputed DAQI agree with the observed DAQI 54% of the total number of days included in this comparison, while in 46% of the days there is a disagreement.

The highest percentage of agreement between imputed and observed DAQI is seen for index values 2 and 3 (31% and 16% of the total number of days respectively), representing the majority of air quality indices in this dataset (year 2018). Similarly in the disagreement cases, the highest percentage of disagreement occurred when the observed DAQI is 1 and the imputed DAQI is either 2 or 3. This disagreement represents 26% and 11% of the total number of days respectively. This is followed by 6% when the imputed DAQI is 3, and the observed DAQI is 2. Other cases are associated with a very small percentage that does not exceed 1% of the total number of days in the dataset. As can be appreciated from the bottom table, the imputed DAQI is higher than the observed DAQI with all cases.

8.2 Results Analysis

This section will analyse some cases of disagreement when there is a high variation between imputed and observed DAQI. This analysis focuses on cases where the imputed DAQI is higher by more than three index values. There are 102 cases/events, included at the bottom of Table 8.1 in rows marked with (*). After analysis, we found that these cases/events occur on 12 specific days at more than one location/station. In addition, these events are mainly caused by either imputed $PM_{2.5}$ or O_3. The total number of stations included in this analysis and correlated to these events is 57 stations, information about these stations is in Appendix H, Table H.1.

These 102 events include 54 events when the imputed DAQI is associated with higher imputed $PM_{2.5}$, 2 events when the imputed DAQI is associated with higher PM_{10}, and 46 events when the imputed DAQI is associated with higher imputed O_3. Table 8.2 shows the detail of these cases, including the date of the event, the pollutant that caused the higher DAQI, and the number of stations associated with each event.

Our selected imputation models (model 8 (Median) and model 9 (Median.ENV))

Table 8.1: Number of days where imputed DAQI agrees/disagrees with observed DAQI in model application, the asterisks represent cases where the imputed DAQI is more than 3 index values (102 cases).

Index Agreement							
Observed DAQI (band (index))	**Imputed DAQI (band (index))**	**Number of days**	**Percentage of Days**	**Observed DAQI (band (index))**	**Imputed DAQI (band (index))**	**Number of days**	**Percentage of Days**
Low (1)	Low (1)	1557	3.16	Moderate (6)	Moderate (6)	70	0.14
Low (2)	Low (2)	15744	31.91	High (7)	High (7)	19	0.04
Low (3)	Low (3)	8151	16.52	High (8)	High (8)	10	0.02
Moderate (4)	Moderate (4)	853	1.73	High (9)	High (9)	1	0.00
Moderate (5)	Moderate (5)	264	0.54	Very High (10)	Very High (10)	3	0.01
Total agreement							26,672
Total Percentage							54%
Index Disagreement							
Observed DAQI (band (index))	**Imputed DAQI (band (index))**	**Number of days**	**Percentage of Days**	**Observed DAQI (band (index))**	**Imputed DAQI (band (index))**	**Number of days**	**Percentage of Days**
*Low (1)	High (7)	5	0.01	*Low (2)	Moderate (6)	8	0.02
*Low (1)	High (8)	3	0.01	*Low (3)	High (7)	2	0.00
*Low (1)	High (9)	1	0.00	*Low (3)	High (8)	1	0.00
Low (1)	**Low (2)**	**13236**	**26.8**	Low (3)	Moderate (4)	61	0.12
Low (1)	**Low (3)**	**5598**	**11.3**	Low (3)	Moderate (5)	21	0.04
Low (1)	Moderate (4)	325	0.66	Low (3)	Moderate (6)	4	0.01
*Low (1)	Moderate (5)	50	0.10	Moderate (4)	High (7)	1	0.00
*Low (1)	Moderate (6)	20	0.04	Moderate (4)	Moderate (5)	6	0.01
*Low (2)	High (7)	1	0.00	Moderate (4)	Moderate (6)	4	0.01
*Low (2)	High (8)	10	0.02	Moderate (5)	High (7)	1	0.00
*Low (2)	High (9)	1	0.00	Moderate (5)	Moderate (6)	4	0.01
Low (2)	**Low (3)**	**3002**	**6.1**	Moderate (6)	High (8)	2	0.00
Low (2)	Moderate (4)	265	0.54	High (7)	High (8)	1	0.00
Low (2)	Moderate (5)	39	0.08				
Total disagreement							22,672
Total Percentage							46%

Table 8.2: Days analysis where there is high disagreement between imputed and observed DAQI.

Date	Main pollutant associated with DAQI	Number of stations
03/03/2018	$PM_{2.5}$	32
03/03/2018	PM_{10}	2
04/03/2018	$PM_{2.5}$	5
06/05/2018	O_3	2
07/05/2018	O_3	23
26/06/2018	O_3	5
29/06/2018	O_3	1
01/07/2018	O_3	3
02/07/2018	O_3	4
03/07/2018	O_3	4
05/07/2018	O_3	1
26/07/2018	O_3	3
05/11/2018	$PM_{2.5}$	17
Total events		102

are based on the median imputed value from all proposed imputation models, as previously explained in Chapter 3. These models use the spatial and temporal similarity between stations in the imputation. These models underestimate the high concentrations and overestimate the low concentrations, as shown in model evaluation in the previous chapter, in Section 7.2. This is also confirmed by the comparison between the imputed and the observed DAQI, as the analysis showed that the imputed DAQI is lower than the observed DAQI in 73% of the total disagreement cases.

In imputation models that are based on the clustering average (model 1 (CA), model 2 (CA+ENV), and model 3 (CA+REG)), if a station within a cluster has some episodes (peaks) of high concentrations of any pollutant, the clustering imputation models will reproduce these episodes in the imputation. Hence, the cluster centroid represents its own cluster and illustrates the pollutant behaviour for all stations within the cluster, any high pollutant episodes (peaks) should be captured by the clusters centroid. Transferring these high peaks to the imputation using these models, will give an estimation of pollutant that are missing in some stations within those clusters.

To further analyse these cases, we study the association between stations where

these cases/events happened and observed pollutants concentrations within the clusters for pollutants that caused these events ($PM_{2.5}$, PM_{10}, and O_3). Then, we compare the time and the location of these events with Defra's report about air pollution in 2018, [3]), in which any high pollution episode recorded by AURN stations is reported every year.

In the next sections we analyse these 12 days based on the observations of each pollutant at each cluster using the clustering average (cluster centroid) and the cluster hourly pollutant concentrations representations. The geographical distribution of these clusters was previously shown in Chapter 6.

8.2.1 Days associated with high imputed $PM_{2.5}$:

From Table 8.2, there are 3 days associated with 54 events caused by high imputed $PM_{2.5}$ on 03/03/2018, 04/03/2018, and 05/11/2018. The number of stations associated with higher imputed DAQI on these days are 43 stations: 46% of these stations belong to cluster 4 (purple) located in the South West, 49% belong to cluster 2 (green) located in the centre of the UK, and the rest (2 stations) belong to cluster 3 (blue) located in the North. Figure 8.1 shows the time variation for each cluster centroid $PM_{2.5}$ concentrations for the whole year 2018, and this is then used within the imputation process. Cluster 2 and 4, with the highest numbers of stations affected by higher imputed values, have the highest concentrations of the pollutant among all other clusters, and have more peaks than other cluster centroids as shown in Figure 8.2, that shows the pattern of hourly concentrations of $PM_{2.5}$ in each cluster centroid for the whole year.

Now, to focus on the high imputation days we look at the concentrations on the key months, and show how the clusters centroid captured those high $PM_{2.5}$ values. Those are then transferred to the imputation at stations within these clusters. Figure 8.3 (top plot) shows the hourly concentrations from the 1st-11th of March only at the centroid of cluster 2 (represented in red) and the modelled $PM_{2.5}$ values

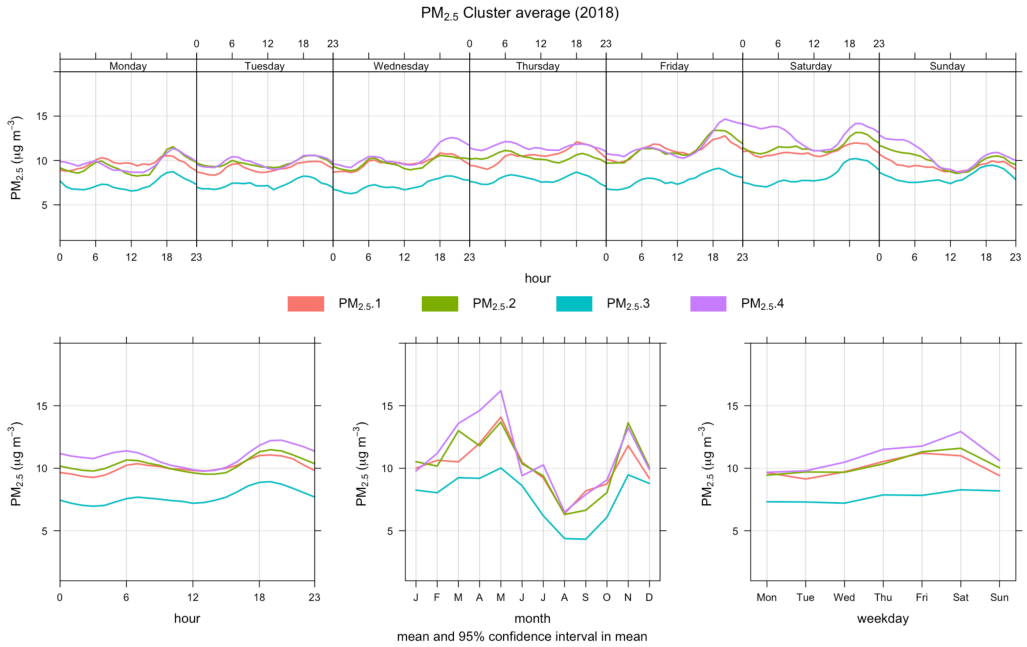


Figure 8.1: Time variation of observed $PM_{2.5}$ for each cluster centroids for the year of 2018.

that are imputed by model 8 (Median) at 14 particular stations (represented in blue) within cluster 2. This shows how the centroid captured the high $PM_{2.5}$ values early in March (03/03/2018 and 04/03/2018), which are then transferred to the imputation within these clusters. Similarly, Figure 8.3 (bottom plot) shows the centroid of cluster 4 (in red) compared to the modelled values using model 8 at 21 particular stations (in blue) within cluster 4 and again we see how the peaks are captured and transferred.

The second peak that causes higher imputed DAQI early on November (05/11/2018) at 17 stations, is also captured by the centroids of cluster 2 and 3. Figure 8.4 (top figure) shows how cluster 2 centroid captures this event (represented in red) and the modelled $PM_{2.5}$ values generated by model 8 at 15 stations (represented in blue) have reflected this event. However, despite the centroid of cluster 3 having the lowest $PM_{2.5}$ concentrations, a small peak is captured by the centroid at this day as shown in Figure 8.4, (bottom figure) and that is then transferred to the imputation at two stations in this cluster.

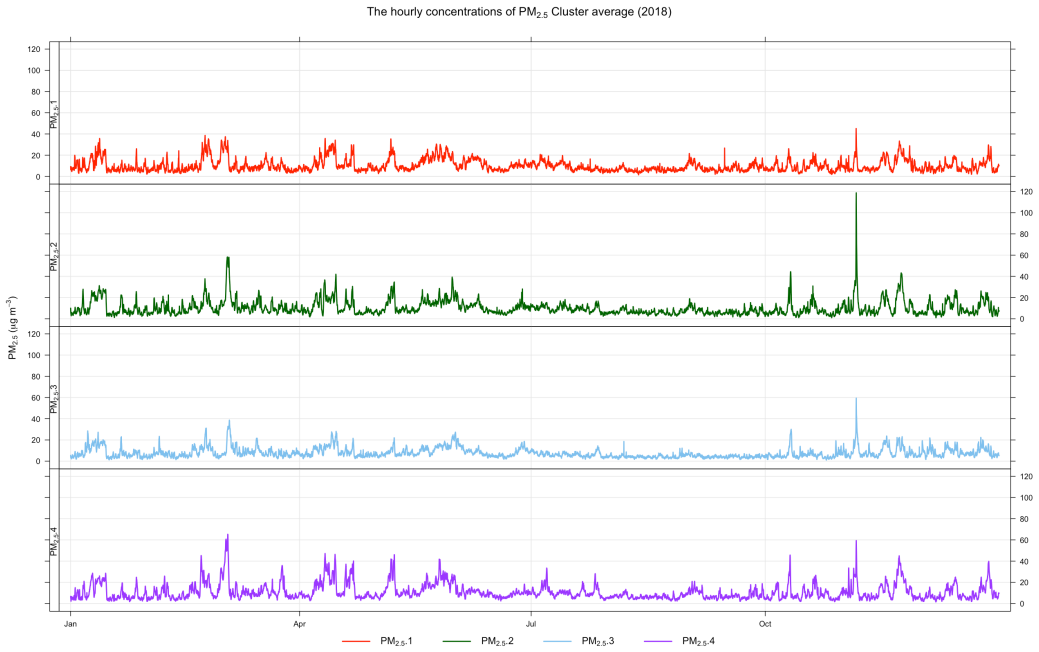


Figure 8.2: Hourly concentrations of observed $PM_{2.5}$ for each cluster for the year of 2018.

8.2.2 Days associated with high imputed PM_{10}:

There is one day (03/03/2018) associated with two events of high imputed DAQI caused by high imputed PM_{10} concentrations. These events correspond to two stations belonging to cluster 4. Figure 8.5, shows that cluster 4 (purple) is associated with the highest PM_{10} concentrations. So, using the cluster average imputation to impute PM_{10} for any stations within that cluster (cluster 4) will reproduce these high PM_{10} concentrations. Figure 8.6 shows the high event of observed PM_{10} in day (03/03/2018) is captured by cluster 4 centroid, and is also transferred to the imputation of two stations within the cluster.

To validate our findings we have looked for specific air pollution events that were reported in 2018. Defra report on Air Pollution in the UK 2018 [3], focuses on events on the year of interest. There were some significant episodes of high particulate pollution recorded early in March and November, and at the end of June of 2018. These episodes of high particulate concentrations ($PM_{2.5}$ and PM_{10}) have different causes such as cold weather (the Beast from the East), wildfires in the moors in

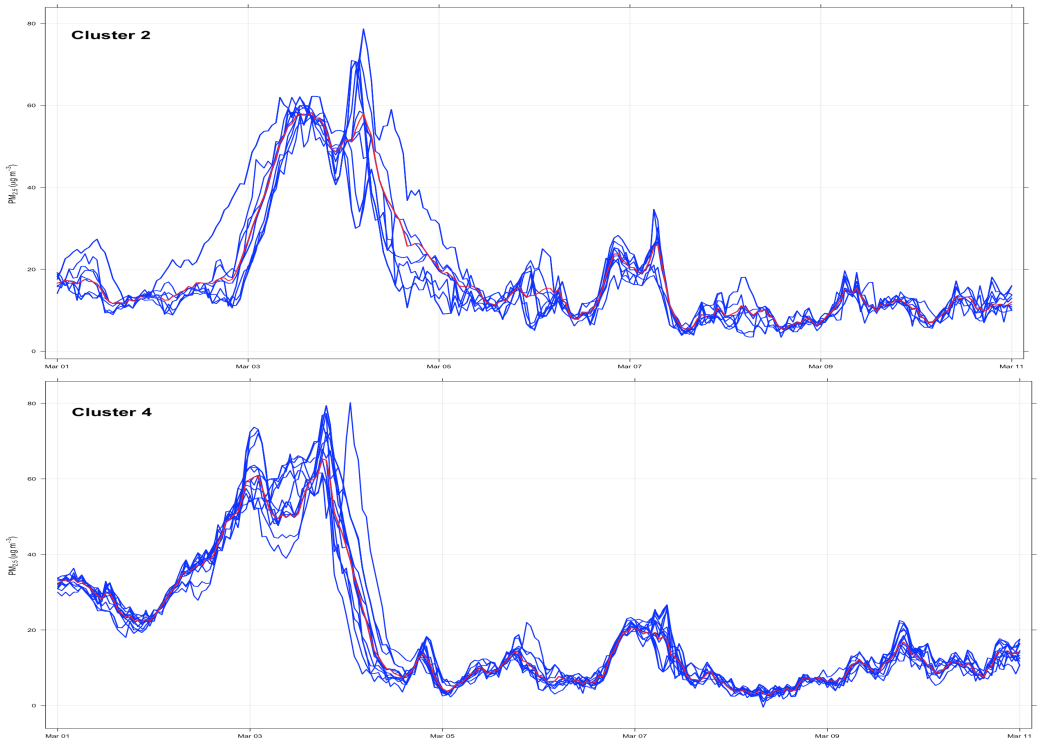


Figure 8.3: Hourly concentrations of observed $PM_{2.5}$ (red) for cluster 2 centroid (top plot) and cluster 4 centroid (bottom plot) from 1st to 11th of March 2018 compared to the modelled $PM_{2.5}$ concentrations (blue) at 14 stations in cluster 2 (top plot) and 21 stations in cluster 4 (bottom plot).

the North West of England, and some events like Bonfire Night.

On the 3rd and the 4th of March 2018, there were two episodes of high $PM_{2.5}$ and PM_{10} associated with a period of extreme cold weather (ice and snow). Particulate matter concentrations increased during these days in some areas of the UK. The highest pollutants episodes were recorded in London, the South East of England, the Midlands, and extended through the rest of England and South Wales (shown in Figure 6-1 in the report, page 101 [3]) and these corroborate our findings.

As shown in our analysis, of the 3rd and 4th March the centroids of cluster 2 and 4 captured these events and the imputation model transfer them to other stations where $PM_{2.5}$ and PM_{10} are not measure. If we compare the time and locations (areas) where high pollutants are recorded (based on the report) with locations of those clusters, we will find that these clusters cover the affected areas. These

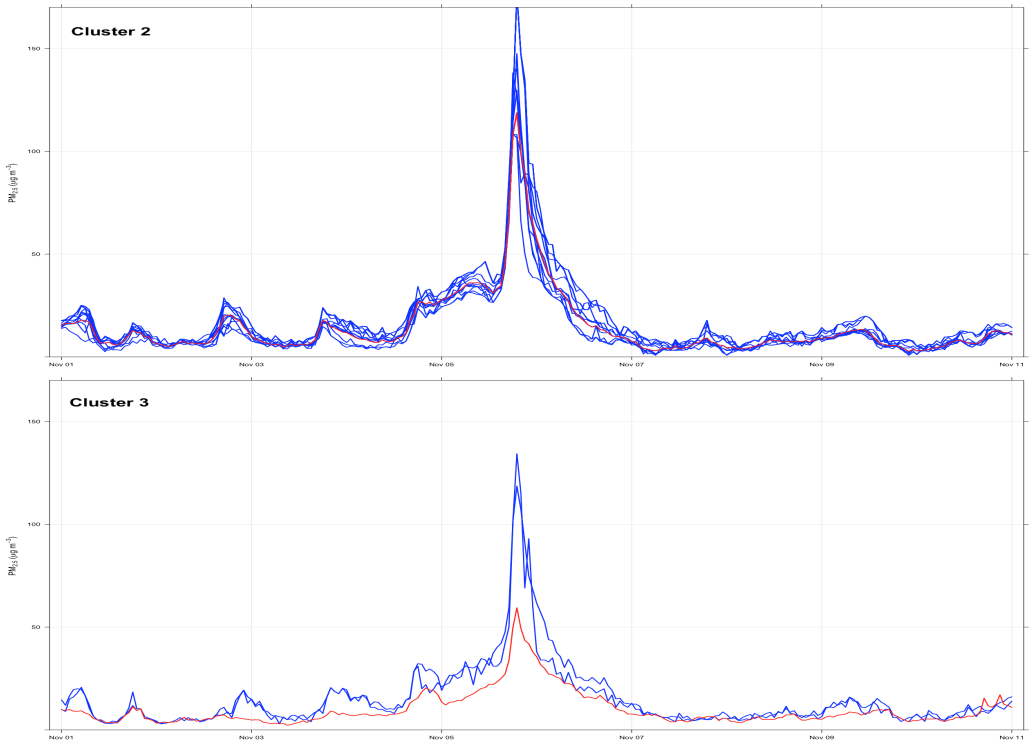


Figure 8.4: Hourly concentrations of observed $PM_{2.5}$ (red) for cluster 2 centroid (top) and cluster 3 centroid (bottom) from 1st to 11th of November 2018 and the modelled $PM_{2.5}$ concentrations (blue) at 15 stations in top plot and 2 stations in bottom plot.

episodes are captured by cluster 2 (green) located in the centre of the UK, covers multiple regions that are West Midlands, East Midlands, Yorkshire & Humberside, and North West and Merseyside. While cluster 4 (purple) located in the South West, covers Eastern, Greater London, and South East.

The second high episode of high particulate matter concentrations was caused by Bonfire Night on 5th November 2018. Pollution was recorded for the majority of the South of England stations including North East, Yorkshire & Humberside, North West and Merseyside, and East Midlands (shown in Figure 6-8, page 108 in the report [3]). In our analysis, this episode was captured by centroids of clusters 2 and 3, then reproduced mainly by the imputation in 15 stations at cluster 2. Two stations only from cluster 3 reproduced this episode, these stations are located in the North East of England.

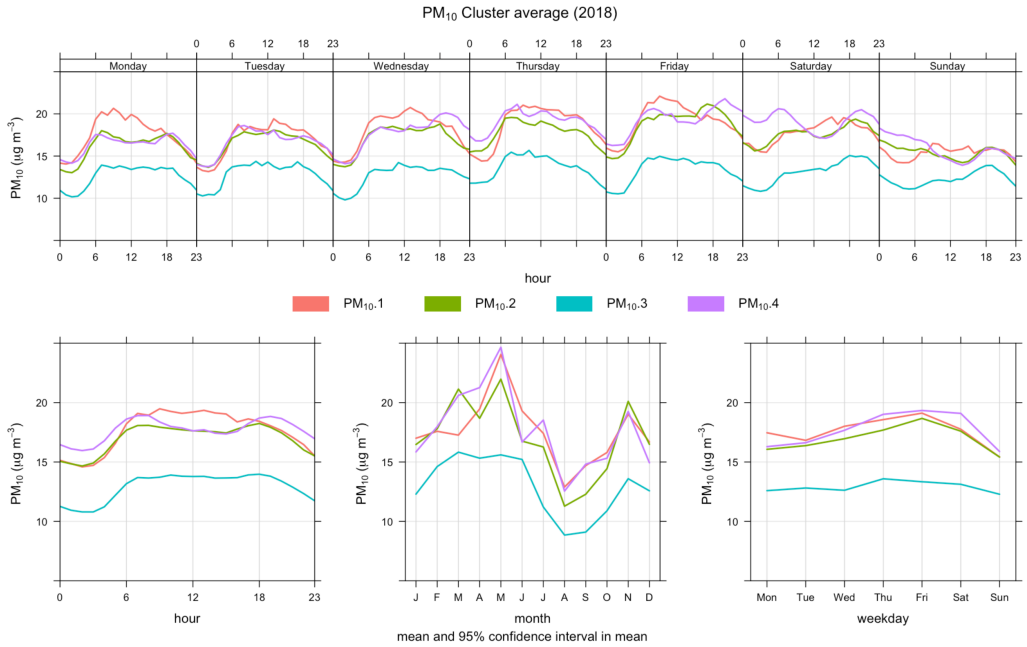


Figure 8.5: Time variation of observed PM_{10} for each cluster centroids for the year of 2018.

The coincidence of the times and the locations of particulate matter episodes/events captured by the cluster centroids with those reported by Defra, confirmed that there were several high pollutant episodes in those areas where pollutants were not measured.

8.2.3 Days associated with high imputed O_3:

There are nine days distributed in May, June, and July which are associated with 46 events of higher imputed DAQI caused by high imputed O_3 in 27 different stations. These stations are spread across all clusters: 25% belong to cluster 1, 44% belong to cluster 2, 7% belong to clusters 3, and 19% belong to cluster 4.

Figure 8.7, shows the time variation of O_3 concentrations of the cluster centroids. It can be observed that clusters 1 and 3 have higher O_3 concentrations, while clusters 2 and 4 have lower concentrations. Figure 8.8, shows the hourly concentrations of O_3 at each cluster, and clearly, we can see some high events at these months that are captured by the clusters centroids. These events are as follows:

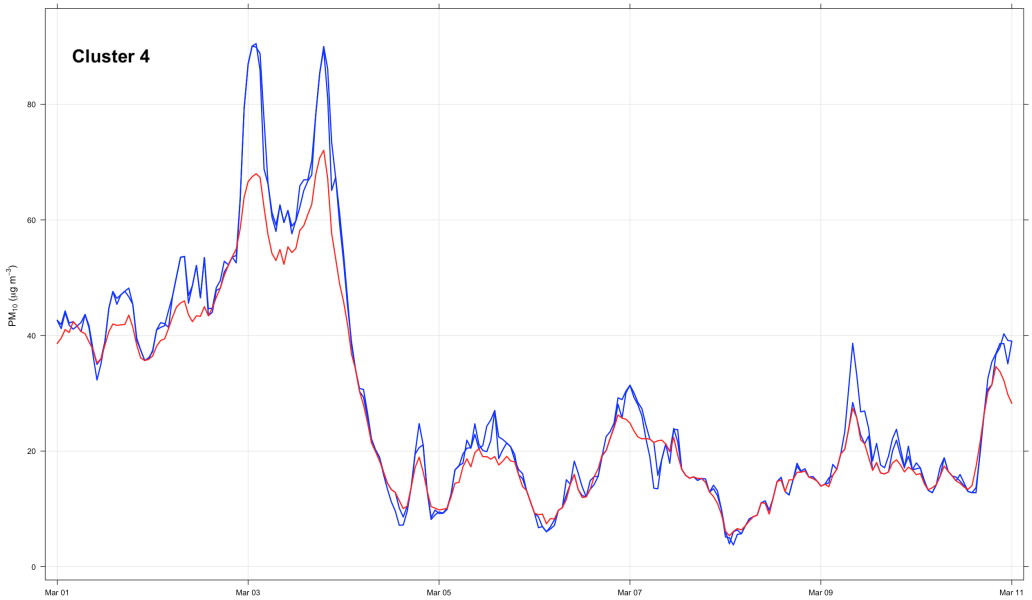


Figure 8.6: Hourly concentrations of observed PM_{10} for cluster 4 centroid from 1st to 11th of March 2018 (red) and the modelled PM_{10} concentrations in 2 stations (blue).

1. Two events of high observed O_3 early on May (6-7/May/2018) that are captured by all four cluster centroids and transferred to the imputation of O_3 at 23 stations using model 9. There are 6 stations that belong to cluster 1, 12 stations that belong to cluster 2, 2 stations that belong to cluster 3, and 3 stations that belong to cluster 4.

   Figure 8.9, show the imputation of O_3 at these stations compared to the concentrations of observed O_3 at clusters centroid that the stations belong to. From these figures, we can see that all cluster centroids captured these events (on day 6-7/May/2018) and the modelled O_3 values generated by model 9 reproduce these events at different stations within these clusters.

2. Two events of high observed O_3 at the end of June (26 and 29/06/2018), are captured by the centroids of cluster 1 and 2. These events are transferred by the imputation of O_3 at 3 stations within cluster 1 and 2 stations within cluster 2. Figure 8.10 shows these events at the centroids of cluster 1 and 2 and the modelled O_3 concentrations at stations where imputed DAQI is

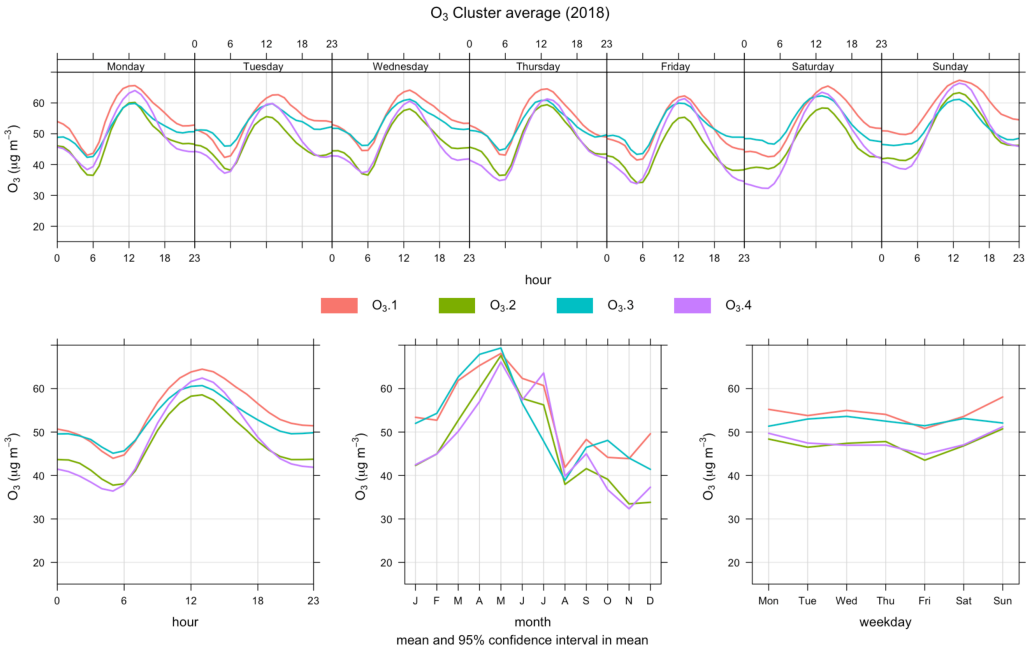


Figure 8.7: Time variation of observed O_3 for each cluster centroids for the year of 2018.

higher.

3. Four events of high observed O_3 early in July (1, 2, 3 and 5/07/2018) are captured by the centroids of cluster 1, 2, and 4. These events are reproduced in the O_3 imputation at some stations within these clusters. Figure 8.11, shows these events in the cluster centroids and then in the imputation. This high events can be observed in cluster 1 (first plot) more than for the other clusters. These events are captured in different clusters centroid as follow:

   - The centroid of cluster 1 captured two of these events (02-03/07/2018) and those are then reproduced by the imputation at 4 stations in this cluster as shown in the first plot.
   - The centroid of cluster 2 captured two events on (01-05/07/2018) that are then reproduced by the imputation at two stations in this cluster as shown in the second plot.
   - The centroid of cluster 4 captured one event on (01/07/2018) that is reproduced by the imputation at four stations in this cluster as shown

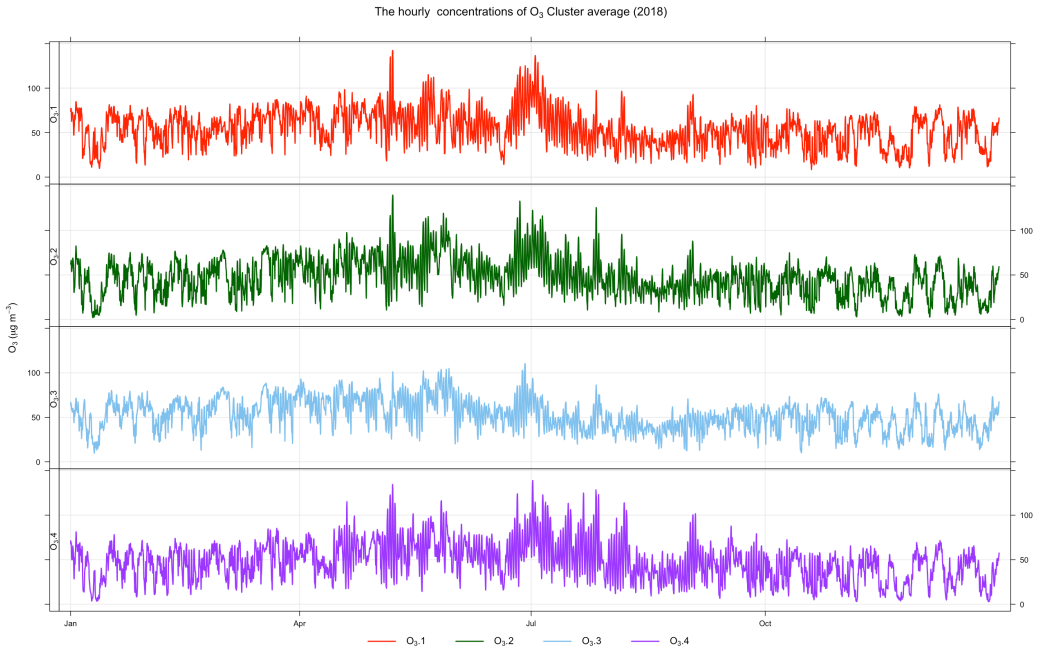


Figure 8.8: Hourly concentrations of observed O_3 for each cluster centroids for the year of 2018.

in the third plot.

4. Finally, one event at the end of July (26/07/2018), is captured by the centroid of cluster 4, as shown in Figure 8.12. These events are reproduced by O_3 imputation at 3 stations in this cluster.

Most of these disagreement cases (102 case) are associated with stations at cluster 2, which has the lowest O_3 concentrations, followed by cluster 1, which is the cluster that has the highest concentrations. However, these clusters are very similar in their peaks (high ozone episodes), as shown in Figure 8.8. Using the clustering imputation will reproduce these events in the imputation of any station within these clusters; that is why most of the stations within these 57 stations reproduce the events included in the analysis belonging to those clusters. Cluster 4 follows a similar pattern but with different wider amplitudes of these events.

As reported by Defra [3], there were several ozone threshold exceedances episodes recorded on the following dates: 5th May, 26th June, 1st July, 2 nd July, 26th July

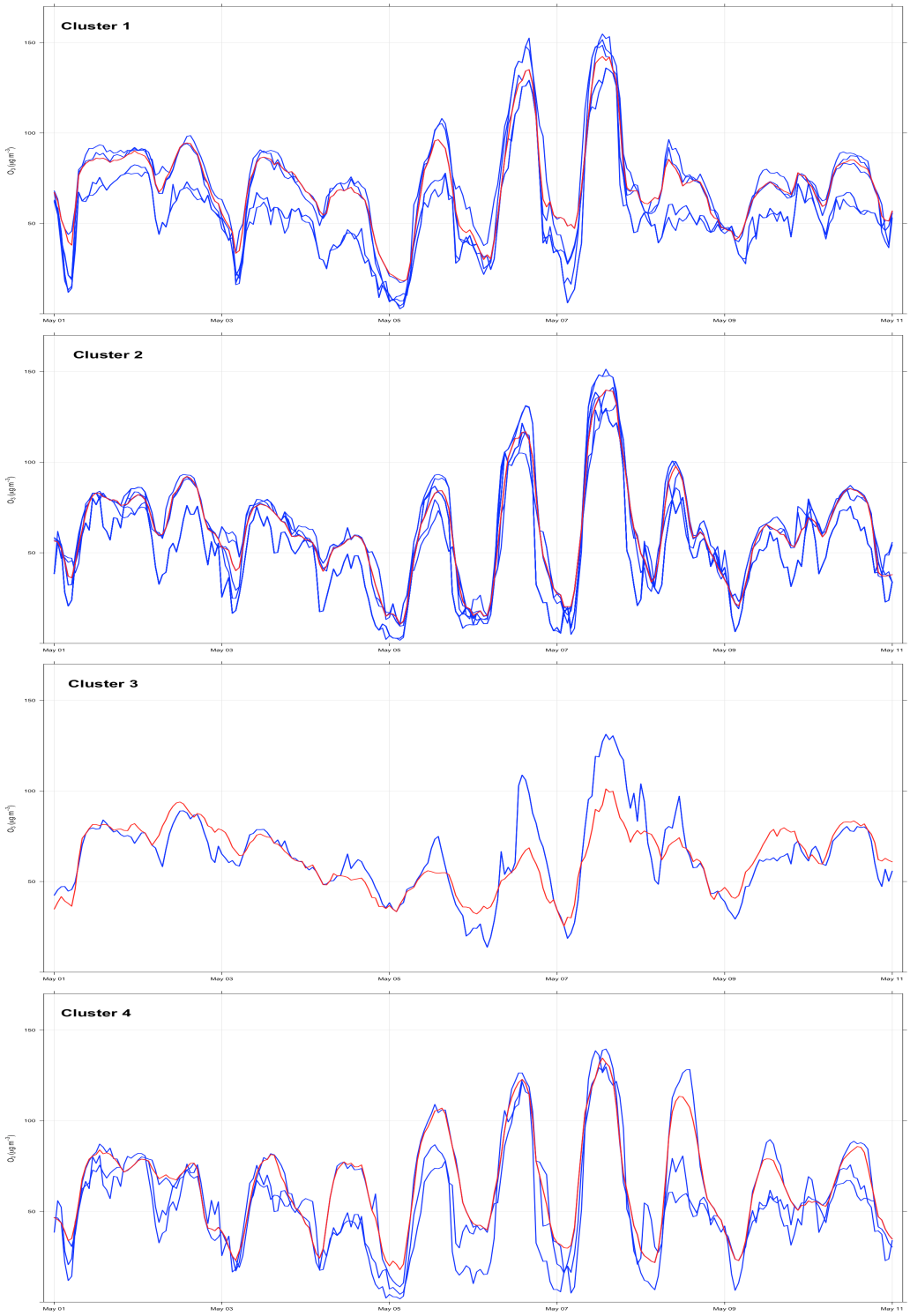


Figure 8.9: Hourly concentrations of observed O_3 for all four clusters' centroid in order from top to bottom from 1st to 11th of May 2018 (red) and the modelled O_3 concentrations in different stations (blue) in these clusters.

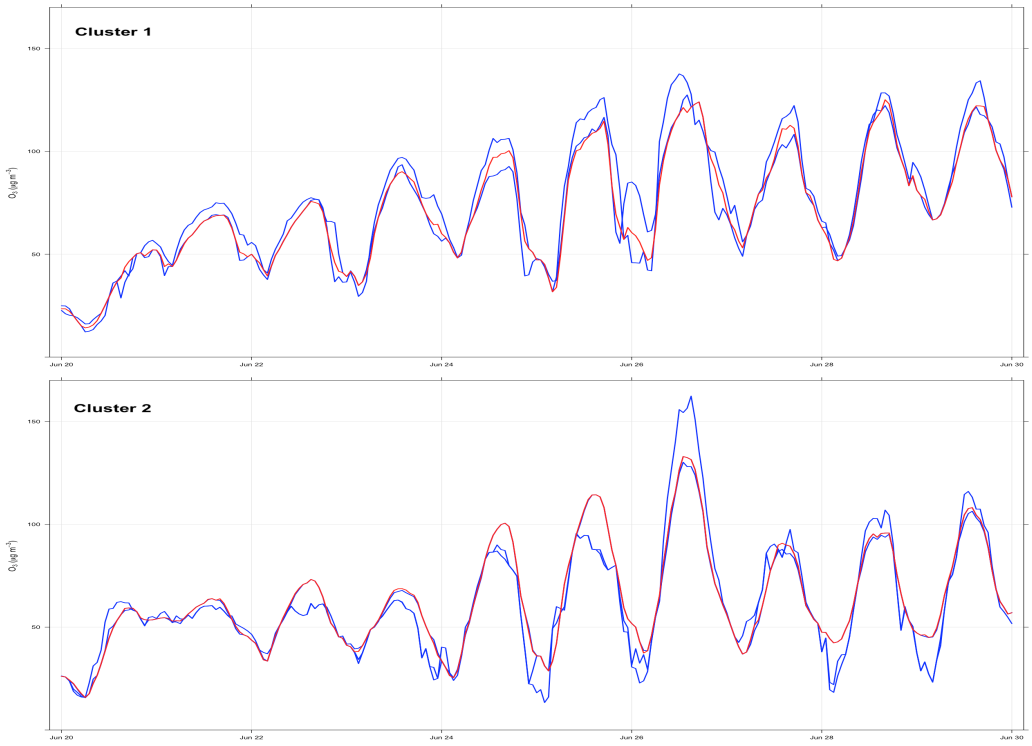


Figure 8.10: Hourly concentrations of observed O_3 (red) for cluster 1 centroid (top) and cluster 2 centroid (bottom) from 20th to 30th of June 2018 and the modelled O_3 concentrations (blue) at 2 stations in top plot and 3 stations in bottom plot in these clusters.

and 27th July, those episode can be in the late afternoon or early evening of those dates.

In our analysis, some clusters' centroids captured these high ozone episodes, which were then reproduced by ozone imputation at different stations in these clusters causing higher imputed DAQI. Now, we will compare the time and the location of these episodes with those included in our analysis.

Early on May, all clusters centroid captured two episodes on the 5th and 6th of May, which indicate higher ozone concentrations on these days across the UK.

A heatwave in June and July contributed to ozone pollution episodes and this also coincided with the moorland wildfires. These episodes occurred at the end of June, and again at the end of July (shown in Figure 6-11 in the report [3]). These pollution episodes were measured in North West and Merseyside, the South East

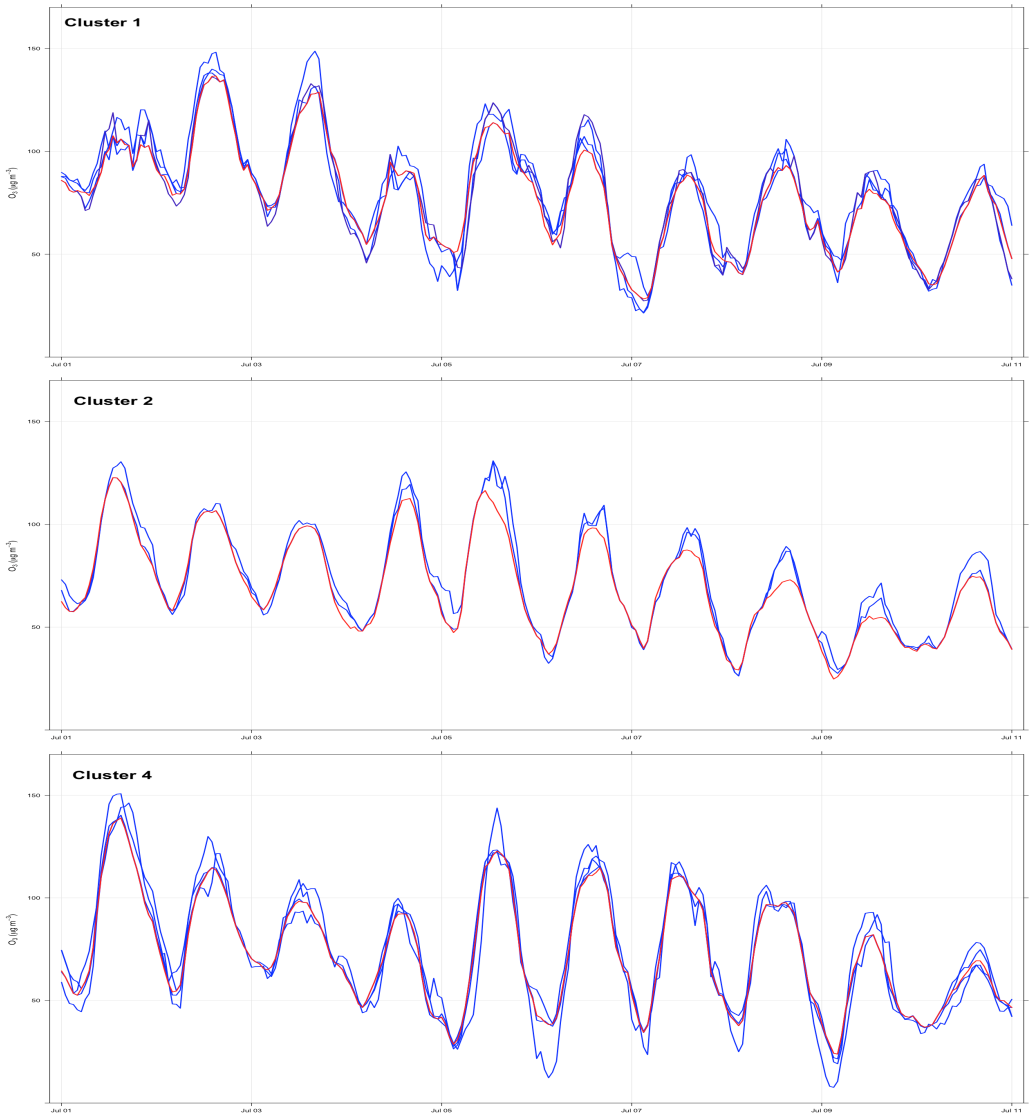


Figure 8.11: Hourly concentrations of observed O_3 (red) for cluster 1, 2, and 4 centroids (top to bottom) from 1st to 11th of July 2018 and the modelled O_3 concentrations (blue) at some stations in these cluster.

and South West.

The centroids of cluster 1 (red) located in South West, including some stations of the South West and cluster 2 (green) that covers most of the regions at the center of the UK, captured high ozone episodes at the end of Jun (26 and 29/6/2018). These clusters cover the affected areas where ozone high episodes were reported (Shown in Figure 6-11 in the report [3]). Similarly, early in July, North West and Merseyside, the South East and South West were affected by high ozone episodes.

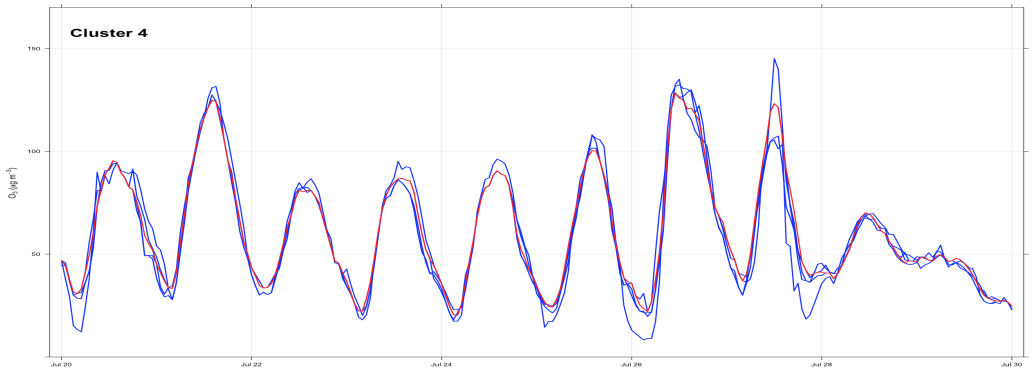


Figure 8.12: Hourly concentrations of observed O_3 (red) for cluster 4 centroid from 20th to 30th of July 2018 and the modelled O_3 concentrations (blue) at 3 stations in this clusters.

These episodes are captured by centroids of clusters covering those areas, that is clusters 1,2 and 4.

On 26th July, the worst pollution was experienced in the Eastern and in the South East regions (shown in Figure 6-13 in the report [3]. This episode was also capture by centroid of cluster 4 (purple) that covers these areas.

In conclusion, these high episodes (102 events) can be seen and were measured in the real data and then captured by the clusters centroid, as shown in the hourly representation of different pollutant concentrations at the cluster centroids. These events are transferred by the imputation to stations within the clusters where they possibly happened but were not measured. Our imputation models is able to reproduce these events in the missing pollutant imputation in stations that are temporally and spatially similar to stations used in the imputation (stations with high episodes). Reproducing these events in the imputation gives higher DAQI indices which may be reflective of true events that were missed at particular stations because of lack of measurements of some pollutants. We cannot verify this entirely as we do not have the real values that would have been measured, but this analysis and corroboration with the Defra reported events has allowed us to understand what changes may be seen in DAQI values if more measurements were taken and what kind of episodes currently missed could be captured.

8.3 Discussion

As shown by the previous analyses, $PM_{2.5}$ and O_3 were associated with the majority of the events when the imputed DAQI is higher than the observed DAQI. This could due to weather effect on those pollutants and other environmental factors (e.g. wild fires and Bonfire Night) which becomes more salient than NO_2. Also, as specified in 2018 air pollution report [3], most of the high pollution were episodes caused by $PM_{2.5}$ and O_3, while NO_2 met the limit value for hourly mean in 41 out of 43 zones in the UK.

For further analysis, to identify the pollutant(s) that has the highest individual DAQI and so determine the overall DAQI, we analyse data at sites where all pollutants are measured (i.e. includes 26 stations out of 167 station in our dataset). This pollutant(s) will have the most influence over the DAQI and can be considered the most important to measure. Table 8.3 includes DAQI analysis for 9,399 days from 26 stations for year 2018. It shows the main pollutant(s) having the highest individual DAQI, number of days associated with that pollutant(s) and the percentage of these days. As shown in the table in most cases O_3 followed by PM determine the overall observed DAQI in these stations.

O_3 determines the overall DAQI in 69% (i.e. 6,501 days) of the number of days in the dataset included in this analysis, followed by 5% (i.e. 479 days) for high O_3 combined with high $PM_{2.5}$, and PM_{10}. Multiple pollutants in this table signify that all pollutants have the same index value.

Similarly with imputed data, we analyse the main pollutant(s) that associate with the highest individual DAQI with all disagreement cases (i.e. 22,672 case) to see what imputed pollutant(s) increases the DAQI. This analysis will show where the weaknesses are in the current measurement and therefore observed DAQI.

Table 8.4 shows that O_3 followed by $PM_{2.5}$ and PM_{10} have the largest percentage compared to NO_2. O_3 is associated with 85% of those cases, while $PM_{2.5}$ and

Table 8.3: Pollutant with the highest individual DAQI from 26 stations that measure all pollutants and the percentage of each pollutant based on the total number of days.

Pollutant with the highest individual DAQI	**Number of days**	**Percentage**
O_3	6,501	**68.96**
O_3, $PM_{2.5}$, PM_{10}	479	**5.08**
NO_2, O_3	436	4.63
O_3, PM_{10}	334	3.54
$PM_{2.5}$, PM_{10}	307	3.26
NO_2, O_3, $PM_{2.5}$, PM_{10}	296	3.14
$PM_{2.5}$	240	2.55
NO_2	233	2.47
NO_2, $PM_{2.5}$, PM_{10}	179	1.90
O_3, $PM_{2.5}$	135	1.43
NO_2, O_3, PM_{10}	90	0.95
PM_{10}	73	0.77
NO_2, O_3, $PM_{2.5}$	36	0.38
NO_2, PM_{10}	31	0.33
NO_2, $PM_{2.5}$	29	0.31
Total days	9,399	

PM_{10} are associated with 4% only. However, $PM_{2.5}$ has more influence on the DAQI compared to PM_{10} in about 2% of those cases.

This indicates that O_3 and $PM_{2.5}$ have most influence over the observed DAQI and imputation may result in higher DAQI at stations where they are not measured, where there are actual high episodes of those pollutants within the clusters (in stations with similar behaviour and correlated geographical location).

The identification of true episodes of high pollution associated with imputed pollutants (unmeasured) that were missed at particular stations shows where there are particular weaknesses of the measurement of both O_3 and $PM_{2.5}$ in the monitoring network. Accordingly, measuring or imputing those pollutants might help to give better indication of the air quality. Hence, it is important to measure O_3 and $PM_{2.5}$ in all sites or use our proposed imputation model as a substitute.

Table 8.4: Pollutant with the highest individual DAQI with disagreement cases between the imputed and the observed DAQI and the percentage of each pollutant based on the total number of disagreement cases.

Pollutant with the highest individual DAQI	**Number of days**	**Percentage**
O_3	19,278	**85.03**
$PM_{2.5}$, PM_{10}	826	**3.64**
O_3, $PM_{2.5}$, PM_{10}	755	3.33
O_3, PM_{10}	729	3.22
$PM_{2.5}$	609	2.69
PM_{10}	235	1.04
O_3, $PM_{2.5}$	159	0.70
NO_2, O_3	56	0.25
NO_2	12	0.05
NO_2, O_3, PM_{10}	10	0.04
NO_2, PM_{10}	3	0.01
Total disagreement	22,672	

8.4 Summary

In this chapter, we show how the selected imputation models (model 8 and model 9) could be used for the assessment of air quality at stations where pollutants are not measured. The evaluation of this application process is based on the comparison between DAQI calculated from observed data (i.e. requiring no imputation) and DAQI after imputing the missing pollutants. Our analysis results showed that imputed DAQI agrees with observed DAQI for more than 50% of the total number of days included in this comparison. This means for those days, additional measurements may not have changed the air quality assessment. However, there is about 46% of disagreement, where imputation results in a higher DAQI value, i.e higher air pollution indices. The majority of index disagreement cases (that represents 44% of the total disagreement) happened when the imputed and observed DAQI fall within the low air quality band (when DAQI's value is between 1 and 3) and disagreement is proportionally low, and hence from a point of view of air pollution may not be a marked difference. Those represents the majority of air quality indices in this dataset (year 2018).

For those days when we saw a higher disagreement between imputed and observed

DAQI, which could be the more problematic, we produced detailed analysis. We found that there were high episodes in the real data and those were captured by the clusters centroids, and would therefore be transferred by the imputation to stations within the clusters. These episodes are transferred mostly by the imputation of $PM_{2.5}$ and O_3 compared to other pollutants. The imputed values which capture those episodes then lead to higher DAQI values which may have been missed. We found episodes that were missed in some stations in March and November of 2018 for $PM_{2.5}$, March of 2018 for PM_{10} and in May/June/July 2018 for O_3. We found from an independent Defra report that our episodes, captured in the imputation were in fact reported and were attributable to different events (e.g. weather, wild fires and Bonfire Night). Our analysis helps us to infer where more measurements for specific pollutants may be beneficial to truly capture pollution episodes.

Chapter 9

Conclusion and Further Work

This chapter presents the conclusion of this research with a discussion and recommended future work to develop the proposed approach.

9.1 Conclusion

In this research, we investigated the problem of missing values in air quality data. The main goal was to reduce the uncertainty of the air quality assessment resulting from missing measurements which may be missing either partially or completely. Enhancing the air quality assessments will help capture any pollution events that may be missed because of inadequate or insufficient measurements.

Our approach is based on time series clustering. We noted in the literature that most existing work in the field of air quality applied data mining and time series analysis approaches to investigate the correlation between air quality and other factors. In addition to other limitations such as dealing with missing data and using a univariate TS (single pollutant) rather than multiple pollutants. In contrast, in our research we have developed two approaches of time series clustering (i.e univariate and multivariate time series clustering), where we cluster stations based on measured pollutants concentration to inform the imputation. For this, we collected the air pollutants concentrations of the main four pollutants included

in the air quality assessment (O_3, NO_2, PM_{10}, and $PM_{2.5}$). The main challenge is that we do not have any previous knowledge about the correlation between these stations or how the pollutants behave in a location when their concentrations are missing or not measured.

Through the research, we applied the multivariate time series clustering using the k-means algorithm with the fused distance that is based on SBD. This proposed clustering approach enabled us to understand the correlation between multiple pollutants and to cluster stations based on measured pollutants only. Then we proposed and applied nine imputation models (i.e in Chapter 3.4.1) to impute unmeasured pollutants concentrations in stations where there are missing. In these models we aggregated the temporal similarity that are derived by the clustering and the spatial similarity that derived by the geographical distance between stations. The advantages of the imputation models that they are used to impute the whole time series (i.e pollutant concentrations) for whole year (2018). The clustering and imputation approaches together helped us to produce an enhanced version of the air pollution dataset to identify where more measurements for some pollutants may be beneficial.

To set up the experiments to achieve the main goal of this thesis, we used the hourly pollutant concentrations for four years (2015-2017) and included 167 stations around the UK belonging to six station environment types to build our models. Then, we applied our proposed models to cluster and impute all missing (unmeasured) pollutants for the following year (2018) in all stations to calculate a new air quality index (DAQI) that is more realistic and based on all contributed pollutants (i.e we called the imputed DAQI). Finally, we analysed any variation between the observed and imputed DAQI to determine where additional measurements may enhance air quality assessment.

At the end of this thesis, we consider our proposed approach achieved the research aim and objectives and answer all the research questions as stated in Chapter 1. For more details, Section 9.2, draws insights from the analysis of our results presented

in the experimental Chapters (Chapters 4 to 8) that answer the research questions addressed in Chapter 1 and Section 9.3 suggests some areas to develop the work in the future.

9.2 Discussion

In Chapter 1, we addressed eight research questions around the aim of clustering stations based on their similarity and imputing the unmeasured pollutants.

In this section, we evaluated how the experimental results have enabled us to answer our research questions. As we discussed in Chapters 5 and 6, through our experiments, we found that time series clustering either univariate or multivariate time series (MVTS) help us to discover more knowledge about the pollutants behaviours and inform the imputation by aggregating the stations based on their temporal similarity. Each approach has its advantages, through the analysis, we found that univariate TS clustering using k-means with SBD helps to understand the individual pollutant behaviour and the concentration of each pollutant around the UK.

For further understanding, we analysed the spatial and temporal distribution of the clustering results and showed how clustering helps us to interpret pollutant behaviour through the analysis of the clusters' centroid (Section 5.2).

Spatially, we found that PM concentrations (PM_{10} and $PM_{2.5}$) have very clear geographical patterns among other pollutants where stations are clustered into well-separated groups around the UK. On the other hand, O_3 clusters have less clear geographical separation than PM but better than NO_2. We found that stations that measured NO_2 are the most difficult to cluster, we believed this is due to traffic urban stations where NO_2 is affected by traffic volume near these stations.

Temporally, through the analysis of time variation on the clusters' centroids, we found that there is clear graduation of the concentrations of PMs across the UK

from the South to the North. For O_3 concentrations, we found two levels of ozone concentrations across the UK, high in the North/West and lower in the Center and the South/East. For NO_2 concentrations, clusters' centroids showed quite similar variations to one another where the South has slightly higher concentrations followed by the Center of the UK, while the lowest concentrations found in the North. Clustering found a group of stations distributed across the UK (with no spatial pattern) where concentrations of NO_2 are very high compared to other groups/clusters. Our analysis of these stations showed that stations are located near to NO_2 sources (i.e. roads, motorways, highways), which is reasonable for NO_2 behaviour.

Univariate TS clustering approach is simple and easy to implement and it helps in understanding the pollutant behaviours around the UK; however, the main drawback is that we need four clustering solutions one for each pollutant, which will make the pollutant imputation limited to stations that are included in the clustering of each pollutant. This means we can only impute a pollutant (let's say O_3) in a station if that is station included in the individual clustering solution of O_3. At the opposite end, if that station does not measure O_3 at all, then it will not be included in the O_3 clustering solution, and therefore we cannot do the imputation of O_3 without knowing to which cluster this station belongs.

On the other hand, as discussed in Chapter 6, the MVTS clustering that is based on the fused distance of the four air pollutants and the k-means clustering gave better clustering results that are geographical compact and well separated compared to the individual clustering. The advantages of this approach, that it helped us to understand the correlation between pollutants through clustering and discover multiple patterns of pollutant behaviour. Also, it increased the imputation models' ability to impute multiple pollutants in a station through one clustering solution. This answered the first and the second research questions.

The MVTS clustering approach that is based on fused distance can be affected by the uncertainty resulting either from missing data or disagreement between

individual pollutant similarity, as discussed in Chapter 3, in our experiments we investigated if the uncertainty impact the clustering results to answer the third research question. We calculated the overall uncertainty generated from the fused distance and unmeasured (missing) pollutants. Then, we implemented two versions of the k-means clustering algorithms (i.e. the weighted and the two-phases k-means algorithms) to cluster the fused distance considering the uncertainty. As a result, we found neither of these versions provided particularly improved clustering results, which confirmed that the uncertainty of the missing data does not impact the clustering results.

Due to the advantages of the MVTS clustering approach over the univariate clustering approach, MVTS has been selected for the next stage of our proposed solution (i.e. missing pollutants imputation). In this stage, we proposed nine imputation models. These models fall under three categories: clustering based models, geographical based models, and ensemble models. Our aim of using the clustering based models is to represent the temporal similarity between stations, while the geographical based models represents the spatial similarity. On the other hand, the ensemble models aggregate the temporal and spatial similarity between stations by taking the median imputed values from other models. We started with the clustering based models and since the clustering results inferred a strong geographically correlation between stations, we included other models that are based on the geographical similarity.

To evaluate the imputation models' ability to reproduce pollutants that are already measured at stations, several models evaluation methods have been used as discussed in Chapter 7. The evaluation process showed that the performance of the imputation models varied from one pollutant to another and from one station type to another. We found that, two imputation models performed better than others: Model 8 (Median) was selected as the best imputation model for PM_{10}, and $PM_{2.5}$, and model 9 (Median.ENV) for O_3 and NO_2 imputation. These models gave the lowest RMSE, the highest index of agreement (IOA) and the highest Correlation

Coefficient (R) between imputed and real values. Also, these models met the minimum requirement criteria using FAC2 for air quality modelling. This answered the fifth research question.

We noticed that the station environmental type plays an important role in O_3 and NO_2 imputation over all proposed approaches (univariate and MVTS clustering). The advantage of these imputation models (model 8 and 9) is that they aggregate the spatial and temporal similarity between stations in their imputation, which confirms that both spatial and temporal similarity are very important to consider in the imputation. This answered the fourth research question.

In general, modelled concentrations from model 8 (Median) and model 9 (Median.ENV) are correlated well with the observed pollutant concentrations based on both statistical and graphical evaluation functions. Even though, these models slightly underestimate the very high concentrations and underestimate the very low concentrations for each pollutant. We found that modelled concentrations of PM_{10}, $PM_{2.5}$, and NO_2 underestimate the observation which is confirmed by the negative model mean bias by (-0.009, -0.012, and -0.028) using NMB respectively for these pollutants, especially at the high levels. While O_3 modelled concentrations slightly overestimate the observations with a model mean bias by (0.010).

We analysed some factors that may affect imputation model performance to answer the sixth research question. We found that pollutant behaviour and its emitted sources impact the model performance, as confirmed by the variation of the model's performance with different environmental types. Imputation models underestimate the concentrations in stations where pollutant concentrations are very high while overestimating the concentrations in stations with very low concentrations. For example, high NO_2 concentrations are correlated to local sources (e.g. read, traffic), we found that the selected model to impute NO_2 (i.e. model 9 (Median.ENV)) underestimate the concentrations in stations near to these sources. Despite this variation of the models' performance, the selected models (models 8 and 9) work equally well under different weather types, as shown in the analysis of models'

performance with Lamb Weather Types (LWTs) using Taylor's diagram analysis.

For further evaluation, we calculated DAQI based on imputed pollutants and compared it with observed DAQI. As confirmed by the comparison in Chapter 7, imputed pollutant concentrations can reproduce well the air quality index. Imputed DAQI agreed with the observed DAQI in 75% of the total number of days in the dataset (that represent 44,118 out of 60,955 days). In comparison, there are 25% of disagreements cases where imputed DAQI is higher/lower than observed DAQI. The majority of disagreements cases happened when both imputed and observed DAQI were within the low air quality band (i.e. DAQI index 1 to 3), while we found 0.1% of disagreement cases with high variations due to some sudden changes in pollutant concentrations for short period of time or due to missing observations during the day.

In relation to the last two research questions, we were able to provide some answers by suggesting which pollutant(s) might help to improve the air quality assessment in stations with unmeasured pollutants as detailed in Chapter 8. The main challenge for this task is that we do not have the ground truth for the imputed (unmeasured) pollutant, to overcome this challenge, the analysis was based on the comparison between imputed and observed DAQI assess the differences between the imputed DAQI and the observed DAQI to highlighting cases where the former is considerably higher than the latter, suggesting that a pollution episode occurred at a location but was not reported by the observed DAQI. Results show that imputation models (model 8 and 9) produced imputed DAQI that was higher than the observed DAQI in 46% of the total number of days included in the comparison. Even though this percentage is considered large, the majority of index disagreements happened when both imputed and observed DAQI fall within the low air quality band and would therefore have less impact.

We gave more attention to cases where the imputed DAQI is higher by more than three index values than observed DAQI, which would indicate a pollution episode within the imputed/unmeasured pollutant values. There were 102 days (represent-

ing 0.5% of the total disagreement) under this category. The analysis of these days showed that $PM_{2.5}$ and O_3 were mostly responsible. Also the time and locations of these episodes coincident with real episodes of high $PM_{2.5}$ and O_3 concentrations at stations where they are measured. We, therefore, concluded that it is important to measure O_3 and $PM_{2.5}$ in all sites or use our proposed imputation model as a substitute.

Lastly, looking at the importance of imputing missing pollutant concentration, the study was limited in investigating the effect of imputing missing (unmeasured) pollutants, which may have improved the air quality assessment. Imputation models through MVTS clustering helped to impute unmeasured pollutants and identify some true events at particular stations because of lack of measurements for some pollutants. Also, the proposed approach achieved this thesis's main goal by allowing for the calculation of DAQI based on all contributed pollutants and inferring where more measurements for specific pollutants may be beneficial to truly capture pollution episodes. It is important to address that since all clustering solutions are geographically consistent with pollutant behaviour, a station can be assigned to a cluster based on its geographical location, which enables pollutant imputation at that location. Using these imputation models over measuring all pollutants in all stations, enable to reduce measurements in the monitoring network and estimate pollutants concentrations in any location.

The most important limitation of our research is that even though the MVTS clustering approach developed in this study has achieved good imputation results based on our extensive evaluation against the univariate time series clustering approach, the performance of the proposed approach needs further evaluation against the traditional ensemble models. In the next section, we suggested some areas where analysis could be extended for future work.

9.3 Further Work

As mentioned, in this study we focus on comparing the proposed two approaches against each other, hence further comparison is needed for the applied MVTS clustering technique against some ensemble clustering methods that are based on univariate TS clustering, then applying the proposed imputation models to evaluate the performance of our MVTS clustering technique compared to traditional ensemble models. On the other hand, to improve the imputation models, several factors that may affect the air pollutants concentrations can be considered such as taking into consideration further information about the stations such as station altitude, location in relation to the weather effects, pollutant emitted sources, correlation between different pollutants, and station environmental type since there is a variation in the concentration of pollutants among the stations' type.

Another area that can improve the assessment of the air quality is looking at the distribution of the monitoring stations by identifying locations where there are more than enough measurements and new locations where we should make measurements to improve the air quality assessments. That can also include assessing in the context of population measurements, where it becomes more strategically important to produce improvements to the network (i.e. prioritise areas of dense population where insufficient measurements are available).

Bibliography

[1] DEFRA air pollution in the UK 2019, uk department for environment, food and rural affairs. https://uk-air.defra.gov.uk/library/annualreport/index2020.

[2] Public Health sources and effects of pm2.5. https://laqm.defra.gov.uk/public-health/pm25.html, 2016.

[3] DEFRA air pollution in the UK 2018. https://uk-air.defra.gov.uk/library/annualreport/, 2019.

[4] Centreforcities cities outlook 2020. https://www.centreforcities.org/publication/cities-outlook-2020/, 2020.

[5] DEFRA about air pollution. https://uk-air.defra.gov.uk/air-pollution/, 2020.

[6] DEFRA air information resource. http://uk-air.defra.gov.uk, 2020.

[7] DEFRA modelled background pollution data. https://uk-air.defra.gov.uk/data/pcm-data, 2020.

[8] National Statistics concentrations of particulate matter pm_{10} and pm_{25}. https://www.gov.uk/

government/publications/air-quality-statistics/concentrations-of-particulate-matter-pm10-and-pm25, 2020.

[9] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. Time-series clustering–a decade review. *Information Systems*, 53:16–38, 2015.

[10] Peerzada Hamid Ahmad and Shilpa Dang. Performance evaluation of clustering algorithm using different datasets. *International Journal of Advance Research in Computer Science and Management Studies*, 3(1):167–173, 2015.

[11] E. Al-Abri. Modelling atmospheric ozone concentration using machine learning algorithms. 2016.

[12] W. Alahamade, I. Lake, C. E. Reeves, and B. De La Iglesia. Evaluation of multivariate time series clustering for imputation of air pollution data. *Geoscientific Instrumentation, Methods and Data Systems*, 10(2):265–285, 2021.

[13] Wedad Alahamade, Iain Lake, Claire E Reeves, and Beatriz De La Iglesia. Clustering imputation for air pollution data. In *International Conference on Hybrid Artificial Intelligence Systems*, pages 585–597. Springer, 2020.

[14] Wedad Alahamade, Iain Lake, Claire E Reeves, and Beatriz De La Iglesia. A multi-variate time series clustering approach based on intermediate fusion a case study in air pollution data imputation. in review, 2021.

[15] Paul D Allison. *Missing data*, volume 136. Sage publications, 2001.

[16] Andrés M Alonso and Daniel Peña. Clustering time series by linear dependency. *Statistics and Computing*, 29(4):655–676, 2019.

[17] China air quality. https://www.healthandsafetyinshanghai.com/china-air-quality.html, 2019.

[18] National air quality index. https://app.cpcbccr.com/AQI_India/, 2019.

[19] Air pollution index of malaysia. `https://apims.doe.gov.my/public_v2/faq.html`, 2019.

[20] Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza, JesúS M PéRez, and IñIgo Perona. An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1):243–256, 2013.

[21] Automatic urban and rural network: Site operator's manual. `https://uk-air.defra.gov.uk/`, 2017.

[22] Report: Statistical evaluation of the input meteorological data used for the uk air quality forecast (uk-aqf). `https://uk-air.defra.gov.uk/library/reports?report_id=770`, 2013.

[23] Elena Austin, Brent Coull, Dylan Thomas, and Petros Koutrakis. A framework for identifying distinct multipollutant profiles in air pollution data. *Environment international*, 45:112–121, 2012.

[24] Elena Austin, Brent A Coull, Antonella Zanobetti, and Petros Koutrakis. A framework to spatially cluster air pollution monitoring sites in us based on the pm2. 5 composition. *Environment international*, 59:244–254, 2013.

[25] Rabia Aziz, CK Verma, and Namita Srivastava. Dimension reduction methods for microarray data: a review. *AIMS Bioengineering*, 4(2):179–197, 2017.

[26] Melissa J Azur, Elizabeth A Stuart, Constantine Frangakis, and Philip J Leaf. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49, 2011.

[27] MA Barrero, JAG Orza, M Cabello, and L Cantón. Categorisation of air quality monitoring stations by evaluation of pm10 variability. *Science of The Total Environment*, 524:225–236, 2015.

[28] Colin Bellinger, Mohomed Shazan Mohomed Jabbar, Osmar Zaïane, and Alvaro Osornio-Vargas. A systematic review of data mining and machine learning for air pollution epidemiology. *BMC public health*, 17(1):907, 2017.

[29] GEORGE EP Box, Gwilym M Jenkins, and G Reinsel. Time series analysis: forecasting and control holden-day san francisco. *BoxTime Series Analysis: Forecasting and Control Holden Day1970*, 1970.

[30] S van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, pages 1–68, 2010.

[31] European citeair index. `https://www.airparif.asso.fr/index.php/en`, 2019.

[32] José Juan Carbajal-Hernández, Luis P Sánchez-Fernández, Jesús A Carrasco-Ochoa, and José Fco Martínez-Trinidad. Assessment and prediction of air quality using fuzzy logic and autoregressive models. *Atmospheric Environment*, 60:37–50, 2012.

[33] David C Carslaw and Karl Ropkins. Openair—an r package for air quality data analysis. *Environmental Modelling & Software*, 27:52–61, 2012.

[34] Cities outlook 2020 - air quality in uk cities. `https://www.centreforcities.org/publication/cities-outlook-2020/`, 2020.

[35] Enhong Chen and Feng Wang. Dynamic clustering using multi-objective evolutionary algorithm. In *International Conference on Computational and Information Science*, pages 73–80. Springer, 2005.

[36] Ho-Wen Chen, Ching-Tsan Tsai, Chin-Wen She, Yo-Chen Lin, and Chow-Feng Chiang. Exploring the background features of acidic and basic air pollutants around an industrial complex using data mining approach. *Chemosphere*, 81(10):1358–1367, 2010.

[37] Jinglong Chen, Jun Pan, Zipeng Li, Yanyang Zi, and Xuefeng Chen. Generator bearing fault diagnosis for wind turbine via empirical wavelet transform using measured vibration signals. *Renewable Energy*, 89:80–92, 2016.

[38] Mei Chen, Pengfei Wang, Qiang Chen, Jiadong Wu, and Xiaoyun Chen. A clustering algorithm for sample data based on environmental pollution characteristics. *Atmospheric Environment*, 107:194–203, 2015.

[39] Sheng Chen, Guangyuan Kan, Jiren Li, Ke Liang, and Yang Hong. Investigating china's urban air quality using big data, information theory, and machine learning. *Polish Journal of Environmental Studies*, 27(2), 2018.

[40] Tapiwa M Chiwewe and Jeofrey Ditsela. Machine learning based estimation of ozone using spatio-temporal data from air quality monitoring stations. In *2016 IEEE 14th International Conference on Industrial Informatics (INDIN)*, pages 58–63. IEEE, 2016.

[41] E Connolly, G Fuller, T Baker, and P Willis. Update on implementation of the daily air quality index. department for environment. *Food and Rural Affairs*, pages 1–11, 2013.

[42] Giorgio Corani and Mauro Scanagatta. Air pollution prediction via multi-label classification. *Environmental modelling & software*, 80:259–264, 2016.

[43] DEFRA air pollution in the UK. `https://uk-air.defra.gov.uk/library/annualreport/`, 2017.

[44] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.

[45] Ton De Waal, Jeroen Pannekoek, and Sander Scholtus. *Handbook of statistical data editing and imputation*, volume 563. John Wiley & Sons, 2011.

[46] Daily air quality index implementation report. `https://uk-air.defra.gov.uk/library/reports?report_id=750/`, 2013.

[47] Gerardo Di Bello, Vincenzo Lapenna, Maria Macchiato, Celestina Satriano, Carmine Serio, Valerio Tramutoli, et al. Parametric time series analysis of geoelectrical signals: an application to earthquake forecasting in southern italy. 1996.

[48] Florencia MR Diaz, M Anwar H Khan, Beth Shallcross, Esther DG Shallcross, Ulrich Vogt, and Dudley E Shallcross. Ozone trends in the united kingdom over the last 30 years. *Atmosphere*, 11(5):534, 2020.

[49] John Abbott Mike Jenkin Paul Willis Dick Derwent, Andrea Fraser and Tim Murrells. Report: Evaluating the performance of air quality models. *Department for Environment, Food and Rural Affairs, London*, 2010.

[50] Evgenia Dimitriadou, Andreas Weingessel, and Kurt Hornik. Voting-merging: An ensemble method for clustering. In *International Conference on Artificial Neural Networks*, pages 217–224. Springer, 2001.

[51] Hui Ding, Goce Trajcevski, Peter Scheuermann, Xiaoyue Wang, and Eamonn Keogh. Querying and mining of time series data: experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment*, 1(2):1542–1552, 2008.

[52] Ling Ding, Daojuan Zhu, Donghong Peng, and Yao Zhao. Air pollution and asthma attacks in children: A case–crossover analysis in the city of chongqing, china. *Environmental Pollution*, 220:348–353, 2017.

[53] Joseph C Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. 1973.

[54] Joseph C Dunn. Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, 4(1):95–104, 1974.

[55] Pierpaolo D'Urso, Livia De Giovanni, and Riccardo Massari. Robust fuzzy clustering of multivariate time trajectories. *International Journal of Approximate Reasoning*, 99:12–38, 2018.

[56] Pierpaolo D'Urso, Elizabeth A Maharaj, and Andrés M Alonso. Fuzzy clustering of time series using extremes. *Fuzzy Sets and Systems*, 318:56–79, 2017.

[57] Iris Eekhout, R Michiel de Boer, Jos WR Twisk, Henrica CW de Vet, and Martijn W Heymans. Missing data: a systematic review of how they are reported and handled. *Epidemiology*, 23(5):729–732, 2012.

[58] Philippe Esling and Carlos Agon. Time-series data mining. *ACM Computing Surveys (CSUR)*, 45(1):12, 2012.

[59] Cristiano Hora Fontes and Hector Budman. A hybrid clustering approach for multivariate time series –a case study applied to failure analysis in a gas turbine. *ISA transactions*, 71:513–529, 2017.

[60] Edward B Fowlkes and Colin L Mallows. A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, 78(383):553–569, 1983.

[61] Andrew Frank and Arthur Asuncion. UCI Machine Learning Repository [http://archive. ics. uci. edu/ml]. Irvine, CA: University of California. *School of information and computer science*, 213(11), 2010.

[62] Wei Gao, Xuexi Tie, Jianming Xu, Rujin Huang, Xiaoqing Mao, Guangqiang Zhou, and Luyu Chang. Long-term trend of o3 in a mega city (shanghai), china: Characteristics, causes, and interactions with precursors. *Science of the Total Environment*, 603:425–433, 2017.

[63] Katherine Gass, Mitch Klein, Howard H. Chang, W. Dana Flanders, and Matthew J. Strickland. Classification and regression trees for epidemiologic research: an air pollution example. *Environmental Health*, 13(1):17, Mar 2014.

[64] Christophe Genolini and Bruno Falissard. KmL: k-means for longitudinal data. *Computational Statistics*, 25(2):317–328, 2010.

[65] Ranjana Waman Gore and Deepa S Deshpande. An approach for classification of health risks based on air quality levels. In *Intelligent Systems and Information Management (ICISIM), 2017 1st International Conference on*, pages 58–61. IEEE, 2017.

[66] Emissions of air pollutants in the UK – sulphur dioxide (so2). `https://www.gov.uk/government/statistics/emissions-of-air-pollutants`, 2021.

[67] Ailish M Graham, Kirsty J Pringle, Stephen R Arnold, Richard J Pope, Massimo Vieno, Edward W Butt, Luke Conibear, Ellen L Stirling, and James B McQuaid. Impact of weather types on uk ambient particulate matter concentrations. *Atmospheric Environment: X*, 5:100061, 2020.

[68] AQE GROUP et al. Fine particulate matter (pm 2.5) in the united kingdom. *Department for Environment, Food and Rural Affairs, London*, 2012.

[69] AQEG GROUP. Aqeg: Ozone in the united kingdom. fifth report of the air quality expert group. *Department for Environment, Food and Rural Affairs, London*, 2009.

[70] Chonghui Guo, Hongfeng Jia, and Na Zhang. Time series clustering based on ICA for stock data analysis. In *2008 4th International Conference on Wireless Communications, Networking and Mobile Computing*, pages 1–4. IEEE, 2008.

[71] M Anwar H. Khan, William C Morris, Matthew Galloway, Beth M A. Shallcross, Carl J Percival, and Dudley E Shallcross. An estimation of the levels of stabilized criegee intermediates in the UK urban and rural atmosphere using the steady-state approximation and the potential effects of these intermediates on tropospheric oxidation cycles. *International journal of chemical kinetics*, 49(8):611–621, 2017.

[72] Petr Hajek, Vladimir Olej, et al. Predicting common air quality index-the case of czech microregions. *Aerosol and Air Quality Research*, 15(2):544–555, 2015.

[73] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of intelligent information systems*, 17(2):107–145, 2001.

[74] Maria Halkidi, Michalis Vazirgiannis, and Yannis Batistakis. Quality scheme assessment in the clustering process. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 265–276. Springer, 2000.

[75] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[76] Weirong Han, Shengjun Zhai, Jia Guo, Seungwon Lee, Kai Chang, and Guihongxuan Zhang. Analysis of no2 and o3 air quality indices and forecasting using machine learning models. 2018.

[77] Wu He, Shenghua Zha, and Ling Li. Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, 33(3):464–472, 2013.

[78] Tracey Holloway, Scott N Spak, Daniel Barker, Matthew Bretl, Claus Moberg, Katharine Hayhoe, Jeff Van Dorn, and Donald Wuebbles. Change in ozone air pollution over chicago associated with global climate change. *Journal of Geophysical Research: Atmospheres*, 113(D22), 2008.

[79] Piotr Holnicki and Zbigniew Nahorski. Emission data uncertainty in urban air quality modeling—case study. *Environmental Modeling & Assessment*, 20(6):583–597, 2015.

[80] Philip K Hopke, Chuanhai Liu, and Donald B Rubin. Multiple imputation for multivariate data with missing and below-threshold measurements: time-

series concentrations of pollutants in the arctic. *Biometrics*, 57(1):22–33, 2001.

[81] Rosaria Ignaccolo, Stefania Ghigo, and E Giovenali. Analysis of air quality monitoring networks by functional clustering. *Environmetrics*, 19(7):672–686, 2008.

[82] Paul Jaccard. Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.*, 44:223–270, 1908.

[83] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[84] AF Jenkinson and FP Collison. An initial climatology of gales over the north sea. *Synoptic climatology branch memorandum*, 62:18, 1977.

[85] Marilena Kampa and Elias Castanas. Human health effects of air pollution. *Environmental pollution*, 151(2):362–367, 2008.

[86] Ganganjot Kaur Kang, Jerry Zeyu Gao, Sen Chiao, Shengqiang Lu, and Gang Xie. Air quality prediction: Big data and machine learning approaches. *International Journal of Environmental Science and Development*, 9(1):8–16, 2018.

[87] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

[88] Jan Kleine Deters, Rasa Zalakeviciute, Mario Gonzalez, and Yves Rybarczyk. Modeling pm2. 5 urban pollution using machine learning and selected meteorological parameters. *Journal of Electrical and Computer Engineering*, 2017, 2017.

[89] Anikender Kumar and Pramila Goyal. Forecasting of air quality in delhi using principal component regression technique. *Atmospheric Pollution Research*, 2(4):436–444, 2011.

[90] Atakan Kurt and Ayşe Betül Oktay. Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks. *Expert Systems with Applications*, 37(12):7986–7992, 2010.

[91] Hubert Horace Lamb. British isles weather types and a register of the daily sequence of circulation patterns 1861-1971. 1972.

[92] Nada Lavrač. Machine learning for data mining in medicine. In *Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making*, pages 47–62. Springer, 1999.

[93] SD Lee, GJR Wolters, LD Grant, and T Schneider. *Atmospheric ozone research and its policy implications*. Elsevier, 1989.

[94] Chen Li and Zhijie Zhu. Research and application of a novel hybrid air quality early-warning system: A case study in china. *Science of the Total Environment*, 626:1421–1438, 2018.

[95] Hailin Li. Multivariate time series clustering based on common principal component analysis. *Neurocomputing*, 349:239–247, 2019.

[96] Hailin Li and Tian Du. Multivariate time-series clustering based on component relationship networks. *Expert Systems with Applications*, 173:114649, 2021.

[97] Hailin Li and Miao Wei. Fuzzy clustering based on feature weights for multivariate time series. *Knowledge-Based Systems*, 197:105907, 2020.

[98] T Warren Liao. Clustering of time series data—a survey. *Pattern recognition*, 38(11):1857–1874, 2005.

[99] Alexander Lin, Yingzhuo Zhang, Jeremy Heng, Stephen A Allsop, Kay M Tye, Pierre E Jacob, and Demba Ba. Clustering time series with nonlinear dynamics: A bayesian non-parametric and particle-based approach. In *The*

22nd International Conference on Artificial Intelligence and Statistics, pages 2476–2484. PMLR, 2019.

[100] Chun Lin, Xiaofan Feng, and Mathew R Heal. Temporal persistence of intra-urban spatial contrasts in ambient no2, o3 and ox in edinburgh, uk. *Atmospheric Pollution Research*, 7(4):734–741, 2016.

[101] Roderick JA Little and Donald B Rubin. Single imputation methods. *Statistical analysis with missing data*, pages 59–74, 2002.

[102] Bing-Chun Liu, Arihant Binaykia, Pei-Chann Chang, Manoj Kumar Tiwari, and Cheng-Chin Tsao. Urban air quality forecasting based on multi-dimensional collaborative support vector regression (svr): A case study of beijing-tianjin-shijiazhuang. *PloS one*, 12(7):e0179763, 2017.

[103] Tong Liu, Alexis KH Lau, Kai Sandbrink, and Jimmy CH Fung. Time series forecasting of air quality based on regional numerical modeling in hong kong. *Journal of Geophysical Research: Atmospheres*, 123(8):4175–4196, 2018.

[104] Helen L Macintyre, Clare Heaviside, Lucy S Neal, Paul Agnew, John Thornes, and Sotiris Vardoulakis. Mortality and emergency hospitalizations associated with atmospheric particulate matter episodes across the UK in spring 2014. *Environment international*, 97:108–116, 2016.

[105] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[106] Elizabeth Ann Maharaj and Pierpaolo D'Urso. Fuzzy clustering of time series in the frequency domain. *Information Sciences*, 181(7):1187–1211, 2011.

[107] Gina M Mazzuca, Xinrong Ren, Christopher P Loughner, Mark Estes, James H Crawford, Kenneth E Pickering, Andrew J Weinheimer, and Rus-

sell R Dickerson. Ozone production and its sensitivity to nox and vocs: Results from the discover-aq field experiment, houston 2013. 2016.

[108] Karl Øyvind Mikalsen, Filippo Maria Bianchi, Cristina Soguero-Ruiz, and Robert Jenssen. Time series cluster kernel for learning similarities between multivariate time series with missing data. *Pattern Recognition*, 76:569–581, 2018.

[109] Aalaa Mojahed and Beatriz de la Iglesia. An adaptive version of k-medoids to deal with the uncertainty in clustering heterogeneous data using an intermediary fusion approach. *Knowledge and Information Systems*, 50(1):27–52, 2017.

[110] Steffen Moritz and Thomas Bartz-Beielstein. imputets: time series missing value imputation in r. *The R Journal*, 9(1):207–218, 2017.

[111] Panagiotis T Nastos, Athanasios G Paliatsos, Michael B Anthracopoulos, Eleftheria S Roma, and Kostas N Priftis. Outdoor particulate matter and childhood asthma admissions in athens, greece: a time-series study. *Environmental Health*, 9(1):45, 2010.

[112] V Naveen and N Anu. Time series analysis to forecast air quality indices in thiruvananthapuram district, kerala, india. 2017.

[113] XY Ni, Hong Huang, and WP Du. Relevance analysis and short-term prediction of pm2.5 concentrations in beijing based on multi-source data. *Atmospheric environment*, 150:146–161, 2017.

[114] Venkatadri M Niharika and Padma S Rao. A survey on air quality forecasting techniques. *International Journal of Computer Science and Information Technologies*, 5(1):103–107, 2014.

[115] Mohamed Noor Norazian, Yahaya Ahmad Shukri, Ramli Nor Azam, and Abdullah Mohd Mustafa Al Bakri. Estimation of missing values in air pollution data using single imputation techniques. *ScienceAsia*, 34(3):341–345, 2008.

[116] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[117] Esma Ergüner Özkoç. Clustering of time-series data. In *Data Mining-Methods, Applications and Systems*. IntechOpen, 2020.

[118] Shraddha Pandit, Suchita Gupta, et al. A comparative study on distance measuring approaches for clustering. *International Journal of Research in Computer Science*, 2(1):29–31, 2011.

[119] John Paparrizos and Luis Gravano. k-shape: Efficient and accurate clustering of time series. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1855–1870. ACM, 2015.

[120] Kay I Penny and Graeme D Smith. The use of data-mining to identify indicators of health-related quality of life in patients with irritable bowel syndrome. *Journal of clinical nursing*, 21(19pt20):2761–2771, 2012.

[121] A Plaia and AL Bondi. Single imputation method of missing values in environmental pollution data sets. *Atmospheric Environment*, 40(38):7316–7330, 2006.

[122] Girish Punj and David W Stewart. Cluster analysis in marketing research: Review and suggestions for application. *Journal of marketing research*, 20(2):134–148, 1983.

[123] Trivellore E Raghunathan, James M Lepkowski, John Van Hoewyk, Peter Solenberger, et al. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology*, 27(1):85–96, 2001.

[124] Francisco Ramos, Sergio Trilles, Andrés Muñoz, and Joaquín Huerta. Promoting pollution-free routes in smart cities using air quality sensor networks. *Sensors*, 18(8):2507, 2018.

[125] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[126] Eréndira Rendón, Itzel Abundez, Alejandra Arizmendi, and Elvia M Quiroz. Internal versus external cluster validation indexes. *International Journal of computers and communications*, 5(1):27–34, 2011.

[127] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[128] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3), 1976.

[129] Donald B Rubin. *Statistical analysis with missing data*. Wiley, 1987.

[130] Donald B Rubin. An overview of multiple imputation. In *Proceedings of the survey research methods section of the American statistical association*. Citeseer, 1988.

[131] Donald B Rubin. Multiple imputation after 18+ years. *Journal of the American statistical Association*, 91(434):473–489, 1996.

[132] Donald B Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.

[133] Y. Rybarczyk and R. Zalakeviciute. Machine learning approach to forecasting urban pollution. In *2016 IEEE Ecuador Technical Chapters Meeting (ETCM)*, pages 1–6, Oct 2016.

[134] Yves Rybarczyk and Rasa Zalakeviciute. Machine learning approach to forecasting urban pollution. In *Ecuador Technical Chapters Meeting (ETCM), IEEE*, pages 1–6. IEEE, 2016.

[135] Kidakan Saithan and Jatupat Mekparyup. Clustering of air quality and meteorological variables associated with high ground ozone concentration in

the industrial areas, at the east of thailand. *International Journal of Pure and Applied Mathematics*, 81(3):505–515, 2012.

[136] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978.

[137] G. Salton and M. McGill. Introduction to modern information retrieval, 1983.

[138] Alexis Sardá-Espinosa. Comparing time-series clustering algorithms in r using the dtwclust package. *Vienna: R Development Core Team*, 2017.

[139] Joseph L Schafer. *Analysis of incomplete multivariate data*. Chapman and Hall/CRC, 1997.

[140] Joan Serra and Josep Lluís Arcos. A competitive measure to assess the similarity between two time series. In *International Conference on Case-Based Reasoning*, pages 414–427. Springer, 2012.

[141] Skyler Seto, Wenyu Zhang, and Yichen Zhou. Multivariate time series classification using dynamic time warping template selection for human activity recognition. In *2015 IEEE Symposium Series on Computational Intelligence*, pages 1399–1406. IEEE, 2015.

[142] Dinkal Shah and Narendra Limbad. A literature survey on contrast data mining. 2015.

[143] Nidhi Sharma, Shweta Taneja, Vaishali Sagar, and Arshita Bhatt. Forecasting air pollution load in delhi using data analysis tools. *Procedia computer science*, 132:1077–1085, 2018.

[144] Robert H Shumway and David S Stoffer. *Time series analysis and its applications: with R examples*. Springer, 2017.

[145] Kunwar P Singh, Shikha Gupta, and Premanjali Rai. Identifying pollution sources and predicting urban air quality using ensemble learning methods. *Atmospheric Environment*, 80:426–437, 2013.

[146] Krzysztof Siwek and Stanisław Osowski. Data mining methods for prediction of air pollution. *International Journal of Applied Mathematics and Computer Science*, 26(2):467–478, 2016.

[147] Ping-Wei Soh, Kai-Hsiang Chen, Jen-Wei Huang, and Hone-Jay Chu. Spatial-temporal pattern analysis and prediction of air quality in taiwan. In *2017 10th International Conference on Ubi-media Computing and Workshops (Ubi-Media)*, pages 1–6. IEEE, 2017.

[148] Jeanette A Stingone, Om P Pandey, Luz Claudio, and Gaurav Pandey. Using machine learning to identify air pollution exposure profiles associated with early cognitive skills among us children. *Environmental Pollution*, 230:730–740, 2017.

[149] Mrs S Sujatha and Mrs A Shanthi Sona. Novel initialization technique for k-means clustering using spectral constraint prototype. *Journal of Global Research in Computer Science*, 3(6):46–50, 2012.

[150] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. Pearson Education India, 2016.

[151] Karl E Taylor. Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, 106(D7):7183–7192, 2001.

[152] John Thornes. Impact of the march/april 2014 air pollution episodes' on acute morbidity outcomes using london ambulance data. 2017.

[153] Giulia Toti, Ricardo Vilalta, Peggy Lindner, Barry Lefer, Charles Macias, and Daniel Price. Analysis of correlation between pediatric asthma exacerbation

and exposure to pollutant mixtures with association rule mining. *Artificial intelligence in medicine*, 74:44–52, 2016.

[154] Goksu Tuysuzoglu, Derya Birant, and Aysegul Pala. Majority voting based multi-task clustering of air quality monitoring network in turkey. *Applied Sciences*, 9(8):1610, 2019.

[155] Stef Van Buuren. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research*, 16(3):219–242, 2007.

[156] Stef Van Buuren and Karin Oudshoorn. *Flexible multivariate imputation by MICE*. Leiden: TNO, 1999.

[157] Claudia Vitolo, Marco Scutari, Mohamed Ghalaieny, Allan Tucker, and Andrew Russell. Modeling air pollution, climate, and health data using bayesian networks: A case study of the english regions. *Earth and Space Science*, 5(4):76–88, 2018.

[158] Michail Vlachos, Marios Hadjieleftheriou, Dimitrios Gunopulos, and Eamonn Keogh. Indexing multi-dimensional time-series with support for multiple distance measures. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 216–225, 2003.

[159] Chi-Man Vong, Weng-Fai Ip, Pak-kin Wong, and Jing-yi Yang. Short-term prediction of air pollution in macau using support vector machines. *Journal of Control Science and Engineering*, 2012:4, 2012.

[160] Vuokko Vuori and Jorma Laaksonen. A comparison of techniques for automatic clustering of handwritten characters. In *Object recognition supported by user interaction for service robots*, volume 3, pages 168–171. IEEE, 2002.

[161] Junshan Wang and Guojie Song. A deep spatial-temporal ensemble model for air quality prediction. *Neurocomputing*, 314:198–206, 2018.

[162] Xiaozhe Wang, Kate Smith, and Rob Hyndman. Characteristic-based clustering for time series data. *Data Mining and knowledge Discovery*, 13(3):335–364, 2006.

[163] Dan Wei. Predicting air pollution level in a specific city, 2014.

[164] Eric W Weisstein. Frobenius norm. 2003.

[165] TIAN Wende, HU Minggang, and LI Chuankun. Fault prediction based on dynamic model and grey time series model in chemical processes. *Chinese Journal of Chemical Engineering*, 22(6):643–650, 2014.

[166] Air pollution. `https://www.who.int/health-topics/air-pollution#tab=tab_1`, 2021.

[167] Wikipedia contributors. Air pollution — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Air_pollution&oldid=877082014`, 2019. [Online; accessed 16-January-2019].

[168] Cort J Willmott. On the validation of models. *Physical geography*, 2(2):184–194, 1981.

[169] Cort J Willmott, Scott M Robeson, and Kenji Matsuura. A refined index of model performance. *International Journal of climatology*, 32(13):2088–2094, 2012.

[170] Edmond HC Wu and LH Philip. Independent component analysis for clustering multivariate time series data. In *International Conference on Advanced Data Mining and Applications*, pages 474–482. Springer, 2005.

[171] Wei-Feng Ye, Zhong-Yu Ma, Xiu-Zhen Ha, Hai-Chao Yang, and Zhi-Xiong Weng. Spatiotemporal patterns and spatial clustering characteristics of air quality in china: A city level analysis. *Ecological Indicators*, 91:523–530, 2018.

[172] Fu Yongjian. Data mining: tasks, techniques and applications. *IEEE Potentials*, 16(4):18–20, 1997.

[173] Yu Zheng, Furui Liu, and Hsun-Ping Hsieh. U-air: When urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1436–1444. ACM, 2013.

[174] Yu Zheng, Xiuwen Yi, Ming Li, Ruiyuan Li, Zhangqing Shan, Eric Chang, and Tianrui Li. Forecasting fine-grained air quality based on big data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2267–2276. ACM, 2015.

[175] Pei-Yuan Zhou and Keith CC Chan. A model-based multivariate time series clustering algorithm. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 805–817. Springer, 2014.

[176] Dixian Zhu, Changjie Cai, Tianbao Yang, and Xun Zhou. A machine learning approach for air quality prediction: model regularization and optimization. *Big data and cognitive computing*, 2(1):5, 2018.

[177] Marinka Žitnik and Blaž Zupan. Data fusion by matrix factorization. *IEEE transactions on pattern analysis and machine intelligence*, 37(1):41–53, 2014.

Appendix A

AURN Details

Table A.1: Numbers of AURN stations based on the UK air quality zones.

	Region	Air Quality Zone	Station	Zone Type
1	Greater London	Greater London Urban Area	15	Agglomeration
2	South East	Brighton,Littlehampton	2	Agglomeration
		Portsmouth Urban Area	2	Agglomeration
		Southampton Urban Area	2	Agglomeration
		Reading/Wokingham Urban Area	2	Agglomeration
		South East	10	Non-Agglomeration
3	South West	Bristol Urban Area	2	Agglomeration
		Bournemouth Urban Area	2	Agglomeration
		South West	10	Non-Agglomeration
4	West Midlands	Coventry/Bedworth	2	Agglomeration
		West Midlands Urban Area	6	Agglomeration
		ThePotteries	2	Agglomeration
		West Midlands	5	Non-Agglomeration
5	North West	Blackpool Urban Area	1	Agglomeration
		Preston Urban Area	1	Agglomeration
		Liverpool Urban Area	2	Agglomeration
		Greater Manchester Urban Area	5	Agglomeration
		Birkenhead Urban Area	2	Agglomeration
		North West,Merseyside	6	Non-Agglomeration
6	North East	Teesside Urban Area	3	Agglomeration
		Tyneside	2	Agglomeration
		North East	4	Non-Agglomeration
7	Yorkshire and Humberside	West Yorkshire Urban Area	4	Agglomeration
		Kingston upon Hull	2	Agglomeration
		Sheffield Urban Area	3	Agglomeration
		Yorkshire,Humberside	7	Non-Agglomeration
8	East Midlands	Nottingham Urban Area	2	Agglomeration
		Leicester Urban Area	2	Agglomeration
		East Midlands	7	Non-Agglomeration
9	East of England	Southend Urban Area	1	Agglomeration
		Easternt	11	Non-Agglomeration
10	Scotland regions- Highland	Highland	3	Non- Agglomeration
11	Scotland regions- North East Scotland	North East Scotland	4	Non- Agglomeration
12	Scotland regions-Central	Glasgow Urban Area	4	Agglomeration
		Edinburgh Urban Area	2	Agglomeration
		Central Scotland	6	Agglomeration
		Scottish Borders	3	Non-Agglomeration
13	South Wales	Swansea Urban Area	2	Agglomeration
		Cardiff Urban Area	2	Agglomeration
		South Wales	5	Non-Agglomeration
14	North Wales	North Wales	2	Non- Agglomeration
15	Northern Ireland	Belfast Metropolitan Urban Area	2	Agglomeration
		Northern Ireland	5	Non-Agglomeration
16	Scottish Borders	Scottish Borders	2	Non- Agglomeration

Figure A.1: The UK zones for ambient air Quality reporting in 2017. Source:[6]

Appendix B

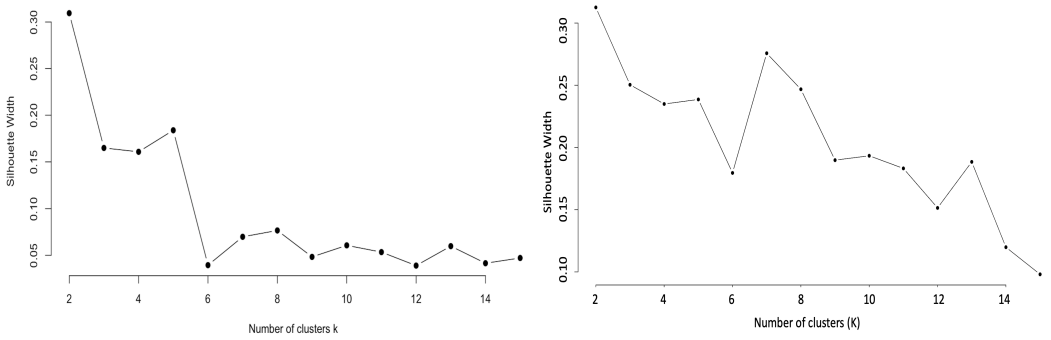
Comparison between clustering results using PAM with DTW and SBD.

These clustering results are based on PAM clustering algorithm with DTW and SBD. Datasets used for these clustering experiments are the daily mean concentrations of each pollutant and missing values within TS imputed using MICE. Stations with missing values with more than 25% of our study period are removed to reduce the effect of imputation process in the clustering results.

B.0.1 Clustering O_3 Dataset

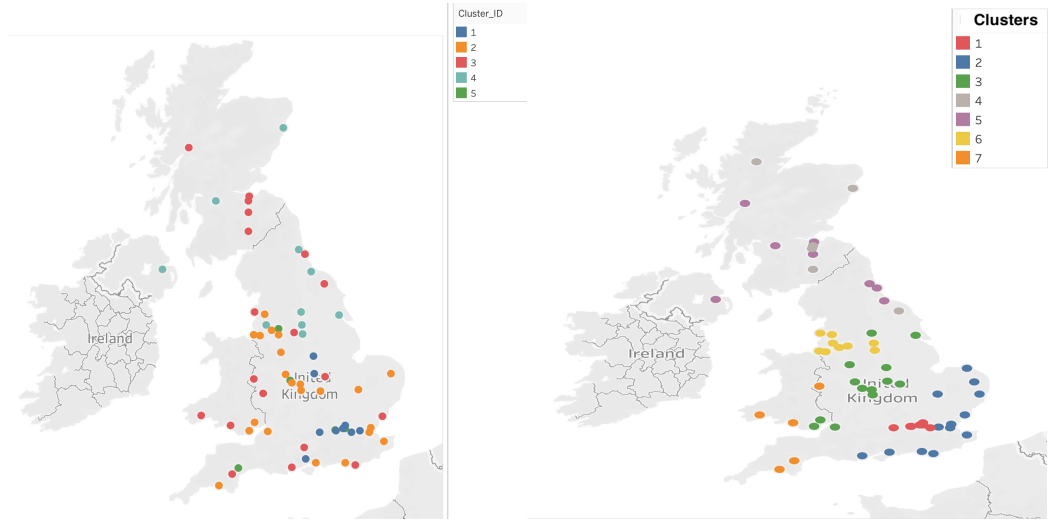
The total number of stations that measure the ozone O_3 is 70 stations, we removed 5 stations and include 65 stations in our clustering. The optimal number of clusters using SBD is K=7 and K=5 with DTW.

The average silhouette width for each clustering algorithm is shown in Figure B.1. The results of clustering this dataset using both DTW and SBD are shown in Figure B.2.



a b

Figure B.1: Average silhouette widths for 2 to 15 clusters with O_3 dataset. (A) PAM+DTW and (B) PAM+SBD.

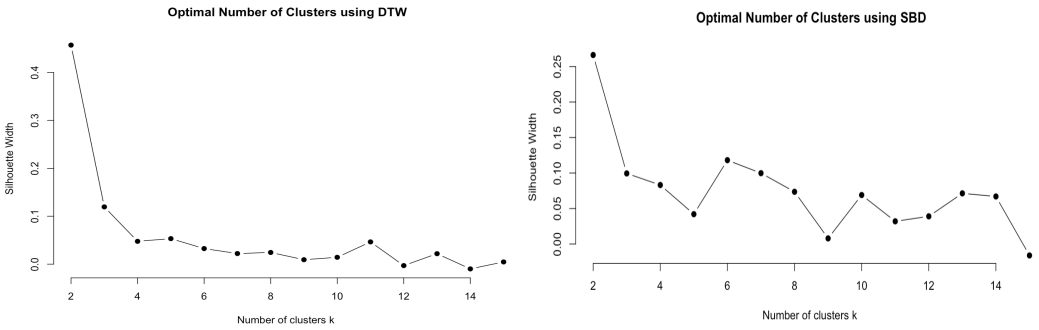


a b

Figure B.2: Geographical distribution of PAM clustering results on O_3 dataset. (A) clustering results based on PAM+DTW and (B) clustering results based on PAM+SBD.

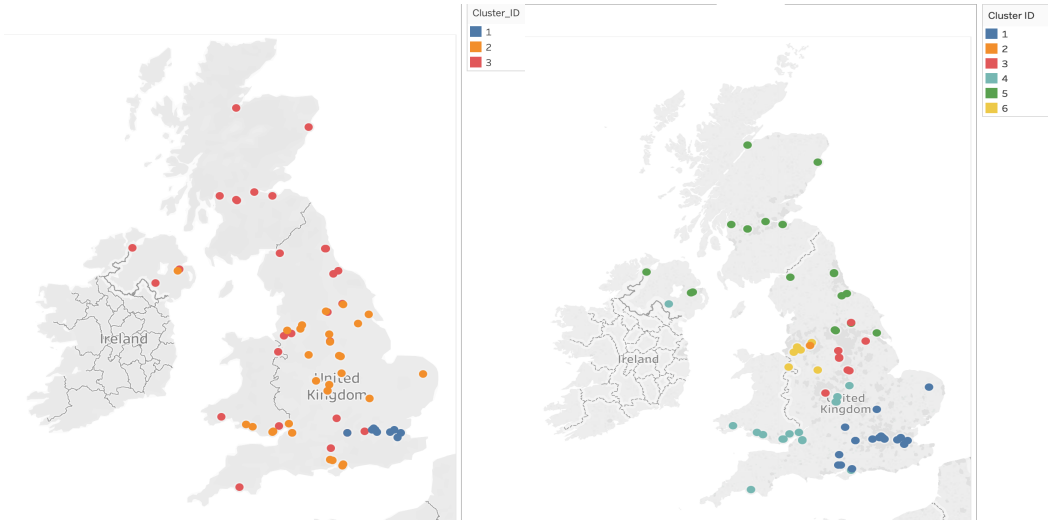
B.1 Clustering PM_{10} Dataset

The total number of stations that measure PM_{10} is 75 stations. The optimal number of clusters for DTW is K=3, and K=6 with SBD. The average silhouette width for each clustering algorithm is shown in Figure B.3. The results of clustering the PM_{10} dataset using both DTW and SBD are shown in Figure B.4.



a b

Figure B.3: Average silhouette widths for 2 to 15 clusters with PM_{10} dataset. (A) PAM+DTW and (B) PAM+SBD.

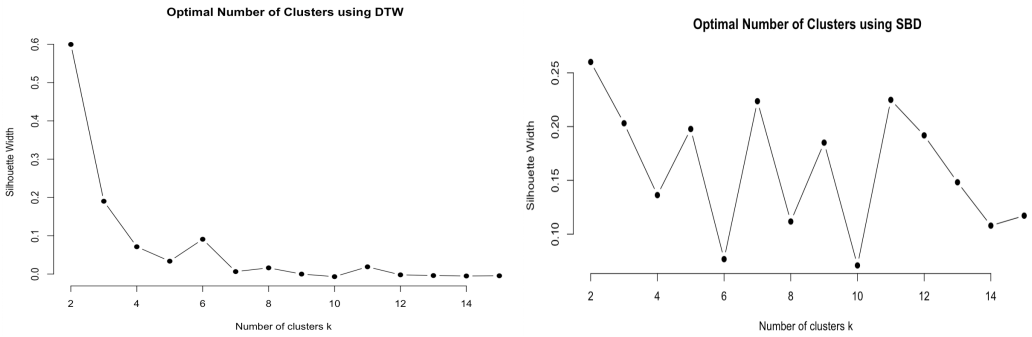


a b

Figure B.4: Geographical distribution of PAM clustering results on PM_{10} dataset. (A) clustering results based on PAM+DTW and (B) clustering results based on PAM+SBD.

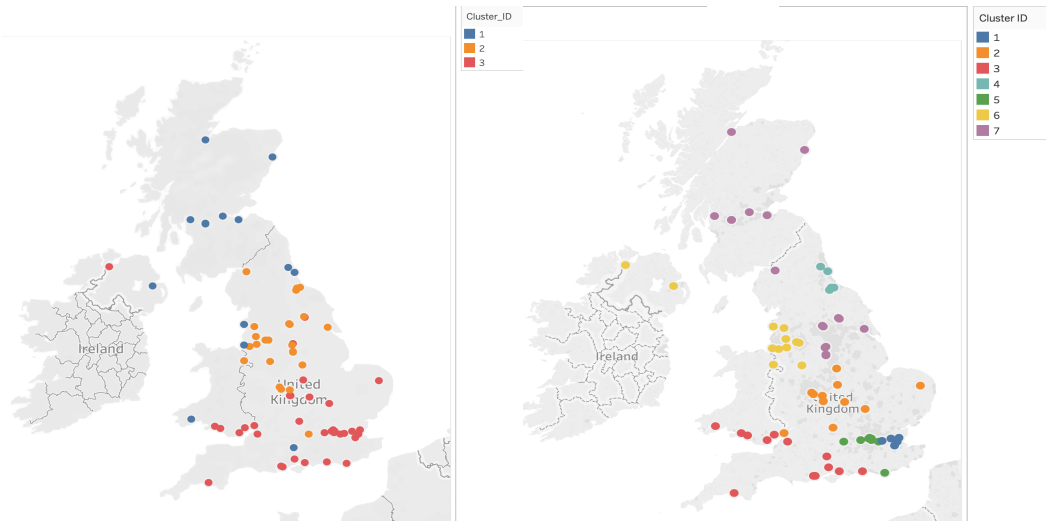
B.2 Clustering $PM_{2.5}$ Dataset

The total number of stations that measure $PM_{2.5}$ is 77 stations. The optimal number of clusters using SBD is K=7 and K=3 with DTW. The average silhouette width for each clustering algorithm is shown in Figure B.5. The results of clustering this dataset using both DTW and SBD are shown in Figure B.6.



a b

Figure B.5: Average silhouette widths for 2 to 15 clusters with $PM_{2.5}$ dataset. (A) PAM+DTW and (B) PAM+SBD

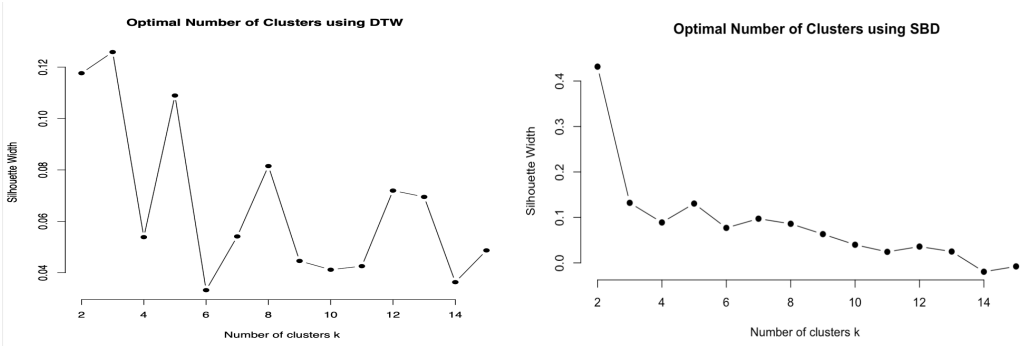


a b

Figure B.6: Geographical distribution of PAM clustering results on $PM_{2.5}$ dataset. (A) clustering results based on PAM+DTW and (B) clustering results based on PAM+SBD.

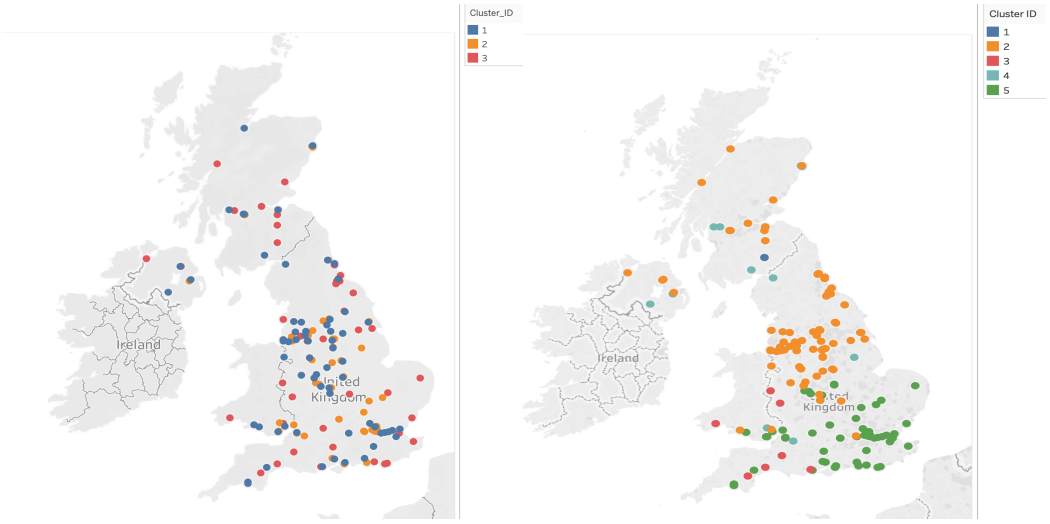
B.3 Clustering NO_2 Dataset

The total number of stations that measure NO2 is 157 stations. The optimal number of clusters using SBD is K=5, and k=3 using DTW. The average silhouette width for each clustering algorithm is shown in Figure B.7. The results of clustering this dataset using both DTW and SBD are shown in Figure B.8.



a b

Figure B.7: Average silhouette widths for 2 to 15 clusters with NO_2 dataset.(A) PAM+DTW and (B) PAM+SBD.



a b

Figure B.8: Geographical distribution of PAM clustering results on NO_2 dataset. (A) clustering results based on PAM+DTW and (B) clustering results based on PAM+SBD.

Appendix C

Results of pollutant imputation based on clustering for ozone dataset.

Table C.1: The average rank and RMSE and Standard Deviation (Std) between observed and imputed TS four imputation models and two dataset MICE dataset (left table) and SMA dataset (right table), including number of stations contributed in each imputation from exploratory experiment.

	MICE Dataset								**SMA dataset**							
	CA		CM		1NN		2NN		CA		CM		1NN		2NN	
Station	RMSE	Rank	RMSE	Rank	RMSE	Rank	RMSE	Rank	RMSE	Rank	RMSE	Rank	RMSE	Rank	RMSE	Rank
Aberdeen	18.66	4	15.909	3	12.312	1	13.433	2	13.469	2	14.034	4	12.619	1	13.645	3
Aston.Hill	11.268	2	7.702	1	17.275	3	17.905	4	8.22	2	7.955	1	17.215	3	17.781	4
Auchencorth.Moss	5.801	4.33			5.04	2 .33	5.209	3 .33	5.141	3.33			5.11	2.33	5.404	4.33
Barnsley.Gawber	6.059	3	8.03	4	4.928	1	5.299	2	7.769	3	13.952	4	5.043	1	5.375	2
Belfast.Centre	13.257	2	19.8	4	14.26	3	10.046	1	22.271	4	20.638	3	14.306	2	10.105	1
Birmingham.Acocks.Green	6.382	4	6.096	2.5	6.096	2.5	4.566	1	6.187	2	6.235	3.5	6.235	3.5	4.752	1
Blackpool.Marton	12.839	3	13.879	4	10.597	1	10.609	2	13.225	3	14.523	4	11.229	2	11.177	1
Bournemouth	10.976	1	13.307	3	13.95	4	11.605	2	9.743	2.33			14.179	4.33	11.942	3.33
Brighton.Preston.Park	9.033	2	12.222	3	15.037	4	7.524	1	15.113	3.83			15.113	3.833	7.811	2.33
Bristol.St.Paul.s	8.404	3	9.205	4	8.27	2	6.01	1	9.436	1.5	10.653	3.5	10.653	3.5	9.436	1.5
Bush.Estate	7.792	4	5.04	1.5	5.04	1.5	6.807	3	7.23	4	5.11	1.5	5.11	1.5	6.827	3
Canterbury	7.268	4	6.714	1.5	6.714	1.5	6.859	3	8.87	4	8.787	3	7.945	2	7.84	1
Cardiff.Centre	9.091	2	9.942	4	9.729	3	7.472	1	8.855	1.5	9.966	3.5	9.966	3.5	8.855	1.5
Coventry.Allesley	4.04	2.33			6.417	4 .33	4.24	3 .33	4.184	2.33			6.571	4.33	4.405	3.33
Cwmbran	10.59	3	10.853	4	9.729	2	8.29	1	8.823	2.83			9.966	4.33	8.823	2.83
Edinburgh.St.Leonards	7.91	1	11.087	2	11.625	4	11.362	3	7.91	1	11.411	3	11.63	4	11.369	2
Eskdalemuir	7.681	3	6.58	2	8.006	4	6.134	1	6.904	3	6.683	2	8.671	4	6.51	1
Fort.William	13.301	3	10.512	2	15.579	4	8.757	1	11.382	3	10.894	2	15.51	4	9.047	1
Glasgow.Townhead	9.78	1	17.278	2	19.032	4	18.886	3	14.01	1	17.395	2	19.097	4	18.966	3
Glazebury	5.208	2.33			7.519	3 .33	8.575	4 .33	7.859	2.33			9.998	3.33	11.749	4.33
High.Muffles	13.069	1	13.078	2	19.747	4	17.708	3	13.756	2	13.024	1	20.087	4	18.07	3
Hull.Freetown	10.064	2	10.601	3	17.47	4	8.136	1	9.488	2	14.889	3	17.847	4	8.61	1
Leamington.Spa	6.637	3	6.417	1.5	6.417	1.5	7.326	4	6.604	3	6.571	1.5	6.571	1.5	7.424	4
Leeds.Centre	12.31	4	11.994	3	9.347	1	9.65	2	10.884	3	18.23	4	9.359	1	9.627	2
Leicester.University	7.961	1	8.026	2	16.528	4	8.884	3	7.715	1	8.144	2	16.582	4	8.801	3
Lerwick	16.902	1	18.252	2	27.854	4	18.29	3	20.555	3	19.117	2	28.664	4	19.077	1
Liverpool.Speke	5.812	2	6.53	4	5.431	1	6.28	3	5.959	1	8.863	4	7.662	3	6.972	2
London.Bloomsbury	8.065	1	12.726	3	19.514	4	8.472	2	11.069	2	11.526	3	19.14	4	8.464	1
London.Eltham	9.404	3	5.987	1	10.83	4	6.749	2	7.325	2.83			11.526	4.33	7.325	2.83
London.Haringey.Priory.Park.South	9.45	2 .33			12.726	3 .33	19.353	4 .33	10.486	2.33			13.005	3.33	19.081	4.33
London.Harlington	4.599	1	9.224	3	11.764	4	5.688	2	5.217	1	8.788	3	12.099	4	6.236	2
London.Hillingdon	12.454	2	17.474	4	11.764	1	14.22	3	11.539	1	17.177	4	12.099	2	14.44	3
London.Marylebone.Road	22.996	2	27.885	4	19.514	1	23.01	3	23.028	3	27.244	4	19.14	1	22.76	2
London.N..Kensington	8.722	2	4.472	1	27.312	4	18.591	3	10.691	2	6.042	1	27.279	4	18.841	3
Lough.Navar	16.472	4	14.274	3	14.26	2	11.639	1	14.418	4	14.363	3	14.306	2	11.874	1
Lullington.Heath	13.077	1	13.785	2	15.037	3	16.926	4	15.113	2	15.113	2	15.113	2	16.93	4
Mace.Head	19.08	1	20.117	2	26.948	3	31.159	4	18.389	2	17.145	1	27.269	3	31.415	4
Manchester.Piccadilly	19.852	4	16.949	1.5	16.949	1.5	18.062	3	20.561	4	19.444	2.5	19.444	2.5	19.131	1
Market.Harborough	12.608	2	12.573	1	16.528	3	17.071	4	13.189	2	12.74	1	16.582	3	17.325	4
Middlesbrough	8.4	1	15.369	4	9.386	2	13.688	3	7.358	1	11.942	2.5	11.942	2.5	14.852	4
Narberth	7.689	2 .33			13.204	3 .33	16.109	4 .33	6.264	2.33			13.467	3.33	16.233	4.33
Newcastle.Centre	11.037	2	18.799	4	11.853	3	9.489	1	7.56	1	14.219	3.5	14.219	3.5	10.434	2
Norwich.Lakenfields	8.997	1	11.746	2	12.01	3	13.521	4	9.117	1	10.05	2	12.081	3	13.463	4
Nottingham.Centre	9.355	3	9.108	2	7.257	1	11.227	4	9.929	3	9.402	2	7.539	1	11.686	4
Peebles	12.322	4 .33			8.018	3 .33	7.824	2 .33	10.881	4.33			8.376	3.33	8.15	2.33
Plymouth.Centre	12.839	2	15.254	4	14.463	3	10.134	1	10.468	3	9.954	1	14.599	4	10.464	2
Port.Talbot.Margam	11.313	2	13.204	4	11.895	3	10.316	1	15.884	4	13.467	3	12.127	2	10.536	1
Portsmouth	13.322	2	15.661	4	13.95	3	10.469	1	13.639	2	14.179	3.5	14.179	3.5	10.648	1
Preston	5.416	1	6.632	3	10.597	4	6.414	2	5.633	1	8.896	3	11.229	4	7.04	2
Reading.New.Town	11.404	2	6.689	1	19.116	4	14.145	3	12.848	2	7.558	1	19.265	4	14.306	3
Rochester.Stoke	6.877	2 .33			7.937	3 .33	10.091	4 .33					8.551	4.5	10.332	5.5
Sheffield.Devonshire.Green	6.717	2	8.935	4	4.928	1	6.768	3	7.35	3	9.963	4	5.043	1	6.786	2
Sibton	9.776	2	10.699	3	12.01	4	9.019	1	7.646	1	8.549	2	12.081	4	9.179	3
Southend.on.Sea	6.736	2	7.937	3.5	7.937	3.5	4.734	1	9.03	4	8.7	3	8.551	2	6.026	1
St.Osyth	5.734	1	7.276	3	8.371	4	6.765	2	5.809	2.33			8.7	4.33	6.868	3.33
Stoke.on.Trent.Centre	6.381	1	7.313	2	10.103	4	8.099	3	6.528	1	7.561	2	10.352	4	8.217	3
Strathvaich	9.399	1	11.565	2	15.579	3	17.369	4	11.718	2	11.531	1	15.51	3	17.275	4
Sunderland.Silksworth	8.745	1	11.382	3	11.853	4	9.664	2	12.221	2.33			14.219	4.33	12.275	3.33
Thurrock	12.325	3	13.246	4	5.625	1	7.959	2	7.311	3	6.384	1.5	6.384	1.5	8.559	4
Walsall.Woodlands	7.599	3	7.849	4	5.945	1	6.26	2	7.943	4	7.888	3	6.116	1	6.332	2
Weybourne	14.396	3	14.269	2	15.905	4	11.046	1	12.093	2	12.569	3	15.803	4	11.025	1
Wicken.Fen	6.682	2	9.231	4	9.072	3	5.29	1	6.353	2	8.497	3	9.13	4	5.785	1
Wigan.Centre	6.4	2	7.519	3.5	7.519	3.5	5.905	1	6.87	1	9.998	3.5	9.998	3.5	7.313	2
Wirral.Tranmere	6.423	3	8.253	4	5.431	2	5.113	1	8.124	3	10.805	4	7.662	2	7.52	1
Yarner.Wood	9.067	2	9.052	1	14.463	3	16.762	4	8.99	1	9.214	2	14.599	3	16.767	4
Average Rank	**2.25**		2.78		2.86		2.41		2.37		2.61		3.05		2.53	
Average Error (RMSE)	**10.003**		11.41		12.116		10.784		10.315		11.692		12.641		11.266	
Standard deviation (Std)	**3.901**		4.544		5.417		5.272		4.218		4.404		5.263		5.121	
Station contributing	65		58		65		65		64		52		65		65	

Appendix D

Univariate TS clustering imputation results

Table D.1: The RMSE between observed and imputed TS using univariate clustering imputation for stations that measure O_3

	Site	model 1 (CA)	model 2 (CA+ENV)	model 3 (CA+REG)	model 4 (1NN)	model 5 (1NN.ENV)	model 6 (2NN)	model 7 (2NN.ENV)	model 8 (Median)	model 9 (Median.ENV)
1	Aberdeen	19.303	17.406	19.303	22.026	23.017	21.212	19.089	18.802	18.865
2	Aston.Hill	17.473	14.021	17.473	23.324	27.953	22.019	20.143	15.697	16.470
3	Auchencorth.Moss	8.843	9.927	12.427	7.982	7.982	8.157	7.136	7.008	6.769
4	Barnsley.Gawber	11.149	11.111	9.787	9.919	9.919	11.211	9.174	9.208	8.730
5	Belfast.Centre	22.799	18.812	22.877	22.877	17.856	16.649	18.745	18.860	18.005
6	Birmingham.A4540.Roadside	13.500	20.289	13.814	15.888	25.449	16.436	18.087	13.910	14.909
7	Birmingham.Acocks.Green	10.747	10.680	8.694	15.888	9.716	11.114	7.923	9.681	8.614
8	Blackpool.Marton	16.462	16.540	15.870	14.122	14.122	14.148	14.148	14.965	14.965
9	Bournemouth	17.204	18.681	17.204	18.842	18.842	13.244	17.138	16.996	17.475
10	Brighton.Preston.Park	13.210	13.738	12.225	19.293	15.163	12.523	14.065	11.718	12.697
11	Bristol.St.Paul.s	14.024	14.233	17.394	13.977	13.977	11.357	11.357	11.847	11.847
12	Bush.Estate	10.648	12.051	12.678	7.982	7.982	9.193	9.451	8.503	8.786
13	Canterbury	14.687	16.640	14.622	12.455	14.476	12.135	15.036	13.395	14.469
14	Cardiff.Centre	14.099	14.647	13.023	14.792	14.792	12.106	12.106	11.713	11.713
15	Chilbolton.Observatory	15.674	14.801	13.124	19.427	17.816	18.201	14.911	13.862	13.394
16	Coventry.Allesley	10.096	9.718	7.975	9.957	9.957	7.668	7.668	7.977	7.977
17	Cwmbran	15.338	15.482	13.943	14.792	14.792	12.810	12.810	12.476	12.476
18	Eskdalemuir	10.662	12.012	13.058	13.058	11.289	10.408	11.710	10.252	10.189
19	Exeter.Roadside	18.748	20.289	18.521	29.358	20.289	23.559	21.189	18.935	19.392
20	Fort.William	16.725	16.725	20.045	20.045	26.997	14.191	27.650	15.048	17.037
21	Glasgow.Townhead	20.363	14.798	22.068	22.332	21.180	22.068	14.847	21.474	17.395
22	Glazebury	10.677	10.677	8.513	11.397	20.419	11.630	23.228	8.709	10.315
23	High.Muffles	17.049	15.688	17.049	22.923	14.606	20.243	18.793	16.358	15.922
24	Hull.Freetown	14.932	14.524	15.100	21.261	17.017	13.345	15.369	13.359	14.475
25	Ladybower	13.811	12.701	13.811	17.803	20.419	17.018	13.764	12.502	12.772
26	Leamington.Spa	10.429	10.405	9.401	9.957	9.957	10.095	10.095	9.110	9.110
27	Leeds.Centre	15.020	14.768	12.620	13.524	13.524	17.408	13.409	13.822	12.765
28	Leicester.University	11.716	11.683	10.850	19.208	11.505	11.774	9.713	11.248	9.952
29	Leominster	13.154	13.154	13.743	23.324	21.670	15.393	18.451	12.562	13.082
30	Liverpool.Speke	10.908	14.529	9.279	9.389	17.070	9.866	14.529	8.556	12.811
31	London.Bloomsbury	17.792	16.040	11.114	23.067	14.287	11.525	14.556	10.935	13.656
32	London.Eltham	10.783	10.783	13.505	14.370	21.670	9.662	23.877	8.871	11.321
33	London.Haringey.Priory.Park.South	10.867	10.131	13.231	16.045	16.045	22.283	10.748	11.581	10.250
34	London.Harlington	14.915	14.915	10.201	15.630	27.699	9.741	23.633	9.655	14.159
35	London.Hillingdon	23.094	21.786	15.366	15.630	21.127	17.412	19.201	17.314	19.207
36	London.Marylebone.Road	32.922	32.922	25.305	23.067	25.449	26.183	25.020	26.758	28.452
37	London.N..Kensington	10.685	9.651	12.089	30.677	14.287	20.966	8.690	12.078	8.523
38	Lough.Navar	19.908	22.764	22.877	22.877	21.097	21.676	20.563	20.759	19.841
39	Lullington.Heath	20.059	15.699	18.395	19.293	19.688	20.417	16.500	17.985	17.000
40	Manchester.Piccadilly	20.267	20.878	22.016	21.145	23.117	19.941	22.117	20.522	21.454
41	Manchester.Sharston	12.231	12.231	11.506	21.145	21.145	13.461	13.461	11.425	11.425
42	Market.Harborough	16.180	11.762	16.180	19.208	14.576	14.060	10.172	13.696	12.462
43	Middlesbrough	14.409	17.406	14.858	13.466	17.070	16.262	17.437	11.896	14.227
44	Narberth	15.933	13.078	15.933	17.995	12.051	19.761	10.652	14.265	12.952
45	Newcastle.Centre	16.181	16.294	14.858	14.904	14.904	13.271	17.993	13.121	13.672
46	Northampton.Spring.Park	14.740	14.339	17.191	13.050	16.965	11.697	15.202	13.008	15.173
47	Norwich.Lakenfields	12.525	14.875	10.936	15.030	14.349	15.774	14.599	10.714	12.547
48	Nottingham.Centre	12.221	12.010	13.377	11.505	11.505	14.071	10.155	11.643	10.680
49	Peebles	13.583	18.122	13.058	12.754	21.180	12.598	20.910	11.741	16.026
50	Plymouth.Centre	18.866	19.623	17.102	19.091	18.971	13.490	17.562	15.631	17.156
51	Port.Talbot.Margam	17.625	19.302	15.351	16.548	19.600	15.351	20.982	15.189	17.614
52	Portsmouth	13.549	12.660	14.577	14.669	14.669	14.478	13.741	12.746	12.551
53	Preston	10.910	11.190	8.346	14.122	14.122	9.785	9.785	8.540	8.540
54	Reading.New.Town	11.538	11.066	13.250	23.021	23.021	17.506	15.039	10.718	10.748
55	Rochester.Stoke	13.260	13.194	13.982	11.847	13.040	13.116	13.835	11.368	11.718
56	Sheffield.Devonshire.Green	11.734	11.409	11.024	17.803	9.919	11.393	10.790	10.455	9.327
57	Sibton	17.240	11.295	12.100	15.030	13.164	11.970	10.197	12.031	11.106
58	Southampton.Centre	14.555	13.483	16.010	14.669	14.669	14.075	14.146	13.710	13.570
59	Southend.on.Sea	9.732	11.387	10.995	11.847	12.947	9.061	10.343	8.277	9.096
60	St.Osyth	13.371	11.284	10.316	12.538	13.040	11.337	10.252	10.345	9.784
61	Stoke.on.Trent.Centre	10.164	10.237	12.214	14.214	15.250	12.197	11.741	10.341	10.527
62	Strathvaich	16.272	14.769	20.045	20.045	16.591	20.305	15.521	17.613	14.426
63	Sunderland.Silksworth	14.499	16.117	14.499	14.904	14.904	12.106	16.576	12.385	12.786
64	Thurrock	11.814	10.721	17.795	10.057	12.947	11.174	10.320	10.851	11.107
65	Walsall.Woodlands	11.647	11.548	10.739	18.442	10.223	12.617	10.101	11.484	10.139
66	Weybourne	22.028	15.650	17.983	19.389	14.433	15.351	15.216	17.120	15.966
67	Wicken.Fen	13.312	12.203	11.958	14.683	14.576	11.570	12.128	11.134	11.249
68	Wigan.Centre	12.924	12.958	10.531	11.397	11.108	9.615	14.230	10.427	11.261
69	Wirral.Tranmere	11.543	11.866	10.340	9.389	14.078	9.580	10.898	9.536	10.174
70	Yarner.Wood	19.009	16.334	19.009	29.358	14.891	22.485	13.442	17.922	16.350
	Average RMSE	14.776	14.496	14.381	16.729	16.345	14.521	14.772	13.062	13.311
	Std	4.060	3.890	3.845	5.075	4.779	4.323	4.604	3.751	3.755

Table D.2: The RMSE between observed and imputed TS using univariate clustering imputation for stations that measure NO_2

	Site	model 1 (CA)	model 2 (CA+ENV)	model 3 (CA+REG)	model 4 (1NN)	model 5 (1NN.ENV)	model 6 (2NN)	model 7 (2NN.ENV)	model 8 (Median)	model 9 (Median.ENV)
1	Aberdeen	15.656	15.786	17.420	29.015	17.420	29.743	16.399	14.812	16.164
2	Aberdeen.Union.Street.Roadside	19.448	19.063	20.172	20.172	20.172	18.484	19.434	16.530	17.241
3	Aberdeen.Wellington.Road	23.587	22.938	20.172	20.172	20.172	23.632	23.440	20.971	20.990
4	Armagh.Roadside	18.317	18.813	33.238	33.238	33.238	20.966	18.502	21.772	20.090
5	Aston.Hill	19.578	4.777	19.836	7.981	14.394	11.974	9.068	11.194	11.854
6	Ballymena..Antrim.Road	13.441	14.196	12.448	12.611	35.085	11.615	20.904	11.469	11.927
7	Ballymena..Ballykeel	10.006	9.639	9.077	12.611	18.465	13.404	10.340	8.956	9.502
8	Barnsley.Gawber	7.589	7.421	10.735	11.929	11.929	18.000	12.721	9.737	9.081
9	Bath.Roadside	15.854	15.717	25.045	29.760	29.760	24.441	19.736	19.087	17.110
10	Belfast.Centre	17.078	18.385	17.695	31.714	18.465	14.922	20.113	15.491	18.083
11	Belfast.Stockman.s.Lane	27.759	26.993	33.238	31.714	35.085	32.374	33.452	30.023	31.705
12	Billingham	9.712	11.574	9.104	12.812	10.809	10.556	9.670	9.026	9.038
13	Birkenhead.Borough.Road	11.557	11.757	10.467	10.394	19.220	11.010	17.262	9.657	11.059
14	Birmingham.A4540.Roadside	11.809	12.797	10.548	19.237	10.501	12.085	11.724	9.944	9.984
15	Birmingham.Acocks.Green	9.135	8.782	7.495	19.237	9.544	18.594	7.439	8.397	7.298
16	Blackburn.Accrington.Road	7.438	13.599	7.968	8.318	12.362	8.729	12.597	7.470	8.754
17	Blackpool.Marton	9.717	9.799	13.369	12.975	12.975	11.413	10.217	10.950	10.603
18	Borehamwood.Meadow.Park	10.472	10.182	11.114	7.789	7.789	21.907	9.911	9.813	9.440
19	Bournemouth	11.627	10.786	10.122	14.811	21.300	20.013	15.455	11.158	11.204
20	Bradford.Mayo.Avenue	32.026	24.698	29.922	27.127	20.725	23.145	23.734	27.537	25.891
21	Brighton.Preston.Park	10.530	9.959	8.908	28.225	9.934	20.827	8.485	10.310	8.612
22	Bristol.St.Paul.s	10.187	10.226	11.877	29.303	10.991	24.023	11.765	9.861	10.282
23	Bristol.Temple.Way	31.606	22.332	36.174	29.303	29.760	26.222	27.726	29.609	27.413
24	Burton.on.Trent..Horninglow	8.261	7.742	9.169	22.144	9.530	13.775	7.973	7.709	7.491
25	Bury.Whitefield.Roadside	12.413	12.885	10.305	11.259	12.296	10.561	9.590	10.030	9.405
26	Bush.Estate	16.385	16.385	20.377	54.774	6.600	35.168	6.342	20.448	15.341
27	Cambridge.Roadside	14.159	15.253	13.409	21.381	13.409	12.888	17.335	10.280	13.354
28	Camden.Kerbside	28.295	27.783	32.757	47.877	47.877	24.260	26.842	24.493	26.428
29	Cannock.A5190.Roadside	10.557	12.663	12.536	12.130	16.021	9.608	15.724	9.666	10.247
30	Canterbury	11.705	10.974	9.464	7.519	11.073	9.815	13.188	9.507	10.327
31	Cardiff.Centre	8.746	8.533	8.063	17.187	9.145	11.146	9.055	7.704	7.735
32	Cardiff.Newport.Road	17.035	13.195	19.251	17.187	47.435	16.042	31.590	16.278	13.215
33	Carlisle.Roadside	11.317	11.273	11.317	18.298	18.298	13.313	19.477	10.500	10.259
34	Charlton..Mackrell	16.989	5.199	16.384	39.022	6.798	39.348	4.121	17.563	6.746
35	Chatham.Roadside	17.725	18.615	17.104	15.568	15.375	11.974	19.559	14.506	16.161
36	Chepstow..A48	13.561	13.596	39.904	23.496	28.248	23.189	28.398	13.328	20.677
37	Chesterfield.Loundsley.Green	10.391	10.616	6.020	10.829	14.589	11.300	16.284	8.908	10.332
38	Chesterfield.Roadside	8.932	14.634	13.350	10.829	23.535	8.951	22.085	8.618	13.602
39	Chilbolton.Observatory	13.564	5.545	10.946	33.535	6.798	27.664	6.160	12.817	5.158
40	Christchurch.Barrack.Road	12.972	17.935	12.656	14.811	29.079	16.774	23.416	12.527	15.919
41	Coventry.Allesley	9.495	9.015	8.165	13.653	9.901	7.992	8.410	7.603	8.137
42	Coventry.Binley.Road	12.577	13.581	11.832	13.653	16.540	14.507	9.574	10.324	10.484
43	Crumlyn	11.698	10.667	9.866	10.853	10.853	34.379	10.373	11.057	9.733
44	Derby.St..Alkmund.s.Way	11.519	11.924	13.077	22.144	13.048	15.004	15.948	12.455	11.491
45	Derry.Rosemount	12.916	12.174	16.198	19.120	11.941	14.640	15.993	14.392	12.845
46	Dewsbury.Ashworth.Grove	9.016	9.392	6.312	27.127	11.570	15.573	7.220	6.292	6.747
47	Doncaster.A630.Cleveland.Street	13.574	9.776	10.509	13.466	15.490	11.368	10.816	9.839	9.626
48	Dumbarton.Roadside	10.979	13.995	10.029	25.019	25.019	19.419	19.419	12.160	12.160
49	Dumfries	17.447	17.395	17.447	35.285	18.298	25.088	16.739	17.289	17.256
50	Dundee.Mains.Loan	10.920	10.149	17.420	13.660	13.660	28.912	9.612	11.138	9.709
51	Eastbourne	11.992	11.237	9.149	7.685	9.934	5.022	12.947	8.696	10.229
52	Edinburgh..Nicolson.Street	29.715	29.043	30.101	40.976	32.013	47.554	29.359	30.435	27.658
53	Edinburgh.St.Leonards	9.647	10.233	8.803	40.976	16.673	15.683	14.071	8.464	9.676
54	Eskdalemuir	20.942	5.646	20.942	35.285	6.600	19.907	4.942	18.733	7.901
55	Exeter.Roadside	16.223	16.464	16.898	30.439	19.973	28.418	14.559	16.081	14.614
56	Fort.William	14.334	14.334	14.334	17.285	11.036	22.107	11.985	14.334	14.334
57	Glasgow..Great.Western.Road	18.014	15.085	16.658	41.164	41.164	20.265	22.463	13.642	13.661
58	Glasgow..High.Street	19.070	16.071	17.506	11.606	40.379	19.070	21.368	14.969	14.528
59	Glasgow..Kerbside	37.774	37.140	35.823	45.561	40.379	42.655	40.373	37.795	36.909
60	Glasgow..Townhead	11.892	13.599	9.534	11.606	13.151	26.413	16.710	8.756	12.469
61	Glazebury	9.639	13.542	12.638	15.081	12.291	12.153	13.074	10.313	11.895
62	Grangemouth	9.835	8.530	10.020	6.580	6.580	24.526	8.167	8.445	7.332
63	Grangemouth.Moray	10.285	9.544	9.905	6.580	6.580	22.956	8.832	8.401	8.144
64	Greenock.A8.Roadside	19.186	19.040	31.171	25.019	25.019	21.498	21.498	18.039	18.039
65	Hafod..y..rys.Roadside	41.024	40.202	39.904	61.327	47.435	57.779	42.258	43.771	41.123
66	Haringey.Roadside	22.842	17.825	15.073	19.227	22.593	14.150	20.500	16.025	15.259
67	Hartlepool.St.Abbs.Walk	10.572	10.115	11.956	10.855	9.884	9.449	12.093	9.509	10.009
68	High.Muffles	18.258	5.342	23.563	15.394	5.766	15.172	8.621	14.294	8.691
69	Honiton	15.367	14.604	14.315	26.750	15.904	12.372	14.617	14.527	14.018
70	Horley	11.404	11.404	11.113	17.552	16.926	23.223	22.181	11.543	11.456
71	Hull.Freetown	18.558	14.637	9.930	10.755	14.547	7.377	12.420	9.622	9.370
72	Hull.Holderness.Road	16.248	17.404	11.259	10.755	15.573	14.946	15.024	10.599	13.422
73	Immingham.Woodlands.Avenue	10.354	10.774	14.368	14.547	14.547	17.185	11.091	11.763	11.049
74	Inverness	23.718	24.625	23.718	17.285	30.429	16.591	30.638	23.588	24.170
75	Ladybower	16.789	4.401	12.849	21.372	12.291	28.398	8.155	16.139	9.480
76	Leamington.Spa	9.170	8.383	7.460	7.008	9.901	9.750	8.596	7.153	7.620
77	Leamington.Spa.Rugby.Road	7.733	19.209	5.467	7.008	16.540	6.254	17.456	5.543	11.862
78	Leeds.Centre	14.099	11.591	12.650	13.576	11.570	10.587	14.257	10.209	10.403
79	Leeds.Headingley.Kerbside	18.777	12.502	15.579	13.576	20.725	10.704	12.671	13.479	12.345
80	Leicester.A594.Roadside	13.836	14.242	13.836	18.537	14.226	25.537	12.292	13.714	12.755
81	Leicester..University	11.313	11.466	17.373	18.537	12.185	8.955	10.203	10.595	10.783
82	Leominster	14.124	14.124	13.185	7.981	13.498	5.258	16.495	11.425	12.739
83	Lincoln..Canwick.Road	18.189	17.981	20.108	28.438	19.931	28.448	18.372	19.106	17.944
84	Liverpool.Speke	9.154	10.101	10.681	25.872	11.601	14.711	10.101	10.555	9.596
85	London..Bexley	9.483	10.179	12.258	10.179	10.179	7.873	14.600	8.243	9.984
86	London..Bloomsbury	20.263	21.851	11.579	60.794	10.066	29.223	11.179	11.568	12.551
87	London..Eltham	8.058	10.179	18.684	10.179	10.179	18.989	6.281	10.825	9.607
88	London..Haringey..Priory.Park.South	9.113	9.235	13.081	19.227	19.279	27.583	14.112	11.049	11.264
89	London..Harlington	16.527	24.017	12.959	26.836	24.017	14.532	22.382	11.258	21.706
90	London..Hillingdon	28.026	32.521	33.374	26.836	29.916	27.752	30.920	25.835	27.869
91	London..Marylebone.Road	59.696	59.254	45.416	60.794	47.877	52.978	49.733	55.816	50.983
92	London..N..Kensington	14.032	14.902	8.695	33.020	14.098	46.663	11.143	8.602	8.558
93	London..Westminster	14.932	16.498	6.658	10.066	10.066	34.871	7.148	6.537	6.725
94	Lullington..Heath	16.227	5.571	13.763	7.685	10.664	10.592	7.613	9.617	9.106
95	Luton.A505.Roadside	30.837	21.759	33.596	32.314	28.494	29.159	24.145	30.519	27.201
96	Manchester..Piccadilly	18.184	19.113	15.629	14.776	14.776	13.424	16.529	16.249	16.741
97	Manchester.Sharston	10.794	10.794	9.529	16.838	16.926	11.929	10.979	9.543	10.416
98	Market.Harborough	15.343	5.527	13.322	19.904	5.476	26.298	4.811	14.550	5.073
99	Middlesbrough	9.256	9.760	10.749	10.809	10.809	8.688	9.359	8.375	8.609
100	Narberth	19.644	7.462	19.602	25.479	3.082	19.107	2.824	18.307	7.265
101	Newcastle.Centre	14.576	16.742	12.934	19.199	18.790	10.899	18.379	12.039	15.808
102	Newcastle.Cradlewell.Roadside	30.079	27.098	28.635	19.199	24.585	21.064	26.125	26.289	27.357
103	Newport	8.319	8.439	7.381	10.853	10.853	24.616	8.118	7.313	7.355
104	Northampton.Spring.Park	11.196	10.151	8.199	9.130	15.786	8.199	12.073	7.591	9.695
105	Norwich.Lakenfields	11.801	10.851	13.374	7.617	11.760	9.223	14.225	10.184	11.754
106	Nottingham.Centre	13.568	10.933	13.045	14.885	12.862	14.447	11.149	12.162	9.039
107	Nottingham.Western.Boulevard	12.760	13.231	10.583	14.885	13.048	11.210	14.595	10.842	11.270
108	Oldbury.Birmingham.Road	11.122	11.743	9.377	10.501	10.501	12.892	10.499	9.105	8.964
109	Oxford.Centre.Roadside	16.374	16.543	19.033	28.897	21.313	22.906	17.949	17.433	15.915
110	Oxford.St.Ebbes	8.936	7.902	7.654	28.897	14.867	23.322	8.886	9.114	8.300
111	Peebles	15.451	14.774	15.451	5.550	16.673	26.960	18.921	15.027	15.114
112	Plymouth.Centre	10.914	11.089	10.153	12.551	15.690	12.545	11.474	9.845	10.231
113	Plymouth.Tavistock.Road	22.499	23.410	21.837	12.551	19.973	14.058	16.210	18.896	20.198
114	Port.Talbot.Margam	11.782	24.017	10.245	16.840	13.342	13.052	12.671	12.077	11.263
115	Portsmouth	9.949	9.829	9.819	18.308	16.657	15.639	10.328	9.471	9.492
116	Portsmouth.Anglesea.Road	15.650	16.234	14.234	18.308	21.866	14.891	15.623	13.358	14.722
117	Preston	8.621	8.630	7.820	8.318	12.975	9.332	10.776	7.822	8.378
118	Reading..London.Road	16.654	12.133	19.847	12.496	21.313	22.168	25.177	13.668	12.587
119	Reading..New..Town	9.722	10.744	12.636	12.496	33.883	17.375	16.286	9.553	9.998
120	Rochester..Stoke	11.579	9.191	9.697	9.840	8.011	11.340	8.156	8.485	7.118
121	Salford.Eccles	10.656	10.904	8.667	14.776	14.776	11.008	9.780	8.542	8.743
122	Sandy.Roadside	17.056	18.004	13.409	28.494	28.494	17.002	17.002	16.083	16.083
123	Scunthorpe.Town	8.848	10.944	12.470	12.357	11.163	7.707	9.967	8.752	8.999
124	Shaw.Crompton.Way	13.705	14.440	13.705	12.296	12.296	12.100	12.549	13.623	13.619
125	Sheffield.Barnsley.Road	22.407	15.051	19.231	19.004	23.535	17.241	18.143	18.610	19.157
126	Sheffield..Devonshire.Green	9.216	9.357	7.945	19.004	13.588	13.837	9.812	7.855	8.457
127	Sheffield.Tinsley	13.670	13.688	12.255	18.046	13.588	12.840	13.758	12.038	12.813
128	Southampton..A33	18.182	18.671	19.300	17.744	21.866	24.059	23.223	17.913	18.766
129	Southampton.Centre	14.402	15.183	17.178	17.744	16.657	12.704	18.025	14.235	16.052
130	Southend.on.Sea	7.995	7.820	7.425	9.840	11.880	7.073	7.610	6.729	7.092
131	Southwark.A2.Old.Kent.Road	27.677	18.398	20.806	21.665	21.665	16.966	38.892	17.839	17.620
132	St..Helens..Linkway	19.279	13.728	16.538	14.567	14.567	13.223	13.418	14.006	13.853
133	St.Osyth	12.196	7.833	13.501	10.564	8.011	7.980	6.627	8.912	7.012
134	Stanford.le.Hope..Roadside	13.632	13.794	13.273	10.417	15.375	13.318	15.731	11.680	11.619
135	Stockton.on..Tees.A1305.Roadside	11.835	13.556	10.634	12.812	9.757	9.292	9.825	10.224	9.722
136	Stockton.on..Tees.Eaglescliffe	11.544	16.344	12.559	9.757	9.757	7.718	11.624	10.357	11.343
137	Stoke.on.Trent..A50.Roadside	27.593	26.948	32.653	38.197	38.694	37.917	32.288	32.359	30.765
138	Stoke.on.Trent.Centre	16.340	15.431	17.894	38.197	12.246	20.292	12.198	17.840	12.331
139	Storrington.Roadside	11.993	15.100	13.455	19.971	19.971	11.636	16.926	10.228	10.951
140	Sunderland.Silksworth	9.465	9.507	10.755	12.914	18.790	21.902	11.513	10.118	10.474
141	Sunderland.Wessington.Way	11.428	11.394	9.488	12.914	24.585	10.578	14.308	9.233	9.691
142	Swansea.Roadside	13.049	16.429	13.272	16.840	51.768	20.266	29.888	13.353	15.454
143	Swindon.Walcot	10.123	9.380	9.447	8.209	8.209	17.331	9.767	9.496	8.530
144	Telford.Hollinswood	8.738	8.934	7.497	8.129	8.129	9.256	9.378	7.493	7.794
145	Thurrock	11.140	11.860	10.720	9.974	11.880	8.102	10.481	10.177	10.531
146	Tower.Hamlets.Roadside	31.324	22.893	23.512	21.665	18.600	26.651	18.601	21.811	21.614
147	Walsall.Woodlands	7.874	7.641	5.949	12.130	8.800	15.120	7.156	7.329	6.030
148	Warrington	8.460	10.453	8.921	22.557	11.601	19.932	10.453	8.993	9.852
149	Wicken.Fen	15.139	4.630	17.390	21.381	5.476	22.099	5.347	15.145	5.796
150	Widnes.Milton.Road	23.146	15.590	20.559	25.872	14.567	23.568	14.410	21.568	17.048
151	Wigan.Centre	8.085	7.750	9.551	21.098	13.515	10.396	10.902	8.851	8.783
152	Wirral.Tranmere	8.566	9.110	9.998	10.394	11.074	7.041	10.705	7.748	9.136
153	Worthing.A27.Roadside	14.000	13.541	16.141	19.971	19.971	23.026	16.863	14.315	14.055
154	Wrexham	9.863	15.244	19.836	11.045	12.892	9.975	17.062	9.498	10.963
155	Yarner.Wood	18.664	4.757	18.664	30.439	4.642	23.968	3.560	16.879	5.405
156	York.Bootham	9.632	9.312	12.803	15.332	18.900	15.568	15.635	10.970	11.654
157	York.Fishergate	14.542	15.413	12.398	15.332	15.891	11.652	12.325	11.195	11.681
	Average RMSE	15.037	14.095	15.150	19.447	17.269	18.052	15.254	13.568	13.204
	Std	7.237	7.029	7.601	11.028	9.485	9.229	7.806	7.056	6.784

Table D.3: The RMSE between observed and imputed TS using univariate clustering imputation for stations that measure PM_{10}

	Site	model 1 (CA)	model 2 (CA+ENV)	model 3 (CA+REG)	model 4 (1NN)	model 5 (1NN.ENV)	model 6 (2NN)	model 7 (2NN.ENV)	model 8 (Median)	model 9 (Median.ENV)
1	Aberdeen	8.435	8.508	8.435	10.681	9.360	9.288	9.362	8.313	8.296
2	Armagh.Roadside	19.508	19.245	18.731	18.905	18.905	18.004	19.241	18.646	18.851
3	Auchencorth.Moss	9.879	6.686	6.781	6.180	6.686	6.271	6.643	5.885	5.855
4	Barnstaple.A39	8.192	8.219	7.759	21.738	10.721	14.096	8.219	8.471	8.184
5	Belfast.Centre	8.059	8.631	8.059	8.506	11.448	10.506	9.287	6.958	8.025
6	Belfast.Stockman.s.Lane	8.786	8.885	8.718	8.506	18.905	11.537	11.427	7.481	8.480
7	Birmingham.A4540.Roadside	7.701	8.183	6.759	6.554	8.087	6.141	7.085	6.363	6.690
8	Birmingham.Ladywood	5.474	5.103	4.077	6.554	5.294	6.170	5.014	4.014	4.120
9	Bristol.St.Paul.s	7.212	6.465	10.850	10.850	6.646	8.155	6.724	7.673	5.721
10	Bristol.Temple.Way	10.392	9.678	10.850	10.850	11.542	10.247	9.566	10.086	9.583
11	Bury.Whitefield.Roadside	6.122	6.248	5.385	9.641	8.157	6.432	7.610	5.169	5.360
12	Camden.Kerbside	7.300	6.623	5.759	8.789	8.789	6.449	10.581	5.731	5.726
13	Cardiff.Centre	13.770	13.633	13.340	14.400	13.801	13.373	12.448	13.169	13.047
14	Cardiff.Newport.Road	8.689	8.606	8.583	14.400	9.132	9.809	7.357	8.446	7.430
15	Carlisle.Roadside	6.841	6.943	8.246	12.541	10.035	9.977	9.289	7.029	6.955
16	Chatham.Roadside	10.239	9.273	11.629	11.516	9.483	9.541	8.542	9.950	9.579
17	Chepstow.A48	7.647	8.102	7.826	9.049	11.542	7.472	9.243	6.817	7.082
18	Chesterfield.Loundsley.Green	5.745	5.774	6.359	6.667	5.769	5.040	5.241	4.708	4.562
19	Chesterfield.Roadside	7.893	7.958	7.068	6.667	9.599	6.389	8.443	6.618	7.234
20	Chilbolton.Observatory	7.758	11.963	6.583	8.748	11.963	7.918	7.458	6.834	7.847
21	Coventry.Binley.Road	7.546	7.308	7.393	8.735	8.735	8.475	7.250	7.347	7.051
22	Derry.Rosemount	10.729	10.953	10.095	10.857	11.448	11.995	9.763	9.107	9.607
23	Ealing.Horn.Lane	16.828	15.909	15.799	18.484	16.070	16.753	14.879	16.413	15.631
24	Edinburgh.St.Leonards	7.214	7.305	4.444	6.180	5.401	4.372	6.774	4.090	5.100
25	Glasgow.High.Street	6.672	7.404	5.697	4.978	8.674	5.645	6.445	4.794	5.382
26	Glasgow.Townhead	6.995	7.158	3.942	4.978	5.401	5.466	5.999	3.963	4.836
27	Grangemouth	7.043	8.999	4.744	5.198	10.603	4.827	8.747	4.548	7.692
28	Greenock.A8.Roadside	9.457	10.149	8.117	8.520	8.674	8.229	9.104	7.962	8.374
29	Hull.Holderness.Road	11.078	10.678	9.371	10.592	9.499	8.572	10.048	9.067	9.276
30	Inverness	8.875	10.093	8.875	10.681	8.836	8.199	7.946	8.960	8.925
31	Leamington.Spa	6.411	5.459	5.233	4.289	5.294	5.329	4.282	4.322	4.444
32	Leamington.Spa.Rugby.Road	6.160	8.238	5.129	4.289	8.735	5.311	7.641	4.664	5.992
33	Leeds.Centre	7.025	7.553	5.964	8.032	7.666	6.430	6.413	5.392	6.264
34	Leeds.Headingley.Kerbside	9.641	9.338	8.719	8.032	10.168	8.153	9.216	8.300	8.780
35	Leicester.A594.Roadside	9.522	8.317	9.522	10.011	8.293	8.163	7.683	8.842	8.540
36	Liverpool.Speke	6.163	6.370	6.266	9.222	5.017	6.061	7.492	5.623	5.414
37	London.Bloomsbury	5.622	6.401	5.953	10.043	5.461	7.561	5.090	5.307	4.708
38	London.Harlington	5.471	5.471	7.567	17.868	22.803	9.273	13.891	6.657	5.918
39	London.Marylebone.Road	9.740	8.513	8.509	10.043	8.789	8.719	9.254	8.572	7.904
40	London.N..Kensington	5.575	4.877	7.885	18.484	5.461	12.035	5.823	7.612	4.877
41	Lough.Navar	10.327	6.686	11.688	10.857	6.686	14.156	6.699	8.827	6.338
42	Middlesbrough	7.188	7.752	7.589	9.273	11.890	8.116	8.662	6.763	7.340
43	Narberth	9.221	9.221	14.857	11.196	8.204	14.857	7.238	10.412	8.845
44	Newcastle.Centre	10.269	10.221	10.268	10.853	10.732	10.036	10.673	9.601	9.891
45	Newcastle.Cradlewell.Roadside	7.834	8.018	8.214	10.853	10.967	8.522	8.802	7.590	7.795
46	Newport	6.492	5.938	7.221	8.961	13.801	9.141	9.052	5.894	6.538
47	Norwich.Lakenfields	7.605	7.427	7.803	8.211	9.448	7.619	8.339	7.131	7.565
48	Nottingham.Centre	7.128	7.670	6.114	7.017	7.685	7.255	7.656	5.453	6.638
49	Nottingham.Western.Boulevard	8.540	8.184	7.568	7.017	9.599	7.122	7.247	6.954	7.444
50	Oxford.St.Ebbes	7.270	5.619	7.137	8.894	5.085	6.016	4.323	6.120	5.304
51	Plymouth.Centre	6.972	6.972	5.684	5.215	14.998	5.684	10.008	5.033	5.776
52	Port.Talbot.Margam	20.373	20.373	20.293	19.497	23.335	20.577	23.121	20.014	20.547
53	Portsmouth	6.474	5.936	5.062	7.692	6.043	5.809	5.155	5.078	5.002
54	Portsmouth.Anglesea.Road	7.993	8.255	7.745	7.692	7.836	6.975	7.949	7.247	7.324
55	Reading.London.Road	7.142	8.160	7.231	8.071	18.675	7.236	12.684	6.804	7.858
56	Reading.New.Town	7.083	5.634	6.681	8.071	5.085	6.201	4.497	5.896	5.225
57	Rochester.Stoke	9.708	11.963	9.791	11.516	11.963	9.181	12.406	9.317	11.338
58	Salford.Eccles	10.134	10.797	9.459	9.641	11.687	9.396	11.053	9.502	10.030
59	Saltash.Callington.Road	6.829	6.906	5.328	5.215	8.024	5.328	6.906	5.115	6.358
60	Sandy.Roadside	5.780	6.155	6.756	8.664	8.664	7.084	11.114	5.937	6.035
61	Scunthorpe.Town	10.616	11.398	9.065	10.592	11.890	9.434	10.977	8.828	10.522
62	Sheffield.Devonshire.Green	7.164	7.454	7.595	5.769	5.769	5.697	5.609	6.193	6.232
63	Southampton.A33	6.762	7.493	6.349	6.224	7.836	6.370	6.971	5.973	6.145
64	Southampton.Centre	5.991	6.723	5.361	6.224	6.043	5.675	6.528	5.095	5.531
65	Southwark.A2.Old.Kent.Road	9.426	8.421	7.857	8.995	10.352	8.296	7.652	7.876	7.643
66	St.Helens.Linkway	8.706	8.309	7.759	8.571	8.157	8.542	8.641	7.938	7.968
67	Stanford.le.Hope.Roadside	6.551	6.983	6.134	5.867	9.483	5.872	7.955	5.339	6.040
68	Stockton.on.Tees.Eaglescliffe	9.042	9.360	9.330	9.273	10.967	9.728	9.977	8.427	8.938
69	Stoke.on.Trent.A50.Roadside	9.428	8.747	9.428	9.788	9.788	9.386	8.333	9.025	8.953
70	Swansea.Roadside	8.395	9.253	11.234	19.497	9.973	11.234	9.119	10.944	8.319
71	Thurrock	6.456	7.963	6.290	5.867	7.392	6.157	7.247	5.562	6.447
72	Warrington	5.249	5.635	4.983	8.571	5.017	5.297	6.540	4.490	4.406
73	Wrexham	5.711	6.867	5.711	6.182	10.782	5.390	9.714	5.692	6.346
74	York.Bootham	5.898	5.958	6.577	7.726	7.666	6.462	7.130	5.321	5.668
75	York.Fishergate	8.256	7.848	6.534	7.726	10.168	7.078	7.481	6.709	6.762
	Average RMSE	8.312	8.367	8.050	9.500	9.714	8.391	8.613	7.439	7.579
	Std	2.750	2.739	2.979	3.815	3.830	3.121	3.012	2.925	2.850

Table D.4: The RMSE between observed and imputed TS using univariate clustering imputation for stations that measure $PM_{2.5}$

	Site	model 1 (CA)	model 2 (CA+ENV)	model 3 (CA+REG)	model 4 (1NN)	model 5 (1NN.ENV)	model 6 (2NN)	model 7 (2NN.ENV)	model 8 (Median)	model 9 (Median.ENV)
1	Aberdeen	4.405	4.652	4.405	5.804	4.873	4.681	4.682	4.354	4.392
2	Auchencorth.Moss	4.501	4.501	3.939	3.976	5.029	4.237	4.578	3.863	3.822
3	Barnstaple.A39	4.450	5.563	4.670	7.313	8.626	7.169	5.563	4.797	5.558
4	Belfast.Centre	7.394	7.487	7.356	9.295	9.295	7.356	6.797	6.983	6.665
5	Birmingham.A4540.Roadside	5.261	5.481	4.212	4.734	4.780	4.299	5.182	4.057	4.275
6	Birmingham.Acocks.Green	5.049	4.866	3.968	4.734	5.288	4.249	4.320	3.834	4.053
7	Blackpool.Marton	4.404	4.609	4.313	5.149	5.149	5.046	5.046	4.109	4.109
8	Bournemouth	5.529	5.492	4.499	4.593	6.254	4.434	5.400	4.131	4.698
9	Bristol.St.Paul.s	5.159	5.079	6.328	6.353	4.946	5.089	4.569	4.659	4.541
10	Camden.Kerbside	4.056	4.385	2.945	6.216	6.216	3.383	4.068	3.189	3.271
11	Cardiff.Centre	4.783	6.204	5.051	3.797	3.797	4.081	3.653	4.118	4.077
12	Carlisle.Roadside	5.225	6.580	5.724	6.361	7.224	6.155	5.266	5.396	5.196
13	Chatham.Roadside	7.613	7.344	7.982	8.206	7.941	7.622	6.692	7.472	7.321
14	Chepstow.A48	5.986	6.473	5.106	6.353	8.053	5.206	6.216	5.075	5.265
15	Chesterfield.Loundsley.Green	4.265	4.576	4.245	5.225	5.609	4.390	4.294	3.845	3.615
16	Chesterfield.Roadside	5.790	5.712	5.519	5.225	8.431	5.100	6.892	5.036	5.422
17	Chilbolton.Observatory	4.823	7.147	4.567	5.946	7.147	5.031	4.848	4.374	5.057
18	Christchurch.Barrack.Road	6.059	6.445	4.849	4.593	6.649	4.654	6.387	4.615	5.422
19	Coventry.Allesley	5.168	5.063	4.031	4.545	4.691	4.221	4.460	4.029	4.107
20	Derry.Rosemount	9.986	10.197	8.568	9.754	9.295	8.568	8.828	8.736	8.703
21	Eastbourne	6.073	6.409	5.615	6.710	6.723	5.626	6.464	5.663	5.953
22	Edinburgh.St.Leonards	3.387	4.101	3.060	3.976	3.905	3.239	4.178	2.849	3.379
23	Glasgow.High.Street	4.035	4.434	3.497	3.136	4.442	3.331	4.079	3.298	3.677
24	Glasgow.Townhead	3.437	4.483	2.706	3.136	3.905	2.712	4.353	2.542	3.133
25	Grangemouth	4.149	7.262	4.072	4.505	7.262	4.205	6.530	3.950	5.938
26	Greenock.A8.Roadside	4.006	4.356	3.367	3.840	4.442	3.845	4.732	3.407	3.673
27	Hull.Freetown	7.065	7.074	7.316	6.230	6.667	6.216	6.883	6.388	6.524
28	Inverness	5.278	5.721	5.278	5.804	5.362	5.253	5.278	5.213	5.241
29	Leamington.Spa	4.276	4.070	3.509	3.748	4.691	3.587	3.768	3.059	3.223
30	Leamington.Spa.Rugby.Road	4.309	4.536	3.410	3.748	4.780	3.443	4.038	3.170	3.444
31	Leeds.Centre	5.147	5.376	3.962	3.724	5.342	3.496	4.595	3.628	4.343
32	Leeds.Headingley.Kerbside	5.600	5.382	4.635	3.724	6.108	4.276	5.192	4.286	4.962
33	Leicester.University	4.400	4.266	4.492	4.975	4.975	4.037	4.037	3.430	3.430
34	Liverpool.Speke	4.048	4.003	3.786	3.782	4.003	3.199	4.348	3.166	3.521
35	London.Bexley	4.401	3.966	3.907	3.966	3.966	3.694	3.694	3.628	3.628
36	London.Bloomsbury	4.009	4.296	2.909	7.026	3.780	4.727	2.856	3.181	3.031
37	London.Eltham	3.642	3.966	3.115	3.966	3.966	3.256	3.256	3.050	3.050
38	London.Harlington	3.296	3.296	3.646	3.762	8.155	2.529	6.206	2.944	3.210
39	London.Marylebone.Road	6.980	6.459	6.543	7.026	6.216	6.411	6.743	6.419	6.342
40	London.N..Kensington	3.313	3.373	3.137	3.931	3.260	5.410	3.217	3.134	2.777
41	London.Teddington.Bushy.Park	3.544	3.862	3.319	3.762	4.089	3.700	3.664	3.191	3.345
42	London.Westminster	3.858	4.247	2.929	3.780	3.780	3.710	3.618	2.950	3.091
43	Lough.Navar	8.073	8.073	7.790	9.754	5.029	7.790	4.433	7.367	5.484
44	Manchester.Piccadilly	7.472	7.380	7.250	6.763	6.763	6.700	6.859	7.036	7.038
45	Middlesbrough	5.707	7.262	4.116	5.413	6.734	4.222	6.395	4.124	4.962
46	Narberth	4.236	4.236	5.669	8.369	5.738	6.687	3.953	5.363	4.120
47	Newcastle.Centre	5.114	5.088	4.260	4.312	4.312	4.156	4.339	3.879	3.864
48	Newport	4.683	4.505	5.106	3.797	3.797	3.501	3.649	3.089	3.141
49	Northampton.Spring.Park	4.429	4.353	4.492	4.492	4.492	4.039	4.039	3.948	3.948
50	Norwich.Lakenfields	5.770	5.810	5.833	6.361	6.977	5.892	5.553	5.568	5.559
51	Nottingham.Centre	5.019	5.068	5.960	7.116	4.975	4.899	4.460	5.044	4.412
52	Oxford.St.Ebbes	4.011	3.809	4.823	4.271	4.271	3.815	3.815	3.751	3.751
53	Plymouth.Centre	4.338	6.204	4.350	4.143	6.204	4.350	5.764	4.074	5.565
54	Port.Talbot.Margam	5.695	5.695	5.442	7.084	8.237	5.889	8.151	5.423	5.627
55	Portsmouth	4.895	4.943	4.172	4.612	4.612	3.840	3.973	4.060	4.068
56	Preston	4.429	4.510	4.372	5.149	5.149	4.565	4.565	4.134	4.134
57	Reading.New.Town	4.021	3.872	4.774	3.848	4.271	3.364	3.769	3.681	3.781
58	Rochester.Stoke	5.811	7.147	5.969	5.284	7.147	5.990	7.368	5.417	6.749
59	Salford.Eccles	5.871	5.722	5.325	6.763	6.763	4.663	4.663	5.172	5.172
60	Saltash.Callington.Road	4.317	5.043	3.361	4.143	4.584	3.361	5.043	3.410	4.252
61	Sandy.Roadside	4.496	4.568	5.272	4.970	5.724	4.335	5.623	4.186	4.507
62	Sheffield.Barnsley.Road	8.455	8.101	7.415	6.187	8.431	6.774	7.168	7.388	7.671
63	Sheffield.Devonshire.Green	5.907	6.241	5.465	6.187	5.609	4.191	5.108	4.875	5.228
64	Southampton.Centre	5.172	5.308	4.818	4.612	4.612	4.673	4.768	4.615	4.620
65	Southend.on.Sea	4.929	5.442	4.616	5.284	5.645	4.002	5.205	4.200	4.707
66	Stanford.le.Hope.Roadside	4.731	5.363	4.565	5.333	7.941	4.050	7.048	3.902	4.835
67	Stockton.on.Tees.Eaglescliffe	6.446	7.902	5.118	5.413	6.068	5.130	5.778	5.058	5.104
68	Stoke.on.Trent.Centre	4.540	4.660	4.540	5.217	8.320	5.681	6.972	4.520	4.547
69	Sunderland.Silksworth	4.891	4.935	3.934	4.312	4.312	3.785	4.313	3.687	3.715
70	Swansea.Roadside	6.846	8.092	6.661	7.084	8.626	7.000	7.526	6.690	7.325
71	Warrington	3.725	4.003	3.536	4.003	4.003	3.833	4.424	3.362	3.471
72	Wigan.Centre	5.306	5.244	4.855	5.661	5.409	4.559	4.556	4.920	4.751
73	Wirral.Tranmere	4.379	4.715	4.215	3.782	6.497	3.606	6.345	3.789	4.412
74	Worthing.A27.Roadside	5.313	6.027	5.011	6.710	9.480	5.145	7.867	4.909	5.887
75	Wrexham	4.333	5.959	4.333	4.855	6.384	4.230	5.487	4.511	4.554
76	York.Bootham	4.920	5.064	4.310	3.462	5.342	3.599	4.511	3.619	4.090
77	York.Fishergate	4.645	5.155	4.344	3.462	6.108	3.670	4.992	3.592	4.392
	Average RMSE	5.079	5.432	4.760	5.255	5.814	4.686	5.168	4.436	4.648
	Std	1.253	1.325	1.279	1.529	1.586	1.270	1.299	1.259	1.230

Appendix E

MVTS clustering imputation

E.1 Experiment 1: The basic k-means clustering algorithm.

Table E.1: The RMSE between observed and imputed TS using MVTS clustering imputation from Experiment 1 for stations that measure O_3

	Site	model 1 (CA)	model 2 (CA+ENV)	model 3 (CA+REG)	model 4 (1NN)	model 5 (1NN.ENV)	model 6 (2NN)	model 7 (2NN.ENV)	model 8 (Median)	model 9 (Median.ENV)
1	Aberdeen	18.054	18.894	18.054	22.026	23.017	21.212	19.089	18.027	18.107
2	Aston.Hill	18.557	13.265	18.557	23.324	27.953	22.019	20.143	17.560	17.472
3	Auchencorth.Moss	14.675	9.316	12.427	7.982	7.982	8.157	7.136	8.023	7.633
4	Barnsley.Gawber	12.236	10.893	10.476	9.919	9.919	11.211	9.174	9.240	8.645
5	Belfast.Centre	18.658	18.030	22.877	22.877	17.856	16.649	18.745	17.361	17.211
6	Birmingham.A4540.Roadside	15.252	25.449	13.117	15.888	25.449	16.436	18.087	12.984	17.236
7	Birmingham.Acocks.Green	13.193	13.139	10.636	15.888	9.716	11.114	7.923	11.199	9.472
8	Blackpool.Marton	13.880	15.982	15.870	14.122	14.122	14.148	14.148	14.143	14.143
9	Bournemouth	13.296	14.129	14.520	18.842	18.842	13.244	17.138	12.744	13.338
10	Brighton.Preston.Park	16.135	15.412	13.050	19.293	15.163	12.523	14.065	13.307	12.772
11	Bristol.St.Paul.s	13.255	11.569	15.682	13.977	13.977	11.357	11.357	11.604	11.604
12	Bush.Estate	15.598	11.355	12.678	7.982	7.982	9.193	9.451	9.219	9.249
13	Canterbury	16.681	17.620	14.848	12.455	14.476	12.135	15.036	13.989	15.040
14	Cardiff.Centre	14.366	12.931	15.571	14.792	14.792	12.106	12.106	12.646	12.646
15	Chilbolton.Observatory	13.023	16.826	14.775	19.427	17.816	18.201	14.911	12.841	13.506
16	Coventry.Allesley	11.303	10.918	8.137	9.957	9.957	7.668	7.668	7.865	7.865
17	Cwmbran	12.297	12.779	13.854	14.792	14.792	12.810	12.810	11.524	11.524
18	Eskdalemuir	14.426	10.606	13.058	13.058	11.289	10.408	11.710	10.552	9.977
19	Exeter.Roadside	20.997	20.997	20.743	29.358	20.289	23.559	21.189	21.688	19.516
20	Fort.William	19.567	19.567	20.045	20.045	26.997	14.191	27.650	14.757	19.428
21	Glasgow.Townhead	16.442	15.854	22.068	22.332	21.180	22.068	14.847	21.266	15.906
22	Glazebury	12.751	20.716	8.513	11.397	20.419	11.630	23.228	9.044	16.658
23	High.Muffles	19.625	16.980	24.299	22.923	14.606	20.243	18.793	19.528	17.562
24	Hull.Freetown	16.134	16.493	16.134	21.261	17.017	13.345	15.369	15.982	16.054
25	Ladybower	14.466	13.065	14.466	17.803	20.419	17.018	13.764	13.474	13.812
26	Leamington.Spa	11.633	11.216	9.095	9.957	9.957	10.095	10.095	8.862	8.862
27	Leeds.Centre	17.392	14.463	15.885	13.524	13.524	17.408	13.409	15.017	13.964
28	Leicester.University	11.957	11.605	12.381	19.208	11.505	11.774	9.713	11.791	9.932
29	Leominster	13.184	13.184	13.792	23.324	21.670	15.393	18.451	13.431	12.731
30	Liverpool.Speke	11.393	17.070	9.279	9.389	17.070	9.866	14.529	8.729	13.019
31	London.Bloomsbury	16.656	16.791	11.114	23.067	14.287	11.525	14.556	10.934	13.396
32	London.Eltham	11.293	11.293	13.505	14.370	21.670	9.662	23.877	9.406	11.587
33	London.Haringey.Priory.Park.South	10.876	11.012	13.231	16.045	16.045	22.283	10.748	11.857	10.571
34	London.Harlington	13.500	13.500	10.201	15.630	27.699	9.741	23.633	9.568	13.154
35	London.Hillingdon	21.088	21.267	15.366	15.630	21.127	17.412	19.201	16.947	18.504
36	London.Marylebone.Road	30.739	25.449	25.305	23.067	25.449	26.183	25.020	25.302	25.137
37	London N Kensington	10.826	10.954	12.089	30.677	14.287	20.966	8.690	12.278	9.007
38	Lough.Navar	19.854	20.861	22.877	22.877	21.097	21.676	20.563	20.755	19.352
39	Lullington.Heath	16.525	14.836	17.762	19.293	19.688	20.417	16.500	16.374	15.125
40	Manchester.Piccadilly	23.585	20.547	22.016	21.145	23.117	19.941	22.117	21.108	21.818
41	Manchester.Sharston	14.097	14.097	11.506	21.145	21.145	13.461	13.461	11.795	11.795
42	Market.Harborough	16.698	13.651	15.537	19.208	14.576	14.060	10.172	14.723	12.051
43	Middlesbrough	12.829	17.070	12.063	13.466	17.070	16.262	17.437	11.845	14.707
44	Narberth	17.133	13.642	19.137	17.995	12.051	19.761	10.652	17.220	11.712
45	Newcastle.Centre	15.186	13.196	13.271	14.904	14.904	13.271	17.993	13.049	13.646
46	Northampton.Spring.Park	14.278	14.552	13.052	13.050	16.965	11.697	15.202	11.680	13.511
47	Norwich.Lakenfields	12.684	13.739	11.132	15.030	14.349	15.774	14.599	10.625	12.216
48	Nottingham.Centre	12.979	12.650	14.991	11.505	11.505	14.071	10.155	12.200	11.381
49	Peebles	16.311	20.043	13.058	12.754	21.180	12.598	20.910	11.552	16.962
50	Plymouth.Centre	14.667	16.122	13.171	19.091	18.971	13.490	17.562	13.379	14.974
51	Port.Talbot.Margam	13.796	13.796	12.708	16.548	19.600	15.351	20.982	12.945	13.503
52	Portsmouth	16.868	14.722	16.903	14.669	14.669	14.478	13.741	13.887	13.705
53	Preston	10.611	10.252	8.346	14.122	14.122	9.785	9.785	8.250	8.250
54	Reading.New.Town	11.652	11.461	13.917	23.021	23.021	17.506	15.039	11.253	11.208
55	Rochester.Stoke	15.428	12.883	14.560	11.847	13.040	13.116	13.835	11.928	11.548
56	Sheffield.Devonshire.Green	14.031	12.898	11.631	17.803	9.919	11.393	10.790	11.675	10.387
57	Sibton	19.068	11.822	14.045	15.030	13.164	11.970	10.197	12.993	12.017
58	Southampton.Centre	14.527	14.342	17.916	14.669	14.669	14.075	14.146	14.014	13.948
59	Southend.on.Sea	10.994	12.075	9.896	11.847	12.947	9.061	10.343	8.641	9.553
60	St.Osyth	15.335	10.788	10.702	12.538	13.040	11.337	10.252	10.649	10.083
61	Stoke.on.Trent.Centre	13.207	12.114	13.207	14.214	15.250	12.197	11.741	12.117	12.139
62	Strathvaich	22.864	17.106	20.045	20.045	16.591	20.305	15.521	18.380	15.689
63	Sunderland.Silksworth	12.754	14.454	12.106	14.904	14.904	12.106	16.576	11.952	12.757
64	Thurrock	11.523	11.866	16.254	10.057	12.947	11.174	10.320	10.904	11.249
65	Walsall.Woodlands	14.008	14.581	13.792	18.442	10.223	12.617	10.101	11.508	10.491
66	Weybourne	20.990	18.921	20.990	19.389	14.433	15.351	15.216	18.569	17.885
67	Wicken.Fen	14.052	11.685	12.384	14.683	14.576	11.570	12.128	11.186	11.220
68	Wigan.Centre	14.143	13.272	10.531	11.397	11.108	9.615	14.230	10.410	11.330
69	Wirral.Tranmere	10.903	11.496	10.340	9.389	14.078	9.580	10.898	9.051	9.833
70	Yarner.Wood	16.598	14.078	19.807	29.358	14.891	22.485	13.442	19.208	13.462
	Average RMSE	15.223	14.717	14.648	16.729	16.345	14.521	14.772	13.293	13.368
	Std	3.575	3.528	4.020	5.075	4.779	4.323	4.604	3.742	3.460

Table E.2: The RMSE between observed and imputed TS using MVTS clustering imputation from Experiment 1 for stations that measure NO_2

	Site	model 1 (CA)	model 2 (CA+ENV)	model 3 (CA+REG)	model 4 (1NN)	model 5 (1NN.ENV)	model 6 (2NN)	model 7 (2NN.ENV)	model 8 (Median)	model 9 (Median.ENV)
1	Aberdeen	16.257	16.038	20.459	29.015	17.420	29.743	16.399	20.274	14.957
2	Aberdeen.Union.Street.Roadside	24.699	20.219	21.987	20.172	20.172	18.484	19.434	19.368	19.449
3	Aberdeen.Wellington.Road	29.471	24.274	27.363	20.172	20.172	23.632	23.440	24.411	24.268
4	Armagh.Roadside	13.505	14.224	11.873	33.238	33.238	20.966	18.502	12.854	12.496
5	Aston.Hill	15.548	4.529	15.548	7.981	14.394	11.974	9.068	11.203	11.370
6	Ballymena..Antrim.Road	13.215	16.460	10.649	12.611	35.085	11.615	20.904	10.333	13.606
7	Ballymena..Ballykeel	12.607	10.681	15.276	12.611	18.465	13.404	10.340	11.332	11.196
8	Barnsley.Gawber	10.774	7.530	11.251	11.929	11.929	18.000	12.721	10.810	9.951
9	Bath.Roadside	21.276	17.577	29.232	29.760	29.760	24.441	19.736	20.770	19.049
10	Belfast.Centre	14.551	17.090	12.822	31.714	18.465	14.922	20.113	12.373	15.672
11	Belfast.Stockman.s.Lane	36.555	30.310	37.076	31.714	35.085	32.374	33.452	33.782	34.599
12	Billingham	11.337	10.238	9.104	12.812	10.809	10.556	9.670	9.119	8.885
13	Birkenhead.Borough.Road	11.127	14.045	11.127	10.394	19.220	11.010	17.262	10.294	12.200
14	Birmingham.A4540.Roadside	11.035	11.186	13.735	19.237	10.501	12.085	11.724	11.931	10.099
15	Birmingham.Acocks.Green	14.723	10.948	10.290	19.237	9.544	18.594	7.439	13.363	9.479
16	Blackburn.Accrington.Road	8.831	15.622	7.715	8.318	12.362	8.729	12.597	7.761	9.242
17	Blackpool.Marton	13.264	9.323	13.228	12.975	12.975	11.413	10.217	11.235	11.007
18	Borehamwood.Meadow.Park	12.907	9.363	10.926	7.789	7.789	21.907	9.911	9.915	8.919
19	Bournemouth	7.772	7.854	9.785	14.811	21.300	20.013	15.455	8.909	8.942
20	Bradford.Mayo.Avenue	28.833	22.702	29.286	27.127	20.725	23.145	23.734	26.568	24.706
21	Brighton.Preston.Park	10.671	10.138	9.979	28.225	9.934	20.827	8.485	10.047	9.221
22	Bristol.St.Paul.s	12.349	12.207	12.356	29.303	10.991	24.023	11.765	11.418	11.787
23	Bristol.Temple.Way	35.684	25.987	35.503	29.303	29.760	26.222	27.726	31.265	29.220
24	Burton.on.Trent..Horninglow	12.661	9.034	9.139	22.144	9.530	13.775	7.973	10.968	8.039
25	Bury.Whitefield.Roadside	12.051	14.784	10.256	11.259	12.296	10.561	9.590	10.256	9.861
26	Bush.Estate	20.334	5.799	27.662	54.774	6.600	35.168	6.342	26.248	5.814
27	Cambridge.Roadside	9.297	13.646	11.321	21.381	13.409	12.888	17.335	10.348	9.945
28	Camden.Kerbside	33.311	27.560	25.461	47.877	47.877	24.260	26.842	24.346	24.064
29	Cannock.A5190.Roadside	13.069	14.841	13.888	12.130	16.021	9.608	15.724	10.880	10.734
30	Canterbury	19.195	14.910	16.987	7.519	11.073	9.815	13.188	14.151	14.655
31	Cardiff.Centre	9.073	9.059	8.063	17.187	9.145	11.146	9.055	7.781	8.157
32	Cardiff.Newport.Road	19.722	11.436	19.251	17.187	47.435	16.042	31.590	16.879	12.059
33	Carlisle.Roadside	10.257	13.769	12.673	18.298	18.298	13.313	19.477	10.303	11.773
34	Charlton..Mackrell	12.258	3.444	15.488	39.022	6.798	39.348	4.121	14.972	5.983
35	Chatham.Roadside	11.554	16.787	10.711	15.568	15.375	11.974	19.559	10.353	12.342
36	Chepstow.A48	17.086	14.025	39.904	23.496	28.248	23.189	28.398	16.613	20.943
37	Chesterfield.Loundsley.Green	13.859	10.172	11.460	10.829	14.589	11.300	16.284	10.108	11.325
38	Chesterfield.Roadside	10.553	16.887	9.336	10.829	23.535	8.951	22.085	8.208	14.554
39	Chilbolton.Observatory	9.537	7.231	7.775	33.535	6.798	27.664	6.160	9.653	5.847
40	Christchurch.Barrack.Road	13.347	16.214	12.423	14.811	29.079	16.774	23.416	12.483	14.958
41	Coventry.Allesley	12.354	8.955	7.674	13.653	9.901	7.992	8.410	8.420	7.432
42	Coventry.Binley.Road	9.545	11.560	11.078	13.653	16.540	14.507	9.574	10.750	9.774
43	Crumlyn	8.470	8.075	9.866	10.853	10.853	34.379	10.373	9.128	8.729
44	Derby.St..Alkmund.s.Way	15.771	12.693	17.793	22.144	13.048	15.004	15.948	15.886	14.240
45	Derry.Rosemount	16.646	13.375	22.524	19.120	11.941	14.640	15.993	16.149	14.531
46	Dewsbury.Ashworth.Grove	8.322	9.811	6.198	27.127	11.570	15.573	7.220	6.576	6.519
47	Doncaster.A630.Cleveland.Street	11.854	11.546	10.263	13.466	15.490	11.368	10.816	9.883	9.607
48	Dumbarton.Roadside	13.135	18.855	14.872	25.019	25.019	19.419	19.419	15.744	15.744
49	Dumfries	17.976	15.040	33.314	35.285	18.298	25.088	16.739	25.080	16.808
50	Dundee.Mains.Loan	14.804	11.635	26.863	13.660	13.660	28.912	9.612	14.554	11.913
51	Eastbourne	8.855	9.027	5.465	7.685	9.934	5.022	12.947	5.189	7.462
52	Edinburgh..Nicolson.Street	38.920	33.102	36.817	40.976	32.013	47.554	29.359	38.605	33.675
53	Edinburgh.St.Leonards	12.236	11.325	13.830	40.976	16.673	15.683	14.071	11.902	10.504
54	Eskdalemuir	24.179	6.950	19.907	35.285	6.600	19.907	4.942	19.971	8.063
55	Exeter.Roadside	19.846	12.514	18.543	30.439	19.973	28.418	14.559	20.082	16.622
56	Fort.William	18.063	18.063	17.285	17.285	11.036	22.107	11.985	17.932	14.051
57	Glasgow..Great.Western.Road	16.240	15.259	11.862	41.164	41.164	20.265	22.463	13.010	13.192
58	Glasgow..High.Street	16.844	14.941	11.788	11.606	40.379	19.070	21.368	11.466	12.556
59	Glasgow..Kerbside	46.307	39.632	45.008	45.561	40.379	42.655	40.373	44.097	42.691
60	Glasgow..Townhead	11.885	13.317	9.927	11.606	13.151	26.413	16.710	9.064	11.438
61	Glazebury	13.530	13.086	12.494	15.081	12.291	12.153	13.074	11.010	11.694
62	Grangemouth	13.616	9.547	17.601	6.580	6.580	24.526	8.167	12.189	8.630
63	Grangemouth.Moray	13.053	10.227	16.135	6.580	6.580	22.956	8.832	11.239	8.509
64	Greenock.A8.Roadside	21.154	17.574	19.098	25.019	25.019	21.498	21.498	19.540	19.540
65	Hafod.yr.ynys.Roadside	48.249	41.968	39.904	61.327	47.435	57.779	42.258	48.719	43.420
66	Haringey.Roadside	16.048	13.872	9.979	19.227	22.593	14.150	20.500	12.977	11.510
67	Hartlepool.St.Abbs.Walk	13.543	10.087	11.956	10.855	9.884	9.449	12.093	9.949	10.580
68	High.Muffles	21.463	5.878	23.778	15.394	5.766	15.172	8.621	14.921	8.668
69	Honiton	10.729	11.379	13.475	26.750	15.904	12.372	14.617	12.586	11.933
70	Horley	15.068	15.068	13.847	17.552	16.926	23.223	22.181	14.769	14.760
71	Hull.Freetown	12.464	10.996	9.589	10.755	14.547	7.377	12.420	7.690	9.107
72	Hull.Holderness.Road	13.875	16.783	11.218	10.755	15.573	14.946	15.024	10.901	12.944
73	Immingham.Woodlands.Avenue	13.226	10.106	14.691	14.547	14.547	17.185	11.091	12.311	11.504
74	Inverness	13.709	18.931	17.285	17.285	30.429	16.591	30.638	13.787	17.322
75	Ladybower	20.139	4.644	19.369	21.372	12.291	28.398	8.155	18.677	10.287
76	Leamington.Spa	14.023	10.155	10.285	7.008	9.901	9.750	8.596	9.204	9.302
77	Leamington.Spa.Rugby.Road	13.866	21.778	9.761	7.008	16.540	6.254	17.456	9.267	15.052
78	Leeds.Centre	13.582	16.113	12.519	13.576	11.570	10.587	14.257	11.960	13.115
79	Leeds.Headingley.Kerbside	15.911	12.240	14.824	13.576	20.725	10.704	12.671	12.424	11.518
80	Leicester.A594.Roadside	16.217	13.568	17.797	18.537	14.226	25.537	12.292	17.115	13.778
81	Leicester..University	11.579	10.487	9.314	18.537	12.185	8.955	10.203	9.200	9.634
82	Leominster	9.904	9.904	12.978	7.981	13.498	5.258	16.495	8.629	10.535
83	Lincoln.Canwick.Road	21.452	17.375	29.686	28.438	19.931	28.448	18.372	26.517	18.944
84	Liverpool.Speke	11.389	9.944	10.635	25.872	11.601	14.711	10.101	11.051	9.419
85	London..Bexley	11.183	10.179	19.302	10.179	10.179	7.873	14.600	7.790	9.984
86	London..Bloomsbury	15.042	18.866	11.776	60.794	10.066	29.223	11.179	11.741	11.085
87	London..Eltham	14.787	26.032	10.179	10.179	10.179	18.989	6.281	12.963	10.068
88	London..Haringey..Priory.Park.South	10.923	7.776	19.177	19.227	19.279	27.583	14.112	15.697	13.180
89	London..Harlington	13.239	13.239	15.180	26.836	24.017	14.532	22.382	12.328	13.234
90	London..Hillingdon	30.373	34.518	25.240	26.836	29.916	27.752	30.920	28.013	30.150
91	London..Marylebone.Road	66.473	60.588	60.660	60.794	47.877	52.978	49.733	60.513	58.351
92	London N Kensington	11.443	11.992	13.688	33.020	14.098	46.663	11.143	13.388	9.483
93	London..Westminster	10.782	13.672	11.633	10.066	10.066	34.871	7.148	7.723	6.649
94	Lullington..Heath	12.172	5.657	10.229	7.685	10.664	10.592	7.613	8.370	7.503
95	Luton.A505.Roadside	25.026	20.797	31.882	32.314	28.494	29.159	24.145	28.421	24.859
96	Manchester..Piccadilly	15.659	19.508	15.756	14.776	14.776	13.424	16.529	15.512	16.116
97	Manchester.Sharston	10.682	10.682	9.440	16.838	16.926	11.929	10.979	9.662	9.980
98	Market.Harborough	23.018	6.900	23.930	19.904	5.476	26.298	4.811	21.432	6.613
99	Middlesbrough	12.466	8.897	10.749	10.809	10.809	8.688	9.359	8.586	8.852
100	Narberth	14.951	4.080	19.602	25.479	3.082	19.107	2.824	17.347	4.208
101	Newcastle.Centre	11.365	14.263	12.934	19.199	18.790	10.899	18.379	10.564	14.358
102	Newcastle.Cradlewell.Roadside	25.908	21.744	28.635	19.199	24.585	21.064	26.125	22.790	25.376
103	Newport	8.451	8.661	7.381	10.853	10.853	24.616	8.118	7.460	7.942
104	Northampton.Spring.Park	17.728	13.109	17.711	9.130	15.786	8.199	12.073	11.802	14.194
105	Norwich.Lakenfields	19.070	14.491	13.990	7.617	11.760	9.223	14.225	12.653	13.795
106	Nottingham.Centre	8.774	10.111	8.279	14.885	12.862	14.447	11.149	7.869	9.104
107	Nottingham.Western.Boulevard	15.092	13.206	15.920	14.885	13.048	11.210	14.595	13.220	12.947
108	Oldbury.Birmingham.Road	11.099	10.097	14.630	10.501	10.501	12.892	10.499	10.379	9.641
109	Oxford.Centre.Roadside	18.436	15.480	21.308	28.897	21.313	22.906	17.949	20.191	17.246
110	Oxford.St.Ebbes	16.175	11.439	14.346	28.897	14.867	23.322	8.886	15.430	12.244
111	Peebles	19.720	15.780	14.521	5.550	16.673	26.960	18.921	15.756	15.439
112	Plymouth.Centre	10.108	10.707	9.870	12.551	15.690	12.545	11.474	9.924	10.481
113	Plymouth.Tavistock.Road	16.256	22.230	29.232	12.551	19.973	14.058	16.210	14.744	17.474
114	Port.Talbot.Margam	8.961	8.961	10.245	16.840	13.342	13.052	12.671	10.073	9.181
115	Portsmouth	10.898	10.662	13.375	18.308	16.657	15.639	10.328	10.016	10.093
116	Portsmouth.Anglesea.Road	13.660	14.689	14.300	18.308	21.866	14.891	15.623	13.446	13.838
117	Preston	9.427	9.092	7.627	8.318	12.975	9.332	10.776	7.644	8.355
118	Reading..London.Road	15.317	17.165	15.602	12.496	21.313	22.168	25.177	14.719	16.079
119	Reading..New..Town	10.412	9.382	10.156	12.496	33.883	17.375	16.286	9.128	10.146
120	Rochester..Stoke	18.675	8.122	16.684	9.840	8.011	11.340	8.156	11.130	7.335
121	Salford.Eccles	10.558	11.818	8.645	14.776	14.776	11.008	9.780	8.661	8.801
122	Sandy.Roadside	12.921	16.029	14.467	28.494	28.494	17.002	17.002	14.697	14.697
123	Scunthorpe.Town	11.844	8.967	12.834	12.357	11.163	7.707	9.967	9.154	9.149
124	Shaw.Crompton.Way	12.744	14.179	12.744	12.296	12.296	12.100	12.549	11.704	11.952
125	Sheffield.Barnsley.Road	19.608	14.636	18.847	19.004	23.535	17.241	18.143	17.674	18.468
126	Sheffield..Devonshire.Green	10.197	10.529	8.290	19.004	13.588	13.837	9.812	8.840	8.952
127	Sheffield.Tinsley	12.117	12.889	13.242	18.046	13.588	12.840	13.758	11.142	11.654
128	Southampton.A33	19.191	17.720	21.714	17.744	21.866	24.059	23.223	19.385	19.957
129	Southampton.Centre	11.184	12.928	11.759	17.744	16.657	12.704	18.025	10.502	12.400
130	Southend.on.Sea	13.263	9.559	8.242	9.840	11.880	7.073	7.610	7.645	8.515
131	Southwark.A2.Old.Kent.Road	23.665	21.342	20.458	21.665	21.665	16.966	38.892	18.972	19.631
132	St.Helens.Linkway	16.677	13.347	16.667	14.567	14.567	13.223	13.418	13.255	13.188
133	St.Osyth	19.207	6.911	14.140	10.564	8.011	7.980	6.627	9.593	7.234
134	Stanford.le.Hope..Roadside	12.255	16.204	12.911	10.417	15.375	13.318	15.731	10.734	12.062
135	Stockton.on..Tees.A1305.Roadside	12.388	17.600	10.634	12.812	9.757	9.292	9.825	10.445	9.709
136	Stockton.on..Tees.Eaglescliffe	14.252	21.570	12.559	9.757	9.757	7.718	11.624	11.661	12.204
137	Stoke.on.Trent..A50.Roadside	38.791	31.985	38.197	38.197	38.694	37.917	32.288	37.393	35.736
138	Stoke.on.Trent.Centre	10.029	11.411	38.197	38.197	12.246	20.292	12.198	20.171	10.257
139	Storrington.Roadside	13.360	17.817	11.706	19.971	19.971	11.636	16.926	12.517	14.388
140	Sunderland.Silksworth	12.801	10.042	10.755	12.914	18.790	21.902	11.513	11.241	11.212
141	Sunderland.Wessington.Way	10.876	15.016	9.488	12.914	24.585	10.578	14.308	8.827	10.837
142	Swansea.Roadside	14.433	14.101	13.272	16.840	51.768	20.266	29.888	13.999	13.831
143	Swindon.Walcot	7.381	7.393	9.380	8.209	8.209	17.331	9.767	7.626	7.199
144	Telford.Hollinswood	8.446	8.810	7.466	8.129	8.129	9.256	9.378	7.315	7.392
145	Thurrock	11.246	10.617	10.375	9.974	11.880	8.102	10.481	8.968	9.636
146	Tower.Hamlets.Roadside	25.170	20.786	19.700	21.665	21.665	18.600	26.651	19.642	19.067
147	Walsall.Woodlands	7.919	7.959	6.314	12.130	8.800	15.120	7.156	6.807	6.350
148	Warrington	10.003	11.227	8.900	22.557	11.601	19.932	10.453	9.778	9.474
149	Wicken.Fen	22.892	5.977	17.915	21.381	5.476	22.099	5.347	18.795	6.544
150	Widnes.Milton.Road	20.287	14.953	20.782	25.872	14.567	23.568	14.410	20.638	16.230
151	Wigan.Centre	11.155	8.486	9.548	21.098	13.515	10.396	10.902	9.645	9.374
152	Wirral.Tranmere	11.032	9.276	10.245	10.394	11.074	7.041	10.705	8.257	9.454
153	Worthing.A27.Roadside	17.787	14.101	19.835	19.971	19.971	23.026	16.863	18.372	16.909
154	Wrexham	11.315	17.363	11.315	11.045	12.892	9.975	17.062	11.288	12.110
155	Yarner.Wood	14.439	3.899	17.821	30.439	4.642	23.968	3.560	17.043	4.886
156	York.Bootham	12.430	8.982	12.952	15.332	18.900	15.568	15.635	11.918	12.293
157	York.Fishergate	9.623	10.757	9.872	15.332	15.891	11.652	12.325	9.203	9.622
	Average RMSE	15.805	13.947	16.120	19.447	17.269	18.052	15.254	14.369	13.348
	Std	8.107	7.299	8.608	11.064	9.516	9.258	7.830	7.907	7.395

Table E.3: The RMSE between observed and imputed TS using MVTS clustering imputation from Experiment 1 for stations that measure PM_{10}

	Site	model 1 (CA)	model 2 (CA+ENV)	model 3 (CA+REG)	model 4 (1NN)	model 5 (1NN.ENV)	model 6 (2NN)	model 7 (2NN.ENV)	model 8 (Median)	model 9 (Median.ENV)
1	Aberdeen	8.402	8.488	8.402	10.681	9.360	9.288	9.362	8.297	8.280
2	Armagh.Roadside	19.523	19.217	18.731	18.905	18.905	18.004	19.241	18.664	18.870
3	Auchencorth.Moss	9.541	6.686	6.781	6.180	6.686	6.271	6.643	5.876	5.842
4	Barnstaple.A39	6.822	7.748	6.896	21.738	10.721	14.096	8.219	7.170	7.244
5	Belfast.Centre	7.990	8.613	8.059	8.506	11.448	10.506	9.287	6.968	8.047
6	Belfast.Stockman.s.Lane	8.705	8.702	8.718	8.506	18.905	11.537	11.427	7.488	8.470
7	Birmingham.A4540.Roadside	7.582	7.976	6.759	6.554	8.087	6.141	7.085	6.324	6.668
8	Birmingham.Ladywood	5.758	5.362	4.077	6.554	5.294	6.170	5.014	4.148	4.267
9	Bristol.St.Paul.s	6.584	5.980	7.701	10.850	6.646	8.155	6.724	6.684	5.817
10	Bristol.Temple.Way	10.707	10.184	10.995	10.850	11.542	10.247	9.566	10.392	10.122
11	Bury.Whitefield.Roadside	6.311	6.472	5.385	9.641	8.157	6.432	7.610	5.239	5.412
12	Camden.Kerbside	7.008	6.725	5.759	8.789	8.789	6.449	10.581	5.767	5.756
13	Cardiff.Centre	13.190	12.974	13.859	14.400	13.801	13.373	12.448	13.043	13.011
14	Cardiff.Newport.Road	7.999	7.212	9.017	14.400	9.132	9.809	7.357	7.968	7.267
15	Carlisle.Roadside	6.840	6.866	8.246	12.541	10.035	9.977	9.289	7.014	6.921
16	Chatham.Roadside	10.016	9.419	11.140	11.516	9.483	9.541	8.542	9.860	9.525
17	Chepstow.A48	8.110	8.371	8.110	9.049	11.542	7.472	9.243	7.980	7.963
18	Chesterfield.Loundsley.Green	5.858	6.035	6.667	6.667	5.769	5.040	5.241	4.633	4.098
19	Chesterfield.Roadside	8.109	8.217	6.667	6.667	9.599	6.389	8.443	6.343	7.017
20	Chilbolton.Observatory	8.337	8.204	6.359	8.748	11.963	7.918	7.458	6.043	5.908
21	Coventry.Binley.Road	7.234	7.074	7.393	8.735	8.735	8.475	7.250	7.300	6.992
22	Derry.Rosemount	10.569	10.890	10.095	10.857	11.448	11.995	9.763	9.106	9.594
23	Ealing.Horn.Lane	16.623	15.926	15.799	18.484	16.070	16.753	14.879	16.369	15.614
24	Edinburgh.St.Leonards	6.899	7.075	4.444	6.180	5.401	4.372	6.774	4.091	5.073
25	Glasgow.High.Street	6.490	7.146	5.697	4.978	8.674	5.645	6.445	4.814	5.398
26	Glasgow.Townhead	6.688	6.955	3.942	4.978	5.401	5.466	5.999	3.932	4.808
27	Grangemouth	6.768	8.999	4.744	5.198	10.603	4.827	8.747	4.540	7.648
28	Greenock.A8.Roadside	9.250	9.831	8.117	8.520	8.674	8.229	9.104	7.956	8.330
29	Hull.Holderness.Road	10.404	10.215	10.404	10.592	9.499	8.572	10.048	10.089	10.048
30	Inverness	8.555	9.673	8.555	10.681	8.836	8.199	7.946	8.630	8.597
31	Leamington.Spa	6.624	5.307	5.233	4.289	5.294	5.329	4.282	4.389	4.439
32	Leamington.Spa.Rugby.Road	6.313	7.984	5.129	4.289	8.735	5.311	7.641	4.679	6.037
33	Leeds.Centre	7.263	7.811	5.675	8.032	7.666	6.430	6.413	5.310	6.302
34	Leeds.Headingley.Kerbside	9.840	9.595	8.526	8.032	10.168	8.153	9.216	8.206	8.784
35	Leicester.A594.Roadside	9.037	8.074	9.055	10.011	8.293	8.163	7.683	8.302	7.657
36	Liverpool.Speke	6.125	6.370	6.266	9.222	5.017	6.061	7.492	5.602	5.357
37	London.Bloomsbury	5.548	5.976	5.953	10.043	5.461	7.561	5.090	5.338	4.663
38	London.Harlington	5.669	5.669	7.567	17.868	22.803	9.273	13.891	6.733	6.061
39	London.Marylebone.Road	9.310	8.401	8.509	10.043	8.789	8.719	9.254	8.498	7.831
40	London N Kensington	5.763	4.466	7.885	18.484	5.461	12.035	5.823	7.610	4.901
41	Lough.Navar	10.021	6.686	11.688	10.857	6.686	14.156	6.699	8.790	6.334
42	Middlesbrough	7.287	7.752	7.589	9.273	11.890	8.116	8.662	6.777	7.357
43	Narberth	7.956	8.204	10.024	11.196	8.204	14.857	7.238	8.388	7.029
44	Newcastle.Centre	10.186	10.186	10.268	10.853	10.732	10.036	10.673	9.573	9.842
45	Newcastle.Cradlewell.Roadside	7.890	8.087	8.214	10.853	10.967	8.522	8.802	7.591	7.810
46	Newport	5.591	6.283	7.744	8.961	13.801	9.141	9.052	6.028	6.149
47	Norwich.Lakenfields	7.369	6.928	7.803	8.211	9.448	7.619	8.339	7.092	7.509
48	Nottingham.Centre	6.907	7.118	7.255	7.017	7.685	7.255	7.656	6.246	5.709
49	Nottingham.Western.Boulevard	7.851	7.955	6.615	7.017	9.599	7.122	7.247	6.383	6.534
50	Oxford.St.Ebbes	7.694	5.719	7.926	8.894	5.085	6.016	4.323	6.459	5.538
51	Plymouth.Centre	6.789	8.490	7.019	5.215	14.998	5.684	10.008	5.930	8.053
52	Port.Talbot.Margam	20.444	20.444	20.413	19.497	23.335	20.577	23.121	20.172	20.594
53	Portsmouth	7.741	8.205	6.359	7.692	6.043	5.809	5.155	5.403	5.196
54	Portsmouth.Anglesea.Road	8.317	8.443	7.904	7.692	7.836	6.975	7.949	7.394	7.507
55	Reading.London.Road	7.404	8.281	7.293	8.071	18.675	7.236	12.684	6.919	7.945
56	Reading.New.Town	7.452	5.656	7.400	8.071	5.085	6.201	4.497	6.155	5.357
57	Rochester.Stoke	9.657	9.657	9.719	11.516	11.963	9.181	12.406	9.348	9.687
58	Salford.Eccles	10.213	10.934	9.459	9.641	11.687	9.396	11.053	9.518	10.048
59	Saltash.Callington.Road	6.827	7.751	6.779	5.215	8.024	5.328	6.906	5.891	6.554
60	Sandy.Roadside	5.435	5.872	6.756	8.664	8.664	7.084	11.114	5.777	5.854
61	Scunthorpe.Town	10.945	11.398	9.621	10.592	11.890	9.434	10.977	9.125	10.643
62	Sheffield.Devonshire.Green	7.332	7.702	7.258	5.769	5.769	5.697	5.609	6.152	6.173
63	Southampton.A33	7.056	7.489	6.520	6.224	7.836	6.370	6.971	6.120	6.254
64	Southampton.Centre	6.373	7.136	5.611	6.224	6.043	5.675	6.528	5.267	5.686
65	Southwark.A2.Old.Kent.Road	9.182	8.711	7.857	8.995	10.352	8.296	7.652	7.925	7.719
66	St.Helens.Linkway	8.903	8.530	7.759	8.571	8.157	8.542	8.641	8.018	8.043
67	Stanford.le.Hope.Roadside	6.321	6.898	6.134	5.867	9.483	5.872	7.955	5.393	5.992
68	Stockton.on.Tees.Eaglescliffe	9.084	9.436	9.330	9.273	10.967	9.728	9.977	8.437	8.934
69	Stoke.on.Trent.A50.Roadside	9.726	9.113	9.726	9.788	9.788	9.386	8.333	9.309	9.242
70	Swansea.Roadside	8.371	8.691	8.500	19.497	9.973	11.234	9.119	8.131	8.375
71	Thurrock	6.238	7.658	6.290	5.867	7.392	6.157	7.247	5.554	6.464
72	Warrington	5.268	5.635	4.983	8.571	5.017	5.297	6.540	4.477	4.370
73	Wrexham	5.615	6.691	5.615	6.182	10.782	5.390	9.714	5.572	6.192
74	York.Bootham	6.068	6.177	6.499	7.726	7.666	6.462	7.130	5.235	5.788
75	York.Fishergate	8.616	8.398	7.007	7.726	10.168	7.078	7.481	6.904	7.029
	Average RMSE	8.247	8.283	8.003	9.500	9.714	8.391	8.613	7.398	7.522
	Std	2.720	2.650	2.819	3.815	3.830	3.121	3.012	2.867	2.844

Table E.4: The RMSE between observed and imputed TS using MVTS clustering imputation from Experiment 1 for stations that measure $PM_{2.5}$

	Site	model 1 (CA)	model 2 (CA+ENV)	model 3 (CA+REG)	model 4 (1NN)	model 5 (1NN.ENV)	model 6 (2NN)	model 7 (2NN.ENV)	model 8 (Median)	model 9 (Median.ENV)
1	Aberdeen	5.687	6.098	5.687	5.804	4.873	4.681	4.682	5.663	5.665
2	Auchencorth.Moss	6.109	5.029	3.939	3.976	5.029	4.237	4.578	3.694	4.251
3	Barnstaple.A39	4.867	5.522	5.073	7.313	8.626	7.169	5.563	4.981	5.271
4	Belfast.Centre	6.955	7.076	7.356	9.295	9.295	7.356	6.797	6.983	6.669
5	Birmingham.A4540.Roadside	5.094	5.489	4.212	4.734	4.780	4.299	5.182	4.042	4.269
6	Birmingham.Acocks.Green	4.991	4.716	3.968	4.734	5.288	4.249	4.320	3.844	4.057
7	Blackpool.Marton	4.188	4.284	4.313	5.149	5.149	5.046	5.046	3.957	3.957
8	Bournemouth	4.672	4.953	4.566	4.593	6.254	4.434	5.400	4.170	4.522
9	Bristol.St.Paul.s	4.955	4.764	5.752	6.353	4.946	5.089	4.569	4.711	4.658
10	Camden.Kerbside	3.866	4.278	2.945	6.216	6.216	3.383	4.068	3.164	3.267
11	Cardiff.Centre	4.243	4.491	4.263	3.797	3.797	4.081	3.653	3.624	3.621
12	Carlisle.Roadside	4.590	5.004	5.724	6.361	7.224	6.155	5.266	4.806	4.758
13	Chatham.Roadside	7.626	7.298	8.028	8.206	7.941	7.622	6.692	7.483	7.292
14	Chepstow.A48	6.211	6.623	6.211	6.353	8.053	5.206	6.216	6.088	6.132
15	Chesterfield.Loundsley.Green	4.601	4.710	5.225	5.225	5.609	4.390	4.294	3.793	3.385
16	Chesterfield.Roadside	6.233	6.250	5.225	5.225	8.431	5.100	6.892	4.881	5.109
17	Chilbolton.Observatory	4.485	5.738	5.666	5.946	7.147	5.031	4.848	4.310	4.171
18	Christchurch.Barrack.Road	5.691	6.933	5.711	4.593	6.649	4.654	6.387	5.344	5.731
19	Coventry.Allesley	4.992	4.785	4.031	4.545	4.691	4.221	4.460	3.996	4.080
20	Derry.Rosemount	9.629	9.850	8.568	9.754	9.295	8.568	8.828	8.767	8.726
21	Eastbourne	7.098	6.858	6.569	6.710	6.723	5.626	6.464	6.142	6.232
22	Edinburgh.St.Leonards	5.081	5.657	3.060	3.976	3.905	3.239	4.178	2.993	3.862
23	Glasgow.High.Street	5.096	5.747	3.497	3.136	4.442	3.331	4.079	3.181	3.718
24	Glasgow.Townhead	4.714	5.263	2.706	3.136	3.905	2.712	4.353	2.480	3.293
25	Grangemouth	5.299	6.471	4.072	4.505	7.262	4.205	6.530	3.989	5.799
26	Greenock.A8.Roadside	5.487	6.165	3.367	3.840	4.442	3.845	4.732	3.510	4.087
27	Hull.Freetown	7.770	7.584	7.770	6.230	6.667	6.216	6.883	7.490	7.494
28	Inverness	6.694	7.409	6.694	5.804	5.362	5.253	5.278	6.632	6.609
29	Leamington.Spa	4.175	3.868	3.509	3.748	4.691	3.587	3.768	3.072	3.231
30	Leamington.Spa.Rugby.Road	4.176	4.551	3.410	3.748	4.780	3.443	4.038	3.171	3.468
31	Leeds.Centre	5.750	5.713	3.859	3.724	5.342	3.496	4.595	3.517	4.398
32	Leeds.Headingley.Kerbside	6.166	6.175	4.308	3.724	6.108	4.276	5.192	4.137	5.132
33	Leicester.University	4.135	3.789	3.643	4.975	4.975	4.037	4.037	3.374	3.374
34	Liverpool.Speke	4.099	4.557	3.786	3.782	4.003	3.199	4.348	3.306	3.720
35	London.Bexley	4.293	3.966	3.907	3.966	3.966	3.694	3.694	3.619	3.619
36	London.Bloomsbury	3.883	4.278	2.909	7.026	3.780	4.727	2.856	3.182	2.995
37	London.Eltham	3.611	3.966	3.115	3.966	3.966	3.256	3.256	3.056	3.056
38	London.Harlington	3.391	3.391	3.646	3.762	8.155	2.529	6.206	2.994	3.257
39	London.Marylebone.Road	6.878	6.411	6.543	7.026	6.216	6.411	6.743	6.418	6.342
40	London N Kensington	3.342	3.485	3.137	3.931	3.260	5.410	3.217	3.167	2.791
41	London.Teddington.Bushy.Park	3.617	4.114	3.319	3.762	4.089	3.700	3.664	3.237	3.371
42	London.Westminster	3.808	4.371	2.929	3.780	3.780	3.710	3.618	2.982	3.120
43	Lough.Navar	6.825	5.029	7.790	9.754	5.029	7.790	4.433	7.169	4.555
44	Manchester.Piccadilly	7.815	7.627	7.250	6.763	6.763	6.700	6.859	7.110	7.100
45	Middlesbrough	4.900	5.807	4.116	5.413	6.734	4.222	6.395	4.072	5.432
46	Narberth	4.715	5.738	5.049	8.369	5.738	6.687	3.953	5.038	4.367
47	Newcastle.Centre	4.689	4.913	4.260	4.312	4.312	4.156	4.339	3.854	3.928
48	Newport	3.767	4.123	4.132	3.797	3.797	3.501	3.649	3.207	3.258
49	Northampton.Spring.Park	4.249	4.048	4.721	4.492	4.492	4.039	4.039	3.862	3.862
50	Norwich.Lakenfields	5.521	5.424	5.833	6.361	6.977	5.892	5.553	5.472	5.434
51	Nottingham.Centre	5.591	5.262	5.072	7.116	4.975	4.899	4.460	4.626	4.673
52	Oxford.St.Ebbes	4.159	3.990	4.894	4.271	4.271	3.815	3.815	3.777	3.777
53	Plymouth.Centre	4.498	5.326	4.444	4.143	6.204	4.359	5.764	4.124	4.855
54	Port.Talbot.Margam	5.874	5.874	5.558	7.084	8.237	5.889	8.151	5.494	5.758
55	Portsmouth	4.897	4.862	4.313	4.612	4.612	3.840	3.973	3.755	3.794
56	Preston	4.672	4.573	4.372	5.149	5.149	4.565	4.565	4.188	4.188
57	Reading.New.Town	4.164	4.045	4.901	3.848	4.271	3.364	3.769	3.656	3.782
58	Rochester.Stoke	5.816	5.816	6.001	5.284	7.147	5.990	7.368	5.474	5.781
59	Salford.Eccles	6.214	5.989	5.325	6.763	6.763	4.663	4.663	5.228	5.228
60	Saltash.Callington.Road	4.711	4.984	4.554	4.143	4.584	3.361	5.043	4.033	4.359
61	Sandy.Roadside	4.256	4.507	5.272	4.970	5.724	4.335	5.623	4.130	4.459
62	Sheffield.Barnsley.Road	9.247	9.331	7.091	6.187	8.431	6.774	7.168	7.417	7.827
63	Sheffield.Devonshire.Green	6.631	6.630	5.224	6.187	5.609	4.191	5.108	4.752	5.248
64	Southampton.Centre	5.647	6.009	5.497	4.612	4.612	4.673	4.768	5.022	5.024
65	Southend.on.Sea	4.889	5.608	4.616	5.284	5.645	4.002	5.205	4.255	4.777
66	Stanford.le.Hope.Roadside	4.653	5.283	4.565	5.333	7.941	4.050	7.048	3.931	4.835
67	Stockton.on.Tees.Eaglescliffe	5.451	5.453	5.118	5.413	6.068	5.130	5.778	4.868	4.931
68	Stoke.on.Trent.Centre	4.865	4.873	4.865	5.217	8.320	5.681	6.972	4.753	4.790
69	Sunderland.Silksworth	4.635	4.879	3.934	4.312	4.312	3.785	4.313	3.684	3.848
70	Swansea.Roadside	6.669	7.548	6.737	7.084	8.626	7.000	7.526	6.574	7.084
71	Warrington	4.104	4.710	3.536	4.003	4.003	3.833	4.424	3.368	3.785
72	Wigan.Centre	5.833	5.584	4.855	5.661	5.409	4.559	4.556	5.059	4.862
73	Wirral.Tranmere	4.343	4.510	4.215	3.782	6.497	3.606	6.345	3.682	4.297
74	Worthing.A27.Roadside	5.784	6.357	5.664	6.710	9.480	5.145	7.867	5.340	6.220
75	Wrexham	4.395	5.096	4.395	4.855	6.384	4.230	5.487	4.449	4.479
76	York.Bootham	5.280	5.306	4.824	3.462	5.342	3.599	4.511	3.617	4.080
77	York.Fishergate	4.999	5.098	4.998	3.462	6.108	3.670	4.992	3.479	4.538
	Average RMSE	5.234	5.427	4.860	5.255	5.814	4.686	5.168	4.500	4.698
	Std	1.253	1.226	1.319	1.529	1.586	1.270	1.299	1.331	1.256

E.2 Experiment 2: The basic k-means clustering algorithm.

Table E.5: The RMSE between observed and imputed TS using MVTS clustering imputation from Experiment 2 for stations that measure O_3

	Site	model 1 (CA)	model 2 (CA+ENV)	model 3 (CA+REG)	model 4 (1NN)	model 5 (1NN.ENV)	model 6 (2NN)	model 7 (2NN.ENV)	model 8 (Median)	model 9 (Median.ENV)
1	Aberdeen	17.298	17.159	17.298	22.026	23.017	21.212	19.089	17.280	17.527
2	Aston.Hill	18.983	12.538	18.983	23.324	27.953	22.019	20.143	17.799	17.650
3	Auchencorth.Moss	11.362	8.423	12.427	7.982	7.982	8.157	7.136	7.799	7.227
4	Barnsley.Gawber	9.975	10.059	10.934	9.919	9.919	11.211	9.174	9.168	8.762
5	Belfast.Centre	19.540	18.276	22.877	22.877	17.856	16.649	18.745	17.761	17.407
6	Birmingham.A4540.Roadside	14.275	14.275	13.958	15.888	25.449	16.436	18.087	14.190	12.264
7	Birmingham.Acocks.Green	11.451	11.840	9.071	15.888	9.716	11.114	7.923	10.164	8.877
8	Blackpool.Marton	15.762	16.908	15.870	14.122	14.122	14.148	14.148	15.060	15.060
9	Bournemouth	14.818	15.588	14.520	18.842	18.842	13.244	17.138	13.530	13.950
10	Brighton.Preston.Park	13.458	14.118	12.225	19.293	15.163	12.523	14.065	11.778	12.751
11	Bristol.St.Paul.s	13.606	11.940	15.682	13.977	13.977	11.357	11.357	11.988	11.988
12	Bush.Estate	12.491	10.978	12.678	7.982	7.982	9.193	9.451	9.136	9.066
13	Canterbury	14.853	16.743	14.622	12.455	14.476	12.135	15.036	13.349	14.468
14	Cardiff.Centre	13.955	12.983	15.571	14.792	14.792	12.106	12.106	12.776	12.776
15	Chilbolton.Observatory	15.898	14.801	13.124	19.427	17.816	18.201	14.911	13.921	13.376
16	Coventry.Allesley	10.923	10.824	8.626	9.957	9.957	7.668	7.668	8.302	8.302
17	Cwmbran	12.253	12.628	13.854	14.792	14.792	12.810	12.810	11.845	11.845
18	Eskdalemuir	12.359	11.028	13.058	13.058	11.289	10.408	11.710	10.582	10.378
19	Exeter.Roadside	21.261	21.261	20.743	29.358	20.289	23.559	21.189	21.773	19.752
20	Fort.William	16.287	16.287	20.045	20.045	26.997	14.191	27.650	14.419	16.881
21	Glasgow.Townhead	16.481	14.417	22.068	22.332	21.180	22.068	14.847	21.251	16.092
22	Glazebury	10.453	22.491	8.513	11.397	20.419	11.630	23.228	8.715	18.409
23	High.Muffles	22.384	18.793	22.652	22.923	14.606	20.243	18.793	20.483	18.404
24	Hull.Freetown	14.257	14.349	13.165	21.261	17.017	13.345	15.369	12.788	13.489
25	Ladybower	16.133	11.822	22.190	17.803	20.419	17.018	13.764	15.467	15.282
26	Leamington.Spa	13.686	13.615	13.686	9.957	9.957	10.095	10.095	12.807	12.807
27	Leeds.Centre	14.264	13.626	14.987	13.524	13.524	17.408	13.409	13.973	13.261
28	Leicester.University	12.537	12.220	13.665	19.208	11.505	11.774	9.713	12.434	11.271
29	Leominster	13.540	13.540	13.540	23.324	21.670	15.393	18.451	13.540	13.540
30	Liverpool.Speke	10.938	10.938	9.279	9.389	17.070	9.866	14.529	9.100	10.235
31	London.Bloomsbury	17.769	16.116	11.114	23.067	14.287	11.525	14.556	10.950	13.655
32	London.Eltham	10.885	10.885	13.505	14.370	21.670	9.662	23.877	8.897	11.395
33	London.Haringey.Priory.Park.South	10.777	9.965	13.231	16.045	16.045	22.283	10.748	11.583	10.251
34	London.Harlington	14.715	14.715	10.201	15.630	27.699	9.741	23.633	9.642	14.034
35	London.Hillingdon	22.850	21.384	15.366	15.630	21.127	17.412	19.201	17.289	19.086
36	London.Marylebone.Road	32.707	32.707	25.305	23.067	25.449	26.183	25.020	26.720	28.501
37	London.N Kensington	10.680	9.683	12.089	30.677	14.287	20.966	8.690	12.074	8.548
38	Lough.Navar	19.273	21.499	22.877	22.877	21.097	21.676	20.563	20.733	19.836
39	Lullington.Heath	20.337	15.699	18.395	19.293	19.688	20.417	16.509	18.007	17.003
40	Manchester.Piccadilly	20.549	20.077	22.016	21.145	23.117	19.941	22.117	20.517	21.317
41	Manchester.Sharston	11.953	11.953	11.506	21.145	21.145	13.461	13.461	11.247	11.247
42	Market.Harborough	15.846	11.762	13.050	19.208	14.576	14.060	10.172	12.834	10.967
43	Middlesbrough	15.328	15.328	12.063	13.466	17.070	16.262	17.437	12.930	12.554
44	Narberth	16.695	10.652	19.137	17.995	12.051	19.761	10.652	17.071	10.295
45	Newcastle.Centre	18.131	15.106	13.271	14.904	14.904	13.271	17.993	13.444	15.103
46	Northampton.Spring.Park	14.278	16.213	13.050	13.050	16.965	11.697	15.202	10.932	12.345
47	Norwich.Lakenfields	12.362	14.397	10.936	15.030	14.349	15.774	14.599	10.685	12.399
48	Nottingham.Centre	12.268	11.480	14.436	11.505	11.505	14.071	10.155	11.963	10.989
49	Peebles	13.263	18.746	13.058	12.754	21.180	12.598	20.910	11.257	16.421
50	Plymouth.Centre	13.891	15.808	13.171	19.091	18.971	13.490	17.562	13.207	15.165
51	Port.Talbot.Margam	12.805	12.805	12.708	16.548	19.600	15.351	20.982	12.737	12.852
52	Portsmouth	13.791	13.189	14.577	14.669	14.669	14.478	13.741	12.907	12.739
53	Preston	9.996	10.685	8.346	14.122	14.122	9.785	9.785	8.369	8.369
54	Reading.New.Town	11.402	10.793	13.250	23.021	23.021	17.506	15.039	10.509	10.534
55	Rochester.Stoke	13.426	13.194	13.982	11.847	13.040	13.116	13.835	11.401	11.731
56	Sheffield.Devonshire.Green	10.733	10.639	11.924	17.803	9.919	11.393	10.790	10.645	9.676
57	Sibton	17.281	11.295	12.100	15.030	13.164	11.970	10.197	12.065	11.109
58	Southampton.Centre	14.656	13.748	16.010	14.669	14.669	14.075	14.146	13.765	13.646
59	Southend.on.Sea	9.825	11.374	10.995	11.847	12.947	9.061	10.343	8.357	9.171
60	St.Osyth	13.500	11.284	10.316	12.538	13.040	11.337	10.252	10.338	9.808
61	Stoke.on.Trent.Centre	9.841	10.015	12.300	14.214	15.250	12.197	11.741	10.102	10.387
62	Strathvaich	20.128	16.329	20.045	20.045	16.591	20.305	15.521	18.154	15.638
63	Sunderland.Silksworth	13.195	14.393	12.106	14.904	14.904	12.106	16.576	11.799	12.191
64	Thurrock	11.776	10.694	17.795	10.057	12.947	11.174	10.329	10.846	11.123
65	Walsall.Woodlands	12.030	12.648	11.435	18.442	10.223	12.617	10.101	12.167	10.403
66	Weybourne	21.946	15.650	17.983	19.389	14.433	15.351	15.216	17.114	15.969
67	Wicken.Fen	13.052	12.203	11.958	14.683	14.576	11.570	12.128	11.087	11.239
68	Wigan.Centre	12.104	12.483	10.531	11.397	11.108	9.615	14.230	10.274	11.091
69	Wirral.Tranmere	11.126	12.043	10.340	9.389	14.078	9.580	10.898	9.643	10.246
70	Yarner.Wood	16.425	14.649	19.807	29.358	14.891	22.485	13.442	19.159	13.095
	Average RMSE	14.701	14.125	14.517	16.729	16.345	14.521	14.772	13.237	13.217
	Std	3.968	3.846	3.996	5.075	4.779	4.323	4.604	3.825	3.661

Table E.6: The RMSE between observed and imputed TS using MVTS clustering imputation from Experiment 2 for stations that measure NO_2

	Site	model 1 (CA)	model 2 (CA+ENV)	model 3 (CA+REG)	model 4 (1NN)	model 5 (1NN.ENV)	model 6 (2NN)	model 7 (2NN.ENV)	model 8 (Median)	model 9 (Median.ENV)
1	Aberdeen	15.992	15.815	20.459	29.015	17.420	29.743	16.399	20.244	15.106
2	Aberdeen.Union.Street.Roadside	24.143	19.423	21.987	20.172	20.172	18.484	19.434	19.277	19.329
3	Aberdeen.Wellington.Road	29.142	24.113	27.363	20.172	20.172	23.632	23.440	24.426	24.313
4	Armagh.Roadside	12.650	13.636	11.873	33.238	33.238	20.966	18.502	12.804	12.483
5	Aston.Hill	19.828	3.294	19.828	7.981	14.394	11.974	9.068	11.496	12.111
6	Ballymena..Antrim.Road	12.288	16.026	10.649	12.611	35.085	11.615	20.904	10.217	13.878
7	Ballymena..Ballykeel	12.359	9.596	15.276	12.611	18.465	13.404	10.340	11.341	11.050
8	Barnsley.Gawber	11.099	8.147	11.300	11.929	11.929	18.000	12.721	10.989	10.395
9	Bath.Roadside	23.349	18.897	23.349	29.760	29.760	24.441	19.736	21.574	20.083
10	Belfast.Centre	14.483	18.149	12.822	31.714	18.465	14.922	20.113	12.449	16.424
11	Belfast.Stockman.s.Lane	36.402	30.270	37.076	31.714	35.085	32.374	33.452	33.937	34.584
12	Billingham	12.283	11.147	14.997	12.812	10.809	10.556	9.670	9.446	9.226
13	Birkenhead.Borough.Road	11.401	14.106	10.690	10.394	19.220	11.010	17.262	9.872	12.538
14	Birmingham.A4540.Roadside	13.546	11.035	12.372	19.237	10.501	12.085	11.724	12.377	11.129
15	Birmingham.Acocks.Green	11.159	8.967	11.598	19.237	9.544	18.594	7.439	11.792	8.589
16	Blackburn.Accrington.Road	8.606	14.608	7.948	8.318	12.362	8.729	12.597	8.040	9.819
17	Blackpool.Marton	13.899	10.558	13.708	12.975	12.975	11.413	10.217	11.819	11.626
18	Borehamwood.Meadow.Park	11.317	8.881	10.926	7.789	7.789	21.907	9.911	9.627	8.707
19	Bournemouth	11.415	8.636	9.763	14.811	21.300	20.013	15.455	10.243	10.313
20	Bradford.Mayo.Avenue	28.595	23.681	29.315	27.127	20.725	23.145	23.734	26.722	25.062
21	Brighton.Preston.Park	14.291	11.389	10.882	28.225	9.934	20.827	8.485	13.401	9.903
22	Bristol.St.Paul.s	11.042	12.410	12.260	29.303	10.991	24.023	11.765	10.728	11.323
23	Bristol.Temple.Way	32.280	23.134	35.229	29.303	29.760	26.222	27.726	29.773	27.300
24	Burton.on.Trent..Horninglow	9.069	7.229	10.030	22.144	9.530	13.775	7.973	9.533	7.482
25	Bury.Whitefield.Roadside	10.206	12.378	9.826	11.259	12.296	10.561	9.590	9.488	9.218
26	Bush.Estate	20.650	6.600	27.662	54.774	6.600	35.168	6.342	26.027	6.253
27	Cambridge.Roadside	9.460	16.707	11.321	21.381	13.409	12.888	17.335	10.396	10.096
28	Camden.Kerbside	34.871	24.849	25.461	47.877	47.877	24.260	26.842	23.444	23.005
29	Cannock.A5190.Roadside	9.861	13.349	9.451	12.130	16.021	9.608	15.724	9.140	11.508
30	Canterbury	17.095	13.807	13.007	7.519	11.073	9.815	13.188	12.353	13.232
31	Cardiff.Centre	8.135	8.680	10.772	17.187	9.145	11.146	9.055	9.025	7.925
32	Cardiff.Newport.Road	16.813	13.685	14.872	17.187	47.435	16.042	31.590	15.035	14.009
33	Carlisle.Roadside	10.023	14.117	10.023	18.298	18.298	13.313	19.477	10.101	11.520
34	Charlton..Mackrell	16.564	4.310	15.444	39.022	6.798	39.348	4.121	16.776	6.043
35	Chatham.Roadside	10.903	19.889	10.942	15.568	15.375	11.974	19.559	10.430	12.848
36	Chepstow..A48	23.270	16.898	19.646	23.496	28.248	23.189	28.398	21.308	17.473
37	Chesterfield.Loundsley.Green	14.645	11.341	18.052	10.829	14.589	11.300	16.284	12.311	13.706
38	Chesterfield.Roadside	10.945	16.379	13.462	10.829	23.535	8.951	22.085	10.590	15.576
39	Chilbolton.Observatory	19.522	5.917	15.108	33.535	6.798	27.664	6.160	17.685	5.497
40	Christchurch.Barrack.Road	12.574	19.222	12.487	14.811	29.079	16.774	23.416	12.208	16.969
41	Coventry.Allesley	9.888	8.614	9.548	13.653	9.901	7.992	8.410	8.496	7.653
42	Coventry.Binley.Road	11.726	10.523	10.292	13.653	16.540	14.507	9.574	10.648	10.305
43	Crumbran	11.127	8.327	16.647	10.853	10.853	34.379	10.373	11.912	10.147
44	Derby.St..Alkmund.s.Way	18.064	13.691	17.902	22.144	13.048	15.004	15.948	16.768	15.360
45	Derry.Rosemount	16.716	12.493	22.524	19.120	11.941	14.640	15.993	16.123	14.391
46	Dewsbury.Ashworth.Grove	7.601	8.912	6.534	27.127	11.570	15.573	7.220	6.638	6.728
47	Doncaster.A630.Cleveland.Street	10.533	9.935	10.192	13.466	15.490	11.368	10.816	9.693	9.140
48	Dumbarton.Roadside	11.836	18.261	14.872	25.019	25.019	19.419	19.419	15.755	15.755
49	Dumfries	17.357	14.379	33.314	35.285	18.298	25.088	16.739	25.074	16.521
50	Dundee.Mains.Loan	14.818	10.586	26.863	13.660	13.660	28.912	9.612	14.563	11.676
51	Eastbourne	17.651	14.155	13.309	7.685	9.934	5.022	12.947	11.973	13.245
52	Edinburgh..Nicolson.Street	38.717	32.841	36.817	40.976	32.013	47.554	29.359	38.357	33.673
53	Edinburgh.St.Leonards	10.908	10.141	13.830	40.976	16.673	15.683	14.071	11.350	9.928
54	Eskdalemuir	24.773	6.600	19.907	35.285	6.600	19.907	4.942	19.958	6.981
55	Exeter.Roadside	16.536	12.643	18.341	30.439	19.973	28.418	14.559	18.492	15.889
56	Fort.William	17.971	17.971	17.285	17.285	11.036	22.107	11.985	18.238	13.918
57	Glasgow..Great.Western.Road	15.300	14.335	11.862	41.164	41.164	20.265	22.463	12.736	12.951
58	Glasgow..High.Street	15.498	13.186	11.788	11.606	40.379	19.070	21.368	11.096	11.658
59	Glasgow..Kerbside	46.840	40.831	45.008	45.561	40.379	42.655	40.373	44.277	43.037
60	Glasgow..Townhead	10.571	13.265	9.927	11.606	13.151	26.413	16.710	8.724	11.362
61	Glazebury	13.759	12.755	12.952	15.081	12.291	12.153	13.074	11.262	11.693
62	Grangemouth	13.218	7.188	17.601	6.580	6.580	24.526	8.167	12.560	7.797
63	Grangemouth.Moray	12.645	8.443	16.135	6.580	6.580	22.956	8.832	11.618	7.767
64	Greenock.A8.Roadside	20.364	16.829	19.098	25.019	25.019	21.498	21.498	19.365	19.365
65	Hafod.y.ryns.Roadside	56.299	46.575	53.984	61.327	47.435	57.779	42.258	55.432	49.489
66	Haringey.Roadside	17.120	12.814	9.979	19.227	22.593	14.150	20.500	13.026	11.800
67	Hartlepool.St.Abbs.Walk	14.403	11.523	19.833	10.855	9.884	9.449	12.093	11.287	11.700
68	High.Muffles	22.542	8.621	23.569	15.394	5.766	15.172	8.621	15.224	8.737
69	Honiton	14.459	12.013	13.274	26.750	15.904	12.372	14.617	13.621	13.147
70	Horley	13.466	13.466	11.398	17.552	16.926	23.223	22.181	13.411	13.372
71	Hull.Freetown	10.114	10.202	9.928	10.755	14.547	7.377	12.420	8.399	9.845
72	Hull.Holderness.Road	13.724	14.092	13.659	10.755	15.573	14.946	15.024	12.760	13.677
73	Immingham.Woodlands.Avenue	14.000	11.122	14.653	14.547	14.547	17.185	11.091	13.137	12.381
74	Inverness	13.246	19.027	17.285	17.285	30.429	16.591	30.638	13.804	17.877
75	Ladybower	21.234	6.665	25.099	21.372	12.291	28.398	8.155	21.122	11.583
76	Leamington.Spa	13.271	10.726	13.271	7.008	9.901	9.750	8.596	11.091	10.735
77	Leamington.Spa.Rugby.Road	10.151	17.003	11.185	7.008	16.540	6.254	17.456	9.240	13.326
78	Leeds.Centre	12.211	14.816	11.761	13.576	11.570	10.587	14.257	11.404	12.506
79	Leeds.Headingley.Kerbside	15.497	12.461	15.059	13.576	20.725	10.704	12.671	12.964	12.040
80	Leicester.A594.Roadside	19.347	15.651	18.577	18.537	14.226	25.537	12.292	18.533	15.733
81	Leicester..University	10.613	10.878	10.286	18.537	12.185	8.955	10.203	9.349	9.833
82	Leominster	14.184	14.184	14.184	7.981	13.498	5.258	16.495	14.184	14.184
83	Lincoln.Canwick.Road	21.450	18.249	21.365	28.438	19.931	28.448	18.372	22.122	19.011
84	Liverpool.Speke	12.064	10.209	11.001	25.872	11.601	14.711	10.101	11.547	9.767
85	London..Bexley	9.577	10.179	19.302	10.179	10.179	7.873	14.600	7.793	9.982
86	London..Bloomsbury	15.744	19.522	11.776	60.794	10.066	29.223	11.179	11.761	11.348
87	London..Eltham	12.649	10.179	26.032	10.179	10.179	18.989	6.281	12.343	10.053
88	London..Haringey..Priory.Park.South	9.151	7.196	19.177	19.227	19.279	27.583	14.112	15.440	12.868
89	London..Harlington	13.159	13.159	15.180	26.836	24.017	14.532	22.382	12.027	13.168
90	London..Hillingdon	31.662	35.504	25.240	26.836	29.916	27.752	30.920	28.336	30.581
91	London..Marylebone.Road	68.569	59.092	60.660	60.794	47.877	52.978	49.733	60.305	57.898
92	London N Kensington	10.908	12.136	13.688	33.020	14.098	46.663	11.143	13.339	9.405
93	London..Westminster	10.921	14.159	11.633	10.066	10.066	34.871	7.148	7.324	6.600
94	Lullington..Heath	22.104	6.646	17.711	7.685	10.664	10.592	7.613	10.565	9.890
95	Luton.A505.Roadside	26.709	20.032	31.882	32.314	28.494	29.159	24.145	28.797	25.161
96	Manchester..Piccadilly	14.171	17.930	15.213	14.776	14.776	13.424	16.529	14.868	15.503
97	Manchester.Sharston	10.001	10.001	9.274	16.838	16.926	11.929	10.979	9.521	9.634
98	Market.Harborough	21.177	6.172	9.130	19.904	5.476	26.298	4.811	16.072	5.081
99	Middlesbrough	12.920	12.185	9.022	10.809	10.809	8.688	9.359	8.090	8.380
100	Narberth	19.184	2.854	27.129	25.479	3.082	19.107	2.824	19.600	3.182
101	Newcastle.Centre	12.672	16.739	15.996	19.199	18.790	10.899	18.379	11.871	16.518
102	Newcastle.Cradlewell.Roadside	25.292	21.528	31.167	19.199	24.585	21.064	26.125	22.811	25.375
103	Newport	7.277	8.443	10.017	10.853	10.853	24.616	8.118	6.851	7.391
104	Northampton.Spring.Park	16.223	12.863	9.130	9.130	15.786	8.199	12.073	7.593	10.900
105	Norwich..Lakenfields	17.274	13.900	13.990	7.617	11.760	9.223	14.225	12.468	13.615
106	Nottingham.Centre	8.987	11.286	8.482	14.885	12.862	14.447	11.149	8.094	9.872
107	Nottingham.Western..Boulevard	16.834	12.977	15.871	14.885	13.048	11.210	14.595	14.067	13.790
108	Oldbury.Birmingham.Road	13.857	10.901	12.328	10.501	10.501	12.892	10.499	11.672	10.916
109	Oxford.Centre.Roadside	19.583	14.573	24.022	28.897	21.313	22.906	17.949	21.667	17.276
110	Oxford.St.Ebbes	14.325	10.767	10.603	28.897	14.867	23.322	8.886	13.501	10.949
111	Peebles	20.074	15.383	14.521	5.550	16.673	26.960	18.921	15.981	16.036
112	Plymouth.Centre	9.690	10.557	9.502	12.551	15.690	12.545	11.474	9.538	10.333
113	Plymouth..Tavistock.Road	11.199	20.192	11.763	12.551	19.973	14.058	16.210	11.239	14.781
114	Port..Talbot..Margam	10.034	10.034	15.209	16.840	13.342	13.052	12.671	11.982	10.675
115	Portsmouth	13.107	10.966	10.290	18.308	16.657	15.639	10.328	11.559	10.822
116	Portsmouth.Anglesea.Road	14.172	16.554	16.135	18.308	21.866	14.891	15.623	14.352	14.423
117	Preston	8.981	8.430	7.831	8.318	12.975	9.332	10.776	7.411	7.945
118	Reading..London.Road	15.342	18.565	17.293	12.496	21.313	22.168	25.177	14.192	16.663
119	Reading..New..Town	9.295	9.432	10.582	12.496	33.883	17.375	16.286	8.371	9.544
120	Rochester..Stoke	16.454	8.254	12.743	9.840	8.011	11.340	8.156	10.208	7.287
121	Salford.Eccles	9.506	10.506	8.551	14.776	14.776	11.008	9.780	8.565	8.664
122	Sandy.Roadside	13.712	19.271	14.467	28.494	28.494	17.002	17.002	15.918	15.918
123	Scunthorpe.Town	11.974	9.405	12.556	12.357	11.163	7.707	9.967	9.775	9.136
124	Shaw.Crompton.Way	12.845	12.071	13.906	12.296	12.296	12.100	12.549	11.504	12.058
125	Sheffield.Barnsley.Road	18.284	13.457	18.460	19.004	23.535	17.241	18.143	17.123	17.705
126	Sheffield..Devonshire.Green	9.178	9.260	8.648	19.004	13.588	13.837	9.812	8.789	8.773
127	Sheffield.Tinsley	16.743	12.526	11.323	18.046	13.588	12.840	13.758	10.697	11.821
128	Southampton..A33	20.418	18.779	23.915	17.744	21.866	24.059	23.223	20.364	20.929
129	Southampton.Centre	11.221	13.300	13.299	17.744	16.657	12.704	18.025	10.984	13.393
130	Southend.on.Sea	11.099	8.418	8.242	9.840	11.880	7.073	7.610	7.284	8.011
131	Southwark.A2.Old.Kent.Road	24.390	20.236	20.458	21.665	21.665	16.966	38.892	18.446	19.368
132	St..Helens..Linkway	15.485	12.122	16.119	14.567	14.567	13.223	13.418	12.796	12.583
133	St.Osyth	17.038	6.968	14.140	10.564	8.011	7.980	6.627	9.563	7.182
134	Stanford.le.Hope..Roadside	11.710	18.186	12.911	10.417	15.375	13.318	15.731	10.749	11.925
135	Stockton.on..Tees..A1305.Roadside	13.294	19.004	9.817	12.812	9.757	9.292	9.825	9.931	9.210
136	Stockton.on..Tees.Eaglescliffe	14.800	22.586	10.754	9.757	9.757	7.718	11.624	10.410	11.205
137	Stoke.on.Trent..A50.Roadside	37.648	31.553	38.897	38.197	38.694	37.917	32.288	37.191	35.929
138	Stoke.on.Trent.Centre	9.012	10.142	9.572	38.197	12.246	20.292	12.198	9.424	9.837
139	Storrington.Roadside	11.505	15.650	11.505	19.971	19.971	11.636	16.926	10.422	11.698
140	Sunderland.Silksworth	13.385	10.043	9.521	12.914	18.790	21.902	11.513	10.680	10.739
141	Sunderland.Wessington.Way	12.010	16.504	11.737	12.914	24.585	10.578	14.308	9.705	10.982
142	Swansea.Roadside	12.318	16.183	11.747	16.840	51.768	20.266	29.888	12.118	16.135
143	Swindon.Walcot	10.324	7.931	9.350	8.209	8.209	17.331	9.767	8.967	7.841
144	Telford.Hollinswood	12.031	9.511	13.636	8.129	8.129	9.256	9.378	9.711	9.760
145	Thurrock	10.267	10.552	10.375	9.974	11.880	8.102	10.481	8.963	9.620
146	Tower..Hamlets.Roadside	26.494	18.659	19.700	21.665	21.665	18.600	26.651	19.348	18.387
147	Walsall.Woodlands	11.140	8.106	12.571	12.130	8.800	15.120	7.156	10.889	8.317
148	Warrington	9.918	10.504	9.191	22.557	11.601	19.932	10.453	9.980	9.489
149	Wicken.Fen	20.896	5.201	17.915	21.381	5.476	22.099	5.347	17.888	6.240
150	Widnes.Milton.Road	19.389	14.512	20.228	25.872	14.567	23.568	14.410	19.961	16.014
151	Wigan.Centre	10.689	8.278	9.912	21.098	13.515	10.396	10.902	9.846	9.461
152	Wirral.Tranmere	11.592	9.822	10.460	10.394	11.074	7.041	10.705	8.909	9.866
153	Worthing.A27.Roadside	19.325	14.710	22.774	19.971	19.971	23.026	16.863	19.866	17.499
154	Wrexham	12.234	17.473	12.234	11.045	12.892	9.975	17.062	12.199	12.870
155	Yarner.Wood	18.659	3.337	17.763	30.439	4.642	23.968	3.560	18.301	4.637
156	York.Bootham	12.415	9.421	12.587	15.332	18.900	15.568	15.635	11.960	12.319
157	York.Fishergate	9.628	10.602	9.706	15.332	15.891	11.652	12.325	9.459	9.593
	Average RMSE	16.024	13.965	16.112	19.447	17.269	18.052	15.254	14.561	13.531
	Std	8.443	7.355	8.333	11.064	9.516	9.258	7.830	7.958	7.502

Table E.7: The RMSE between observed and imputed TS using MVTS clustering imputation from Experiment 2 for stations that measure PM_{10}

	Site	model 1 (CA)	model 2 (CA+ENV)	model 3 (CA+REG)	model 4 (1NN)	model 5 (1NN.ENV)	model 6 (2NN)	model 7 (2NN.ENV)	model 8 (Median)	model 9 (Median.ENV)
1	Aberdeen	8.589	8.830	8.589	10.681	9.360	9.288	9.362	8.568	8.538
2	Armagh.Roadside	19.540	19.044	18.731	18.905	18.905	18.004	19.241	18.714	18.927
3	Auchencorth.Moss	8.060	6.686	6.781	6.180	6.686	6.271	6.643	5.866	5.786
4	Barnstaple.A39	7.069	7.585	6.896	21.738	10.721	14.096	8.219	7.177	7.139
5	Belfast.Centre	7.786	8.943	8.059	8.506	11.448	10.506	9.287	7.110	8.412
6	Belfast.Stockman.s.Lane	8.352	8.347	8.718	8.506	18.905	11.537	11.427	7.711	8.500
7	Birmingham.A4540.Roadside	7.568	7.477	6.496	6.554	8.087	6.141	7.085	6.352	6.626
8	Birmingham.Ladywood	5.252	6.155	4.839	6.554	5.294	6.170	5.014	4.428	4.340
9	Bristol.St.Paul.s	6.931	6.244	7.701	10.850	6.646	8.155	6.724	6.827	5.918
10	Bristol.Temple.Way	10.390	9.992	10.995	10.850	11.542	10.247	9.566	10.223	10.029
11	Bury.Whitefield.Roadside	5.684	6.197	5.567	9.641	8.157	6.432	7.610	5.100	5.385
12	Camden.Kerbside	7.027	6.263	5.759	8.789	8.789	6.449	10.581	5.808	5.870
13	Cardiff.Centre	13.170	12.749	13.618	14.400	13.801	13.373	12.448	13.032	12.962
14	Cardiff.Newport.Road	7.587	6.940	8.455	14.400	9.132	9.809	7.357	7.668	7.108
15	Carlisle.Roadside	7.516	7.966	7.516	12.541	10.035	9.977	9.289	7.465	7.338
16	Chatham.Roadside	9.958	8.610	11.629	11.516	9.483	9.541	8.542	9.560	9.193
17	Chepstow.A48	7.367	7.680	7.973	9.049	11.542	7.472	9.243	7.148	7.489
18	Chesterfield.Loundsley.Green	5.701	5.183	7.375	6.667	5.769	5.040	5.241	5.055	4.970
19	Chesterfield.Roadside	7.023	7.523	7.578	6.667	9.599	6.389	8.443	6.644	7.392
20	Chilbolton.Observatory	7.889	11.963	6.583	8.748	11.963	7.918	7.458	6.865	8.039
21	Coventry.Binley.Road	7.542	7.037	6.727	8.735	8.735	8.475	7.250	6.980	6.824
22	Derry.Rosemount	9.666	10.234	10.095	10.857	11.448	11.995	9.763	9.013	9.425
23	Ealing.Horn.Lane	16.836	15.577	15.799	18.484	16.070	16.753	14.879	16.373	15.642
24	Edinburgh.St.Leonards	5.401	6.154	4.444	6.180	5.401	4.372	6.774	4.024	4.640
25	Glasgow.High.Street	5.565	6.646	5.697	4.978	8.674	5.645	6.445	4.891	5.380
26	Glasgow.Townhead	5.140	6.088	3.942	4.978	5.401	5.466	5.999	3.745	4.582
27	Grangemouth	5.371	10.603	4.744	5.198	10.603	4.827	8.747	4.466	8.185
28	Greenock.A8.Roadside	8.310	9.020	8.117	8.520	8.674	8.229	9.104	7.934	8.054
29	Hull.Holderness.Road	9.830	9.670	9.371	10.592	9.499	8.572	10.048	8.881	9.084
30	Inverness	7.000	8.860	7.000	10.681	8.836	8.199	7.946	7.153	7.247
31	Leamington.Spa	7.208	5.942	7.208	4.289	5.294	5.329	4.282	5.948	5.885
32	Leamington.Spa.Rugby.Road	6.109	6.787	6.630	4.289	8.735	5.311	7.641	5.197	6.347
33	Leeds.Centre	6.117	6.560	5.964	8.032	7.666	6.430	6.413	5.359	5.906
34	Leeds.Headingley.Kerbside	8.950	9.139	8.719	8.032	10.168	8.153	9.216	8.323	8.705
35	Leicester.A594.Roadside	9.851	9.204	9.979	10.011	8.293	8.163	7.683	9.225	8.885
36	Liverpool.Speke	6.754	7.492	6.356	9.222	5.017	6.061	7.492	5.854	5.766
37	London.Bloomsbury	5.096	5.941	5.953	10.043	5.461	7.561	5.090	5.222	4.640
38	London.Harlington	5.147	5.147	7.567	17.868	22.803	9.273	13.891	6.557	5.733
39	London.Marylebone.Road	9.563	7.904	8.509	10.043	8.789	8.719	9.254	8.533	7.909
40	London N Kensington	5.181	4.220	7.885	18.484	5.461	12.035	5.823	7.533	4.785
41	Lough.Navar	8.408	6.686	11.688	10.857	6.686	14.156	6.699	8.700	6.239
42	Middlesbrough	8.942	10.603	8.116	9.273	11.890	8.116	8.662	7.218	7.470
43	Narberth	8.538	8.538	9.636	11.196	8.204	14.857	7.238	8.811	8.196
44	Newcastle.Centre	10.733	11.005	11.293	10.853	10.732	10.036	10.673	9.501	9.942
45	Newcastle.Cradlewell.Roadside	8.674	9.043	8.674	10.853	10.967	8.522	8.802	8.541	8.613
46	Newport	5.884	7.118	6.934	8.961	13.801	9.141	9.052	5.888	6.983
47	Norwich.Lakenfields	7.512	7.088	7.803	8.211	9.448	7.619	8.339	7.141	7.577
48	Nottingham.Centre	5.703	6.501	5.719	7.017	7.685	7.255	7.656	5.277	6.083
49	Nottingham.Western.Boulevard	6.939	6.721	6.557	7.017	9.599	7.122	7.247	6.347	6.689
50	Oxford.St.Ebbes	7.650	5.768	7.137	8.894	5.085	6.016	4.323	6.353	5.573
51	Plymouth.Centre	7.051	9.371	7.019	5.215	14.998	5.684	10.008	6.033	8.931
52	Port.Talbot.Margam	20.270	20.270	20.402	19.497	23.335	20.577	23.121	20.167	20.441
53	Portsmouth	6.293	5.718	5.062	7.692	6.043	5.809	5.155	5.009	4.970
54	Portsmouth.Anglesea.Road	8.063	8.674	7.745	7.692	7.836	6.975	7.949	7.249	7.323
55	Reading.London.Road	7.187	8.870	7.231	8.071	18.675	7.236	12.684	6.849	8.503
56	Reading.New.Town	7.188	5.562	6.681	8.071	5.085	6.201	4.497	5.996	5.319
57	Rochester.Stoke	9.209	11.963	9.791	11.516	11.963	9.181	12.406	9.048	11.328
58	Salford.Eccles	10.038	10.492	9.479	9.641	11.687	9.396	11.053	9.397	10.177
59	Saltash.Callington.Road	6.962	7.630	6.779	5.215	8.024	5.328	6.906	5.946	6.574
60	Sandy.Roadside	5.947	7.089	6.756	8.664	8.664	7.084	11.114	6.256	6.497
61	Scunthorpe.Town	9.605	11.883	9.065	10.592	11.890	9.434	10.977	8.765	10.121
62	Sheffield.Devonshire.Green	6.386	6.223	7.595	5.769	5.769	5.697	5.609	5.806	5.896
63	Southampton.A33	6.859	8.127	6.349	6.224	7.836	6.370	6.971	6.008	6.220
64	Southampton.Centre	5.849	6.640	5.361	6.224	6.043	5.675	6.528	5.058	5.563
65	Southwark.A2.Old.Kent.Road	9.168	7.900	7.857	8.995	10.352	8.296	7.652	7.869	7.644
66	St.Helens.Linkway	7.759	7.792	7.774	8.571	8.157	8.542	8.641	7.688	7.621
67	Stanford.le.Hope.Roadside	5.933	6.741	6.134	5.867	9.483	5.872	7.955	5.084	5.911
68	Stockton.on.Tees.Eaglescliffe	10.176	11.140	9.728	9.273	10.967	9.728	9.977	8.637	8.929
69	Stoke.on.Trent.A50.Roadside	7.645	7.304	8.582	9.788	9.788	9.386	8.333	7.647	7.550
70	Swansea.Roadside	8.084	8.479	8.291	19.497	9.973	11.234	9.119	7.956	8.252
71	Thurrock	5.997	7.671	6.290	5.867	7.392	6.157	7.247	5.485	6.409
72	Warrington	5.641	6.540	5.194	8.571	5.017	5.297	6.540	4.724	4.820
73	Wrexham	6.741	7.854	6.741	6.182	10.782	5.390	9.714	6.720	7.312
74	York.Bootham	6.456	6.539	6.577	7.726	7.666	6.462	7.130	5.816	6.079
75	York.Fishergate	7.137	7.125	6.534	7.726	10.168	7.078	7.481	6.441	6.548
	Average RMSE	7.943	8.284	7.989	9.500	9.714	8.391	8.613	7.386	7.625
	Std	2.775	2.751	2.793	3.815	3.830	3.121	3.012	2.815	2.802

Table E.8: The RMSE between observed and imputed TS using MVTS clustering imputation from Experiment 2 for stations that measure $PM_{2.5}$

	Site	model 1 (CA)	model 2 (CA+ENV)	model 3 (CA+REG)	model 4 (1NN)	model 5 (1NN.ENV)	model 6 (2NN)	model 7 (2NN.ENV)	model 8 (Median)	model 9 (Median.ENV)
1	Aberdeen	4.500	5.049	4.500	5.804	4.873	4.681	4.682	4.498	4.507
2	Auchencorth.Moss	4.650	5.029	3.939	3.976	5.029	4.237	4.578	3.679	4.230
3	Barnstaple.A39	4.654	5.328	5.073	7.313	8.626	7.169	5.563	4.822	5.056
4	Belfast.Centre	6.783	6.768	7.356	9.295	9.295	7.356	6.797	7.067	6.806
5	Birmingham.A4540.Roadside	5.229	5.478	4.040	4.734	4.780	4.299	5.182	3.999	4.428
6	Birmingham.Acocks.Green	5.012	5.138	3.875	4.734	5.288	4.249	4.320	3.783	3.988
7	Blackpool.Marton	4.794	4.747	4.494	5.149	5.149	5.046	5.046	4.199	4.199
8	Bournemouth	5.013	5.491	4.566	4.593	6.254	4.434	5.400	4.352	4.605
9	Bristol.St.Paul.s	4.965	4.650	5.752	6.353	4.946	5.089	4.569	4.738	4.659
10	Camden.Kerbside	3.693	4.199	2.945	6.216	6.216	3.383	4.068	3.223	3.391
11	Cardiff.Centre	3.994	4.027	4.008	3.797	3.797	4.081	3.653	3.543	3.502
12	Carlisle.Roadside	4.658	5.175	4.658	6.361	7.224	6.155	5.266	4.764	4.768
13	Chatham.Roadside	7.316	6.894	7.982	8.206	7.941	7.622	6.692	7.205	7.044
14	Chepstow.A48	5.399	6.091	5.272	6.353	8.053	5.206	6.216	5.162	5.478
15	Chesterfield.Loundsley.Green	4.143	4.349	4.322	5.225	5.609	4.390	4.294	3.828	3.799
16	Chesterfield.Roadside	5.729	5.912	5.786	5.225	8.431	5.100	6.892	5.258	5.591
17	Chilbolton.Observatory	5.115	7.147	4.567	5.946	7.147	5.031	4.848	4.449	5.185
18	Christchurch.Barrack.Road	6.049	6.693	5.711	4.593	6.649	4.654	6.387	5.489	5.923
19	Coventry.Allesley	5.343	5.511	4.283	4.545	4.691	4.221	4.460	4.094	4.210
20	Derry.Rosemount	9.382	9.494	8.568	9.754	9.295	8.568	8.828	8.804	8.763
21	Eastbourne	5.893	6.204	5.615	6.710	6.723	5.626	6.464	5.576	5.945
22	Edinburgh.St.Leonards	3.536	4.466	3.060	3.976	3.905	3.239	4.178	2.887	3.394
23	Glasgow.High.Street	3.895	4.241	3.497	3.136	4.442	3.331	4.079	3.216	3.639
24	Glasgow.Townhead	3.313	4.247	2.706	3.136	3.905	2.712	4.353	2.435	3.134
25	Grangemouth	4.187	7.262	4.072	4.505	7.262	4.205	6.530	3.938	5.941
26	Greenock.A8.Roadside	4.084	4.461	3.367	3.840	4.442	3.845	4.732	3.438	3.832
27	Hull.Freetown	6.900	6.906	7.316	6.230	6.667	6.216	6.883	6.367	6.503
28	Inverness	5.016	5.735	5.016	5.804	5.362	5.253	5.278	5.044	5.101
29	Leamington.Spa	4.996	4.788	4.996	3.748	4.691	3.587	3.768	4.682	4.678
30	Leamington.Spa.Rugby.Road	5.364	5.893	3.976	3.748	4.780	3.443	4.038	3.788	3.937
31	Leeds.Centre	5.025	5.320	3.962	3.724	5.342	3.496	4.595	3.618	4.304
32	Leeds.Headingley.Kerbside	5.581	5.743	4.635	3.724	6.108	4.276	5.192	4.356	5.016
33	Leicester.University	4.549	4.619	5.172	4.975	4.975	4.037	4.037	4.195	4.195
34	Liverpool.Speke	4.200	4.003	3.802	3.782	4.003	3.199	4.348	3.162	3.512
35	London.Bexley	3.960	3.966	3.907	3.966	3.966	3.694	3.694	3.604	3.604
36	London.Bloomsbury	3.633	3.930	2.909	7.026	3.780	4.727	2.856	3.150	3.004
37	London.Eltham	3.233	3.966	3.115	3.966	3.966	3.256	3.256	3.049	3.049
38	London.Harlington	3.270	3.270	3.646	3.762	8.155	2.529	6.206	3.045	3.179
39	London.Marylebone.Road	6.790	6.226	6.543	7.026	6.216	6.411	6.743	6.415	6.340
40	London N Kensington	3.166	3.214	3.137	3.931	3.260	5.410	3.217	3.134	2.808
41	London.Teddington.Bushy.Park	3.283	3.545	3.319	3.762	4.089	3.700	3.664	3.144	3.305
42	London.Westminster	3.499	3.824	2.929	3.780	3.780	3.710	3.618	2.930	3.053
43	Lough.Navar	5.264	5.029	7.790	9.754	5.029	7.790	4.433	7.082	4.496
44	Manchester.Piccadilly	7.474	7.416	7.226	6.763	6.763	6.700	6.859	6.969	6.974
45	Middlesbrough	5.843	7.262	4.116	5.413	6.734	4.222	6.395	4.122	4.992
46	Narberth	4.427	4.427	4.978	8.369	5.738	6.687	3.953	4.833	4.196
47	Newcastle.Centre	5.312	5.557	4.260	4.312	4.312	4.156	4.339	3.877	3.883
48	Newport	3.484	3.620	3.716	3.797	3.797	3.501	3.649	3.040	3.089
49	Northampton.Spring.Park	4.897	4.766	4.897	4.492	4.492	4.039	4.039	4.653	4.653
50	Norwich.Lakenfields	5.800	5.804	5.833	6.361	6.977	5.892	5.553	5.599	5.621
51	Nottingham.Centre	4.548	4.560	4.810	7.116	4.975	4.899	4.460	4.650	4.201
52	Oxford.St.Ebbes	4.450	4.198	4.823	4.271	4.271	3.815	3.815	4.002	4.002
53	Plymouth.Centre	4.466	5.347	4.444	4.143	6.204	4.350	5.764	4.138	5.058
54	Port.Talbot.Margam	5.798	5.798	5.685	7.084	8.237	5.889	8.151	5.559	5.782
55	Portsmouth	5.000	4.951	4.172	4.612	4.612	3.840	3.973	4.051	4.041
56	Preston	4.640	4.618	4.468	5.149	5.149	4.565	4.565	4.134	4.134
57	Reading.New.Town	4.247	4.048	4.774	3.848	4.271	3.364	3.769	3.830	3.907
58	Rochester.Stoke	5.499	7.147	5.969	5.284	7.147	5.990	7.368	5.277	6.740
59	Salford.Eccles	5.928	5.802	5.268	6.763	6.763	4.663	4.663	5.077	5.077
60	Saltash.Callington.Road	4.626	5.098	4.554	4.143	4.584	3.361	5.043	4.086	4.419
61	Sandy.Roadside	4.543	5.067	5.272	4.970	5.724	4.335	5.623	4.237	4.727
62	Sheffield.Barnsley.Road	8.173	7.904	7.415	6.187	8.431	6.774	7.168	7.281	7.551
63	Sheffield.Devonshire.Green	5.659	6.085	5.465	6.187	5.609	4.191	5.108	4.822	5.179
64	Southampton.Centre	5.276	5.305	4.818	4.612	4.612	4.673	4.768	4.635	4.634
65	Southend.on.Sea	4.446	5.040	4.616	5.284	5.645	4.002	5.205	4.069	4.584
66	Stanford.le.Hope.Roadside	4.298	5.145	4.565	5.333	7.941	4.050	7.048	3.775	4.897
67	Stockton.on.Tees.Eaglescliffe	6.523	7.316	5.118	5.413	6.068	5.130	5.778	5.029	5.084
68	Stoke.on.Trent.Centre	4.346	4.376	5.369	5.217	8.320	5.681	6.972	4.181	4.216
69	Sunderland.Silksworth	5.217	5.592	3.934	4.312	4.312	3.785	4.313	3.733	3.795
70	Swansea.Roadside	6.678	7.145	6.698	7.084	8.626	7.000	7.526	6.564	7.037
71	Warrington	3.828	4.003	3.591	4.003	4.003	3.833	4.424	3.342	3.458
72	Wigan.Centre	5.367	5.354	4.758	5.661	5.409	4.559	4.556	4.899	4.695
73	Wirral.Tranmere	4.630	4.733	4.295	3.782	6.497	3.606	6.345	3.801	4.495
74	Worthing.A27.Roadside	5.505	6.991	5.011	6.710	9.480	5.145	7.867	5.071	6.617
75	Wrexham	4.343	5.663	4.343	4.855	6.384	4.230	5.487	4.486	4.705
76	York.Bootham	4.943	5.081	4.310	3.462	5.342	3.599	4.511	3.683	4.112
77	York.Fishergate	4.638	5.286	4.344	3.462	6.108	3.670	4.992	3.630	4.340
	Average RMSE	4.986	5.332	4.780	5.255	5.814	4.686	5.168	4.472	4.688
	Std	1.155	1.197	1.246	1.529	1.586	1.270	1.299	1.213	1.194

Appendix F

Conditional quantile plots for each pollutant for each station.

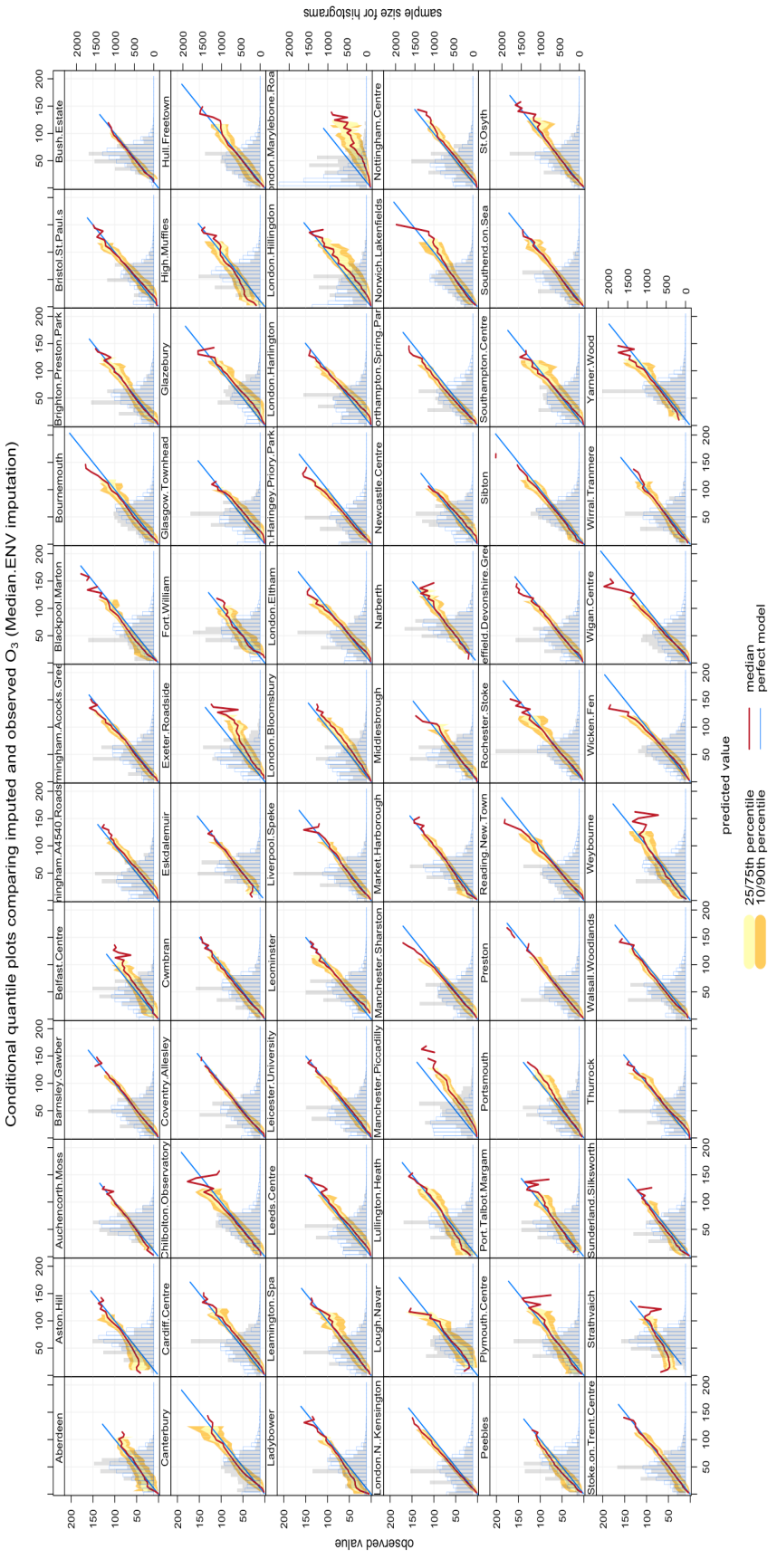


Figure F.1: Conditional quantile plot of modelled and observed pollutants concentrations of O_3 based on model 9 (Median.ENV) for all stations (70 stations).

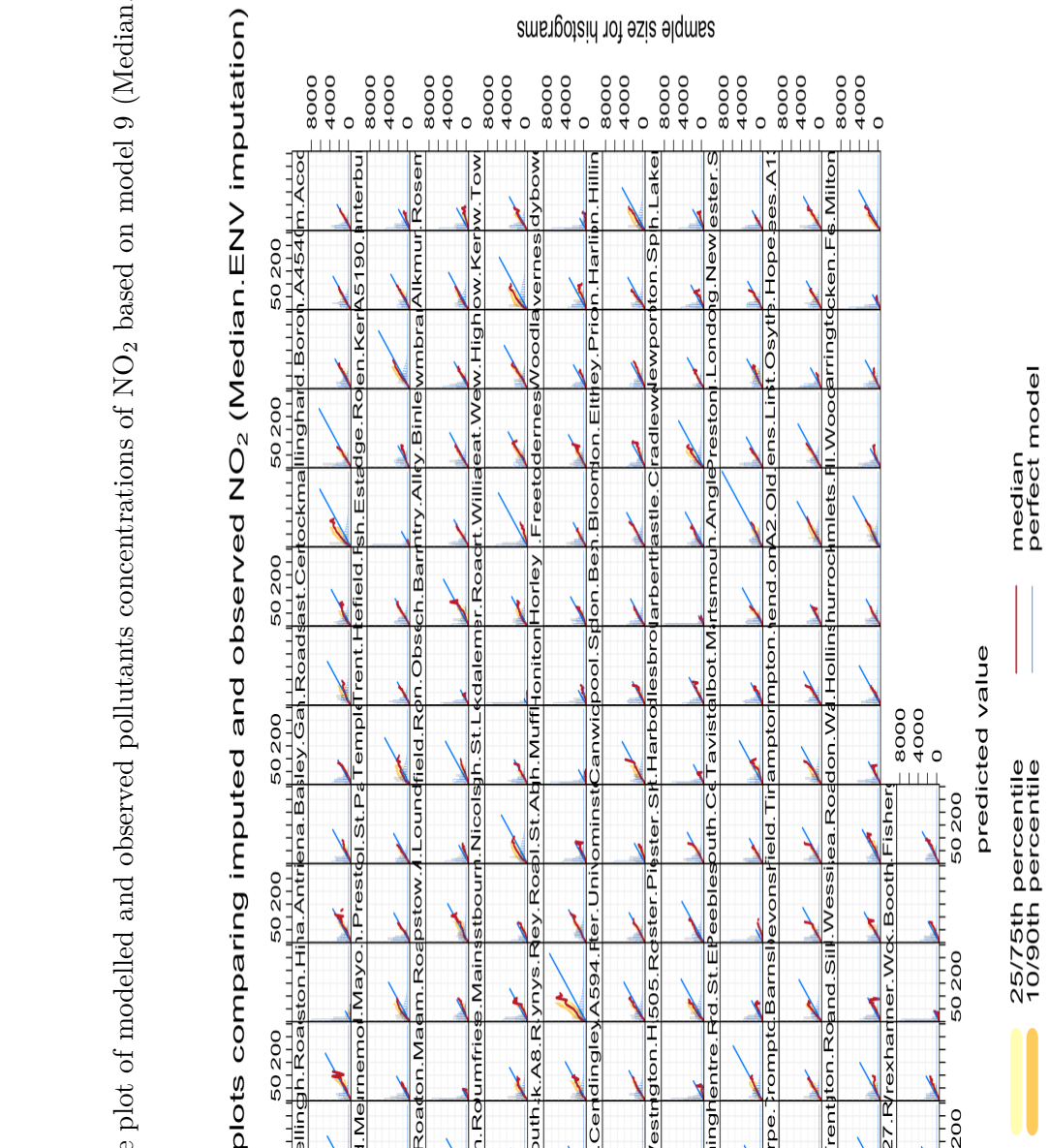


Figure F.2: Conditional quantile plot of modelled and observed pollutants concentrations of NO_2 based on model 9 (Median.ENV) for all stations (175 stations).

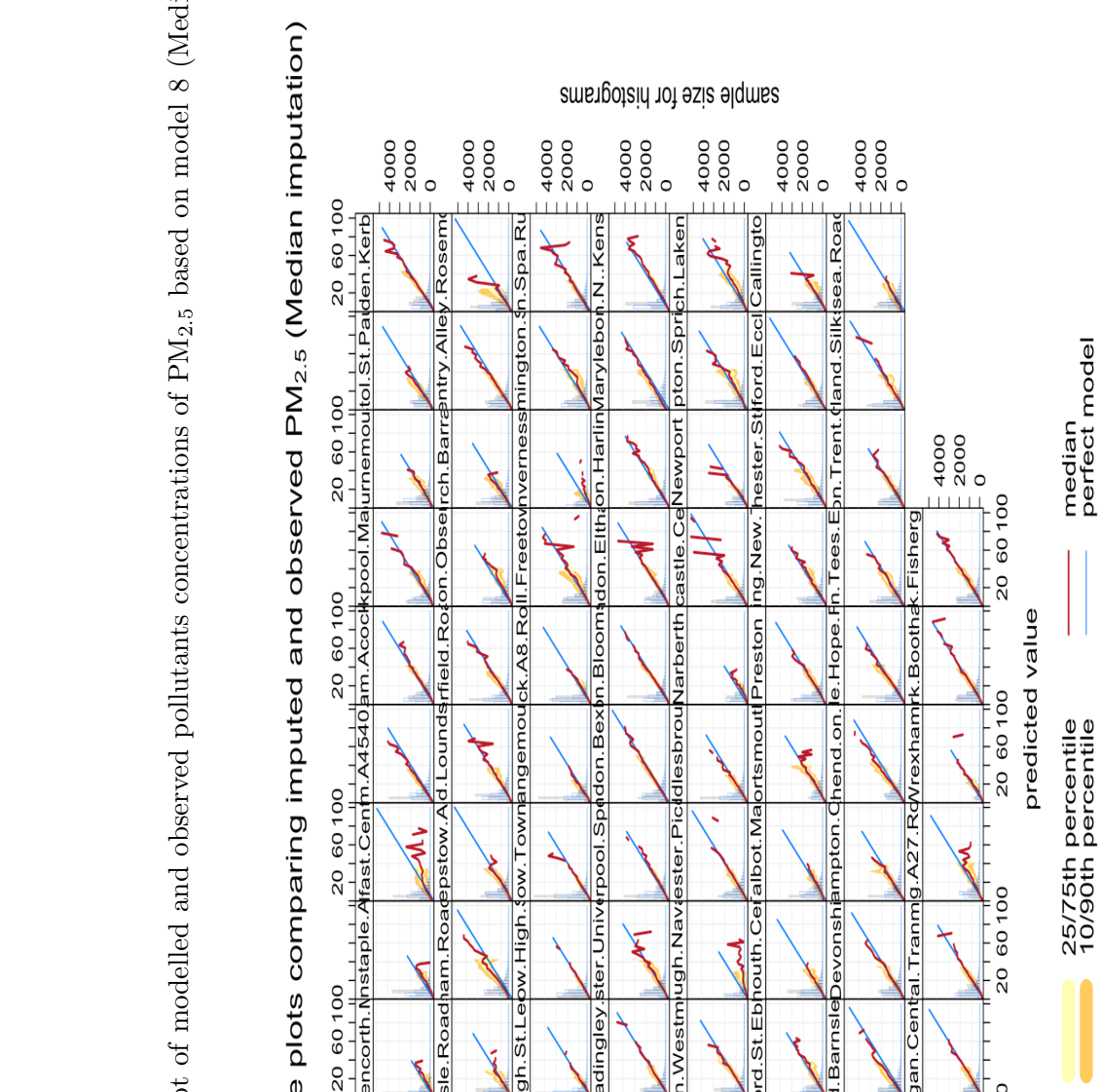


Figure F.3: Conditional quantile plot of modelled and observed pollutants concentrations of $PM_{2.5}$ based on model 8 (Median) for all stations (77 stations).

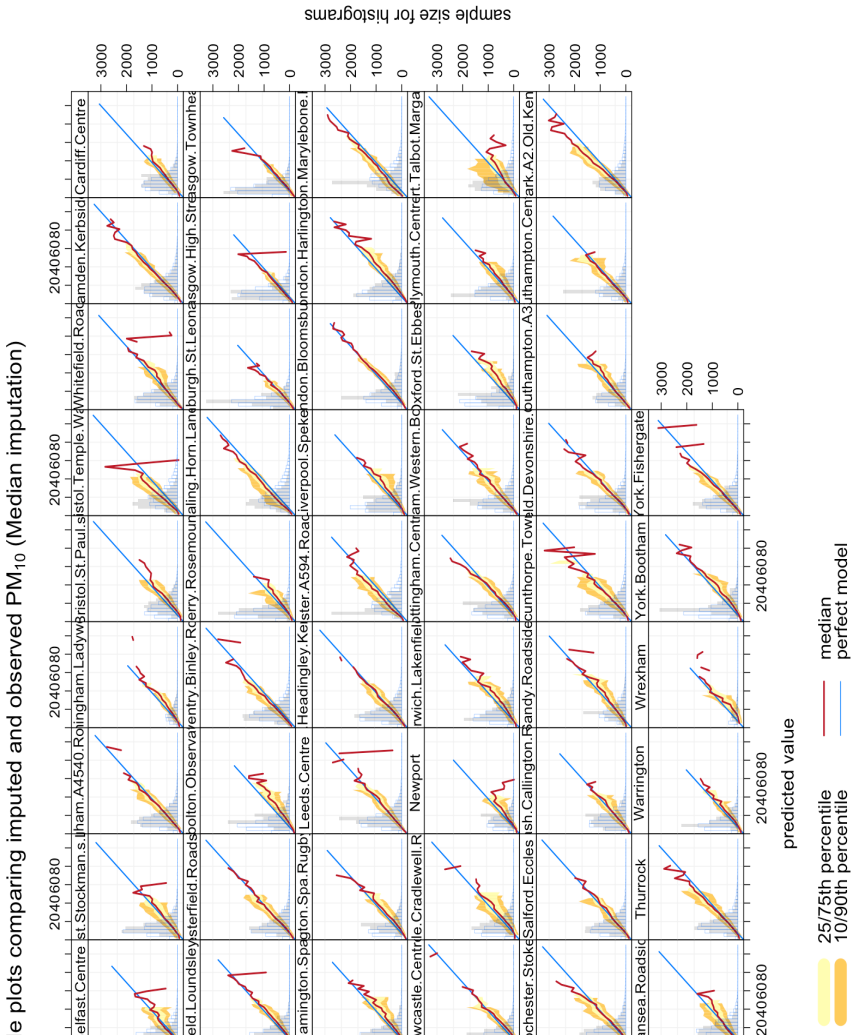


Figure F.4: Conditional quantile plot of modelled and observed pollutants concentrations of PM_{10} based on model 8 (Median) for all stations (75 stations).

Appendix G

Details on DAQI's Analysis.

In these examples, we show some cases where the imputation did not reproduce the DAQI very well and the is a high disagreement between the imputed and the observed DAQI for more than three index values.

Day 1:

- **Site**: Armagh Roadside
- **Date**: 26/06/2018
- **Imputed DAQI**: 2
- **Observed DAQI**: 7
- The observed DAQI's index is based on PM_{10}

Day 2:

- **Site**: Cardiff Centre
- **Date**: 20/08/2018
- **Imputed DAQI**: 2

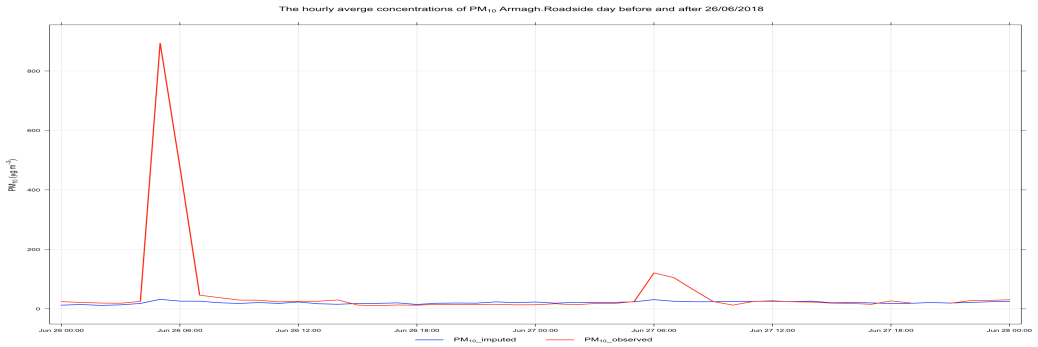


Figure G.1: Day 1: Hourly concentrations of PM_{10} at Armagh Roadside on 26-27/06/2018, showing the difference between imputed and observed concentrations. A sudden peaks values can be seen on observed PM_{10} at 26th Jun that caused higher observed DAQI. These values do not seem normal, and they could be influenced by the emission of PM_{10} from a large car that is running under the monitoring station as this station is a traffic roadside station.

- **Observed DAQI**: 6
- The observed DAQI's index is based on O_3

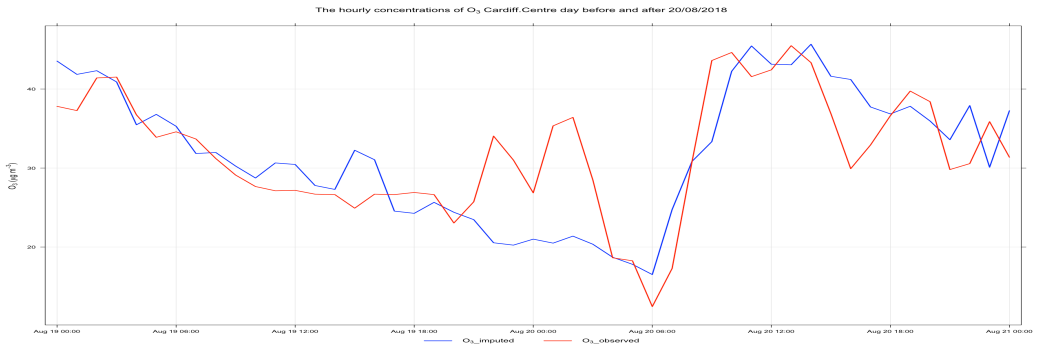


Figure G.2: Day 2: Hourly concentrations of O_3 at Cardiff Centre on 19-20/08/2018, showing the difference between imputed and observed concentrations. A sudden peaks at observed O_3 on 20th August that caused higher observed DAQI.

Day 3:

- **Site**: Chatham Roadside
- **Date**: 11/04/2018
- **Imputed DAQI**: 3

- **Observed DAQI**: 7
- The observed DAQI's index is based on $PM_{2.5}$

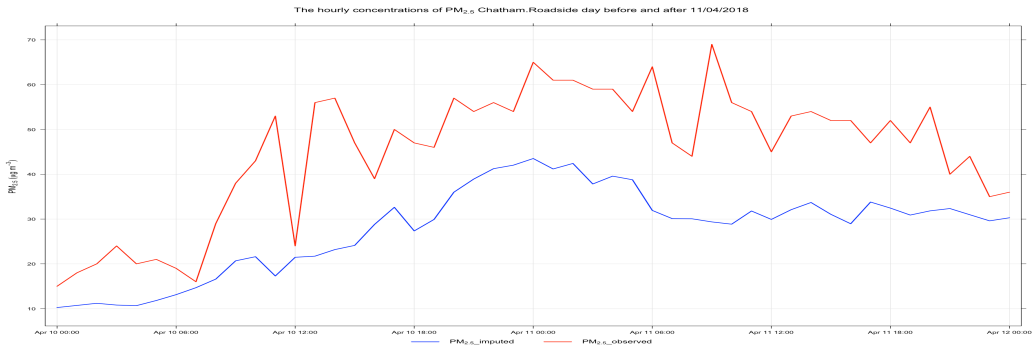


Figure G.3: Day 3: Hourly concentrations of $PM_{2.5}$ at chatham Roadside on 10-11/04/2018, showing the difference between imputed and observed concentrations. As this station is a roadside station.

Day 4:

- **Site**: Derry Rosemount
- **Date**: 07/01/2018
- **Imputed DAQI**: 2
- **Observed DAQI**: 6
- The observed DAQI's index is based on $PM_{2.5}$

Day 5:

- **Site**: Eastbourne
- **Date**: 21/04/2018
- **Imputed DAQI**: 3
- **Observed DAQI**: 8
- The observed DAQI's index is based on $PM_{2.5}$

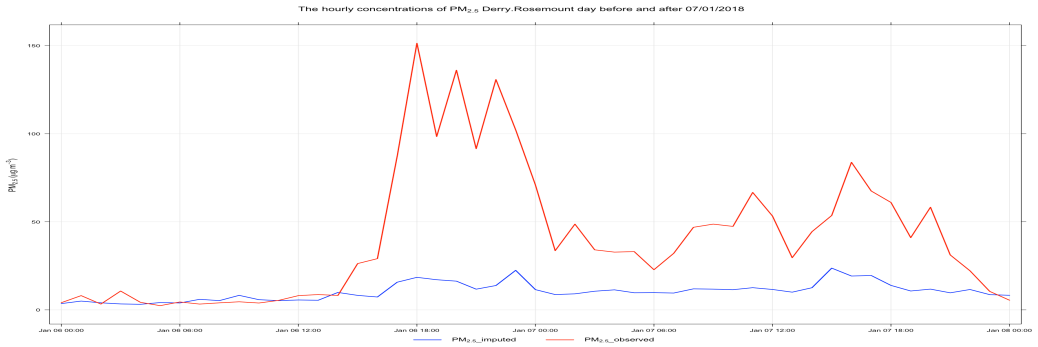


Figure G.4: Day 4: Hourly concentrations of $PM_{2.5}$ at Derry Rosemount on 06-07/01/2018, showing the difference between imputed and observed concentrations. Sudden hourly peaks at observed $PM_{2.5}$ at 7th Jan that caused higher observed DAQI.

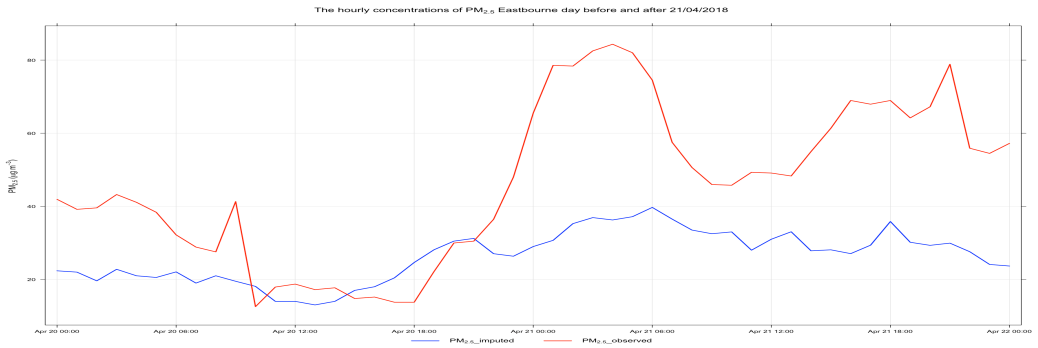


Figure G.5: Day 5: Hourly concentrations of $PM_{2.5}$ at Eastbourne on 21/04/2018, showing the difference between imputed and observed concentrations.

Day 6:

- **Site**: Ladybower
- **Date**: 07/05/2018
- **Imputed DAQI**: 5, DAQI's index is based on O_3
- **Observed DAQI**: 1, DAQI's index is based on NO_2

Day 7:

- **Site**: Salford Eccles
- **Date**: 27/06/2018

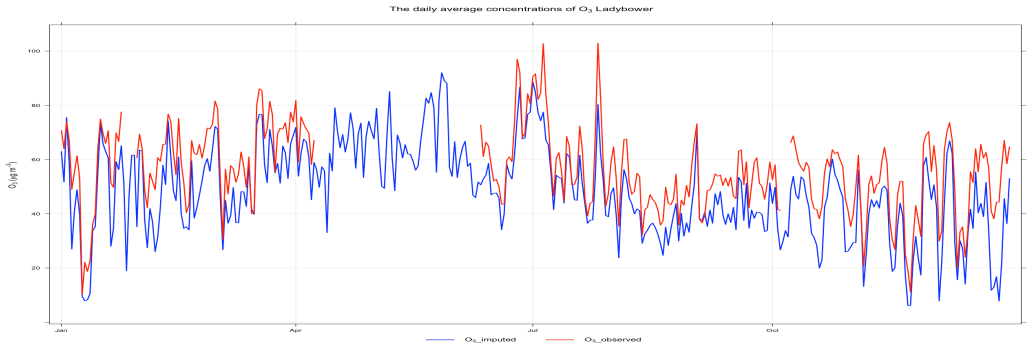


Figure G.6: Day 6: Daily mean concentrations of O_3 at Ladybower for year 2018, showing how the imputation reproduces O_3 observations. Note, that O_3 observations are missing at this day.

- **Imputed DAQI**: 2
- **Observed DAQI**: 8
- The observed DAQI's index is based on $PM_{2.5}$

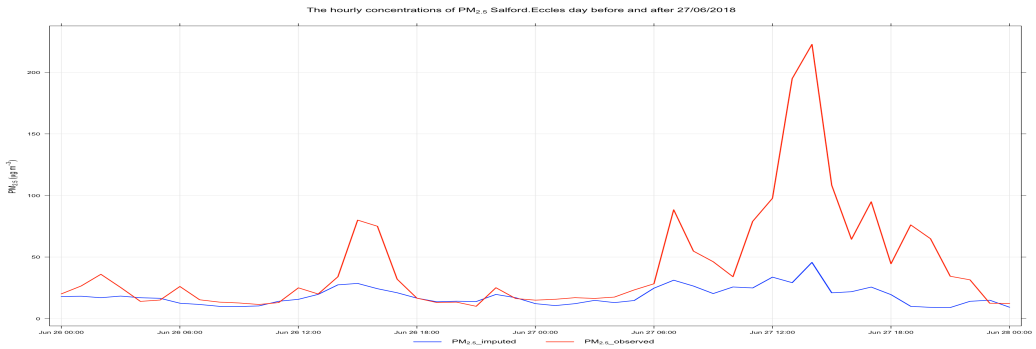


Figure G.7: Day 7: Hourly concentrations of $PM_{2.5}$ at Salford Eccles on 26-27/06/2018, showing the difference between imputed and observed concentrations. Some peaks on 27th Jun, that higher $PM_{2.5}$ level and increase the DAQI.

Day 8:

- **Site**: Sheffield. Barnsley Road
- **Date**:05/11/2018
- **Imputed DAQI**: 6
- **Observed DAQI**: 10

- The observed DAQI's index is based on $PM_{2.5}$

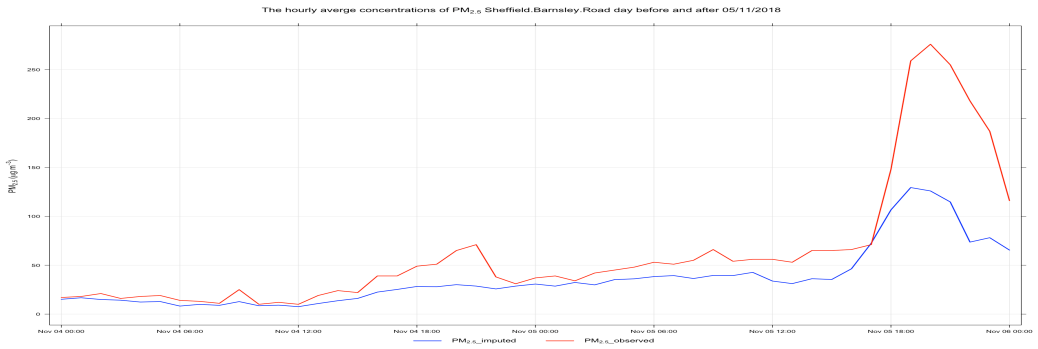


Figure G.8: Day 8: Hourly concentrations of $PM_{2.5}$ at Sheffield. Barnsley Road on 04-05/11/2018, showing the difference between imputed and observed concentrations.

Day 9:

- **Site**: London Westminster
- **Date**: 03/03/2018
- **Imputed DAQI**: 8, DAQI's index is based on $PM_{2.5}$
- **Observed DAQI**: 2, DAQI's index is based on NO_2

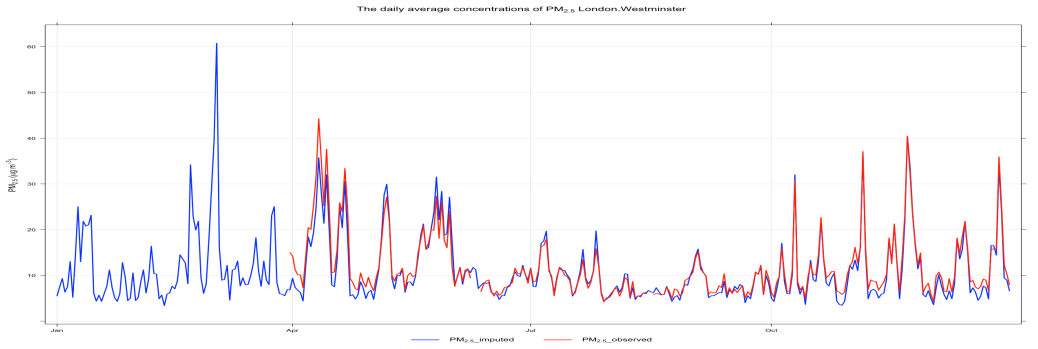


Figure G.9: Day 9: Daily mean concentrations of $PM_{2.5}$ at London Westminster for year 2018, showing how the imputation reproduces $PM_{2.5}$ observations. Note, that $PM_{2.5}$ observations are missing at this day.

Day 10:

- **Site**: Worthing A27 Roadside
- **Date**: 03/03/2018
- **Imputed DAQI**: 6, DAQI's index is based on $PM_{2.5}$
- **Observed DAQI**: 2, DAQI's index is based on NO_2

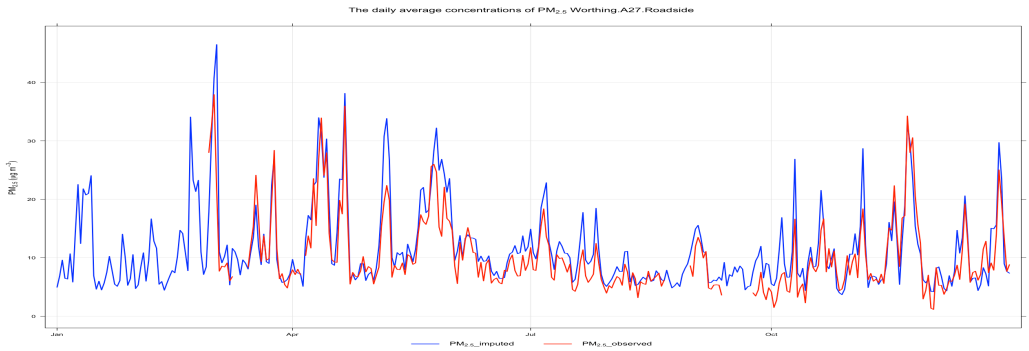


Figure G.10: Day 10: Daily mean concentrations of $PM_{2.5}$ at Worthing A27 Roadside for year 2018, showing how the imputation reproduces $PM_{2.5}$ observations. Note that $PM_{2.5}$ observations are missing at this day.

Day 11:

- **Site**: Wrexham
- **Date**: 03/03/2018
- **Imputed DAQI**: 6, DAQI's index is based on $PM_{2.5}$
- **Observed DAQI**: 1, DAQI's index is based on NO_2

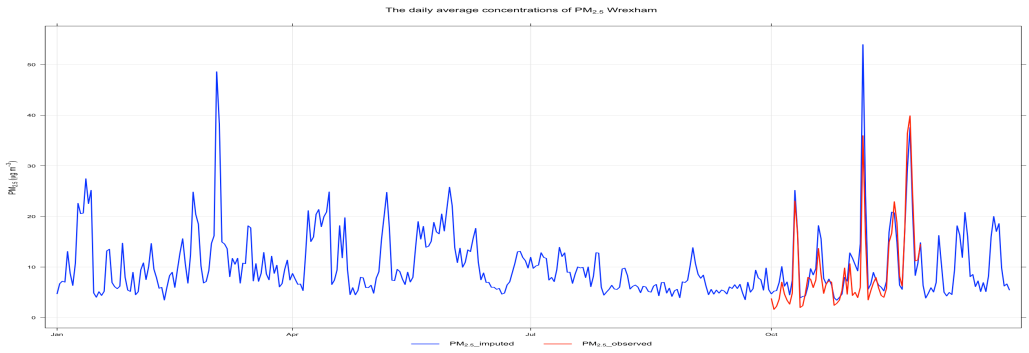


Figure G.11: Day 11: Daily mean concentrations of $PM_{2.5}$ at Wrexham for year 2018, showing how the imputation reproduces $PM_{2.5}$ observations. Note, that $PM_{2.5}$ observations are missing at this day.

Appendix H

Model Application Analysis.

Table H.1: Stations associated with events (102 events) of high variation between observed and imputed DAQI in model application.

Site Name	Environment Type	Zone	Region	NO_2	$PM_{2.5}$	PM_{10}	O_3
Barnsley.Gawber	Background Urban	Yorkshire & Humberside	Yorkshire & Humberside	1	0	0	1
Barnstaple.A39	Traffic Urban	South West	South West	0	1	1	0
Bath.Roadside	Traffic Urban	South West	South West	1	0	0	0
Billingham	Industrial Urban	Teesside Urban Area	North East	1	0	0	0
Birkenhead.Borough.Road	Traffic Urban	Birkenhead Urban Area	North West & Merseyside	1	0	0	0
Blackburn.Accrington.Road	Traffic Urban	North West & Merseyside	North West & Merseyside	1	0	0	0
Borehamwood.Meadow.Park	Background Urban	Eastern	Eastern	1	0	0	0
Bradford.Mayo.Avenue	Traffic Urban	West Yorkshire Urban Area	Yorkshire & Humberside	1	0	0	0
Burton.on.Trent.Horninglow	Background Urban	West Midlands	West Midlands	1	0	0	0
Cambridge.Roadside	Traffic Urban	Eastern	Eastern	1	0	0	0
Cannock.A5190.Roadside	Traffic Urban	West Midlands	West Midlands	1	0	0	0
Canterbury	Background Urban	South East	South East	1	0	0	1
Charlton.Mackrell	Background Rural	South West	South West	1	0	0	0
Derby.St..Alkmund.s.Way	Traffic Urban	East Midlands	East Midlands	1	0	0	0
Dewsbury.Ashworth.Grove	Background Urban	West Yorkshire Urban Area	Yorkshire & Humberside	1	0	0	0
Doncaster.A630.Cleveland.Street	Traffic Urban	Yorkshire & Humberside	Yorkshire & Humberside	1	0	0	0
Glazebury	Background Rural	North West & Merseyside	North West & Merseyside	1	0	0	1
Grangemouth	Industrial Urban	Central Scotland	Central Scotland	1	1	1	0
Grangemouth.Moray	Industrial Urban	Central Scotland	Central Scotland	1	0	0	0
Haringey.Roadside	Traffic Urban	Greater London Urban Area	Greater London Urban Area	1	0	0	0
Hartlepool.St.Abbs.Walk	Background Urban	North East	North East	1	0	0	0
High.Muffles	Background Rural	Yorkshire & Humberside	Yorkshire & Humberside	1	0	0	1
Honiton	Background Urban	South West	South West	1	0	0	0
Horley	Industrial Suburban	South East	South East	1	0	0	0
Immingham.Woodlands.Avenue	Background Urban	Yorkshire & Humberside	Yorkshire & Humberside	1	0	0	0
Ladybower	Background Rural	East Midlands	East Midlands	1	0	0	1
Leamington.Spa.Rugby.Road	Traffic Urban	West Midlands	West Midlands	1	1	1	0
Lincoln.Canwick.Road	Traffic Urban	East Midlands	East Midlands	1	0	0	0
London.Bexley	Background Suburban	Greater London Urban Area	Greater London Urban Area	1	1	0	0
London.Eltham	Background Suburban	Greater London Urban Area	Greater London Urban Area	1	1	0	1
London.Haringey.Priory.Park.South	Background Urban	Greater London Urban Area	Greater London Urban Area	1	0	0	1
London.Hillingdon	Background Urban	Greater London Urban Area	Greater London Urban Area	1	0	0	1
London.Westminster	Background Urban	Greater London Urban Area	Greater London Urban Area	1	1	0	0
Lullington.Heath	Background Rural	South East	South East	1	0	0	1
Luton.A505.Roadside	Traffic Urban	Eastern	Eastern	1	0	0	0
Market.Harborough	Background Rural	East Midlands	East Midlands	1	0	0	1
Newport	Background Urban	South Wales	South Wales	1	1	1	0
Oldbury.Birmingham.Road	Traffic Urban	West Midlands Urban Area	West Midlands	1	0	0	0
Oxford.Centre.Roadside	Traffic Urban	South East	South East	1	0	0	0
Oxford.St.Ebbes	Background Urban	South East	South East	1	1	1	0
Plymouth.Tavistock.Road	Traffic Urban	South West	South West	1	0	0	0
Shaw.Crompton.Way	Traffic Urban	Greater Manchester Urban Area	North West & Merseyside	1	0	0	0
Sheffield.Tinsley	Background Urban	Sheffield Urban Area	Yorkshire & Humberside	1	0	0	0
Sibton	Background Rural	Eastern	Eastern	0	0	0	1
Southampton.A33	Traffic Urban	Southampton Urban Area	South East	1	0	1	0
St.Osyth	Background Rural	Eastern	Eastern	1	0	0	1
Stockton.on.Tees.A1305.Roadside	Traffic Urban	Teesside Urban Area	North East	1	0	0	0
Storrington.Roadside	Traffic Urban	South East	South East	1	0	0	0
Sunderland.Wessington.Way	Traffic Urban	North East	North East	1	0	0	0
Swindon.Walcot	Background Urban	South West	South West	1	0	0	0
Telford.Hollinswood	Background Urban	West Midlands	West Midlands	1	0	0	0
Tower.Hamlets.Roadside	Traffic Urban	Greater London Urban Area	Greater London Urban Area	1	0	0	0
Walsall.Woodlands	Background Urban	West Midlands Urban Area	West Midlands	1	0	0	1
Warrington	Industrial Urban	North West & Merseyside	North West & Merseyside	1	1	1	0
Weybourne	Background Rural	Eastern	Eastern	0	0	0	1
Wicken.Fen	Background Rural	Eastern	Eastern	1	0	0	1
Wrexham	Traffic Urban	North Wales	North Wales	1	1	1	0

*** 0 pollutant is not measured at the station**
*** 1 pollutant is measured at the station**