# Reducing the Errors in High Resolution Environmental Modelling

GABOR MAKRAI

**Doctor of Philosophy**
**University of York**
**Computer Science**
February 2018

# Abstract

Air pollution modelling is one of the key tools for researchers, scientists, and urban planners to support the sustainable development of the urban environment. This modelling tool is critical for the users in the age of rapid urbanization to understand pollution distribution in the modelling area. Recent updates in air quality regulations are challenging the state-of-the-art air pollution modelling techniques by requiring accurate predictions on a high temporal level, i.e. predictions at the hourly level rather than the annual level. Current state-of-the-art models are designed to have good prediction accuracy on the low temporal resolution by assuming that the pollution is in steady state. Making predictions on higher temporal resolution violates this assumption and cause inaccurate predictions. There are existing statistical modelling approaches for air pollution modelling, however, these approaches also struggle to make accurate predictions on higher temporal resolution. This work is looking into the development of a statistical regression based air pollution model which produces accurate high temporal level predictions by utilizing advanced regression algorithm to exploit the hidden knowledge in data with high temporal resolution. The analysis of the predictions of multiple advanced statistical regression algorithms is investigated to determine the most accurate approach hence the Random Forest Regression method is proposed for the given regression task. A novel model ensemble method is then developed to utilize multiple Random Forest Regression models trained on the different subset of the available input data. Motivated by the high computational requirement of the developed methods, this thesis also investigates the scalability and the robustness of the developed methods. Based on the experience gained from this investigation, this work proposes further model ensemble methods to improve the accuracy of the statistical regression approach for air pollution modelling. The developed air pollution model presented in this thesis produces more accurate hourly concentration level predictions than the current state-of-the-art method, hence, the approach gives the opportunity for better understanding of the pollution in the urban area.

# Contents

# List of Tables

# List of Figures

# Acknowledgements

I would like to give my sincere gratitude to the people who supported me throughout the course of my PhD.

Firstly, I would like to thank the guidance of my supervisors, Iain Bate and Steve Cinderby. Thank you for your patience, guidance, support and exceptional mentorship. I am grateful to have had the opportunity to explore and experiment with new ideas and learn the critical thinking to be able to judge these ideas. I am also grateful for the great feedback you provided me on my development during the course of the PhD.

Secondly, I would like to thank the support I have received from Francesco Pilla as he provided my the opportunity for the 3-month long internship in Dublin. His help during this internship was crucial as it helped me to refocus my research and bring it to the right direction.

I would also like to thank the members of the CAPACITIE project (Alistair, Lorraine, Prado, Emily, Xiu, Elena, Fady, Magda, Mayank, Michelle, Xinwei, Rina, Kyle). Thank you all for the interesting discussions and brainstorming sessions we had together to help each other to proceed further on our PhD journeys.

Finally, I would like to thank the faithful and endless support I received from my better half, Viki. Thank you for being available when the most help was needed and thank you for your support in the darkest hours during the course of my PhD.

This work is part of the Cutting-edge Approaches for Pollution Assessment in Cities (CAPACITIE) project that has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 608014.

*Gabor Makrai*

*February 2018, York, United Kingdom*

# Declaration

This thesis has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree other than Doctor of Philosophy of the University of York. This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by explicit references.

I hereby give consent for my thesis, if accepted, to be made available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed .....(candidate)

Date .....

Some of the material contained in this thesis has appeared in the following published or awaiting publication papers:

1. Gabor Makrai and Iain Bate. Signal Selection in a Complex Environmental Distributed Sensing Problem. In *Proceedings of the 13th International Conference on Distributed Computing in Sensor Systems(DCOSS)*, 2017. IEEE Computer Society. This paper consists the brief summary of applying advanced Statistical Regression algorithms to hourly $NO_2$ concentration level predictions presented in Chapter 3 and the majority of the analytic work of these algorithms presented in Chapter 4.

# CHAPTER 1

# Introduction and Motivation

Air pollution modelling is a crucial tool available for scientist, researchers and urban planners in the age of rapid urbanization. The pollution models allow the users to predict the pollution changes caused by the changes in the urban environment (such as building new housing areas or changing the traffic flows in the urban area). The aim of the air pollution models is to accurately predict pollution concentration levels for the complete urban area (often as a pollution concentration level heatmap) which prediction is the base of further environmental analysis. These models help urban planners to investigate the effect of certain changes in the urban environment, therefore, these type of air pollution models must generate pollution concentration levels for the entire urban area considering the changes applied by the urban planners.

According to the European Environment Agency [Guerreiro et al. (2013)] and the World Health Organization [WHO (2009)], one of the most concerning pollutant in the urban area is the Nitrogen Dioxide ($NO_2$). Modelling the concentration levels of $NO_2$ accurately is one of the most challenging tasks compared to modelling other pollutants. These challenges include the fact that the $NO_2$ pollutant has multiple sources (e.g. traffic and industry) and the concentration level depends on many factors (e.g. wind speed, wind direction, traffic volume) and $NO_2$ pollutant is reacting with other gases in the air (e.g. ozone, nitrogen monoxide) it is forming from ozone in some circumstances) [Seinfeld & Pandis (2016)]. The thesis is focusing on this pollutant only because it is one of the most challenging tasks, however, the developed method in this thesis can be applied to any other pollutant. Applying the developed approach to modelling other pollutants is possible because the pollution concentration levels prediction task is based on the similar principle: pollution emission sources are emitting the pollution into the air and the pollution is

dispersed based on the given meteorological conditions.

Air pollution dispersion models are the state-of-the-art model for air pollutant concentration level predictions [Stocker et al. (2012)]. These methods calculate the concentration levels based on the emission rate of the pollutant and using a dispersion technique to distribute the pollution in the modelling area using the weather conditions (e.g. wind speed, wind direction). Air pollution dispersion models are capable of accurately predicting the annual concentration levels in the urban area [Vardoulakis et al. (2007)].

Recent updates in the air pollution regulations define limits on high-temporal (hourly) concentration levels along with the limits on low-temporal (annual) concentration levels [WHO (2000)]. These high-temporal concentration level limits are challenging the state-of-the-art air pollution models as the dispersion models were developed assuming that the pollution is in steady state which assumption is not necessarily valid on the high-temporal level [Berkowicz et al. (2008)]. Also, air pollution dispersion models depend on datasets such as emission inventory databases and traffic amount to calculate the emission levels which datasets often contain uncertain data [Owen et al. (2000)]. The uncertainty in the input data causes uncertainty in the generated predictions, therefore, the air pollution dispersion model struggles to make accurate concentration level predictions on the high-temporal levels [Vardoulakis et al. (2007)].

The Land Use Regression (LUR) method is a different approach for air pollution concentration level prediction in the urban area [Briggs et al. (2000)]. The main idea of the Land Use Regression model is to extract relevant information (for the pollutant) around the monitoring station and turn this problem into a statistical regression task by using the extracted data as input for the regression and use the observed concentration levels as the target of the regression. Land Use Regression models are capable of accurately predicting the annual concentration levels in the urban area without using uncertain data necessary for the air pollution dispersion models [Brauer et al. (2003)]. Studies indicate that using the Land Use Regression struggle to make accurate predictions on high-temporal level due to the nature of the input data and the applied Linear Regression statistical regression method [Hochadel et al. (2006)].

Recent advances in the machine learning field produced new algorithms for solving regression problems more accurately [Nasrabadi (2007)]. Studies indicating that algorithms such as Nearest Neighbour Regression [Nasrabadi (2007)], Neural Network Regression [Gardner & Dorling (1999)], Support Vector Regression [Sánchez et al. (2011)], Decision Tree Regression [Tso & Yau (2007)] and Random Forest Regression [Champendal et al. (2014)] methods can produce more accurate predictions on similar regression task than the standard Linear Regression method. These methods use hyperparameters to build their internal data structures for predictions, therefore, the accuracy of the prediction by these methods are sensitive to these hyperparameters.

The thesis aims to reduce the error of the hourly $NO_2$ concentration level prediction for the urban area by applying high-temporal input data and advanced machine learning regression

algorithms. The current state of the art methods in high temporal resolution air pollution concentration level predictions are using air pollution dispersion techniques which require reliable input data to make accurate predictions. It is difficult to collect accurate input data on high temporal resolution. The method presented in this thesis reduces the prediction error of the current state of the art air pollution models by utilizing advanced machine learning algorithms to efficiently exploit the hidden relationship between the input data and air pollution concentration levels. This method can utilize the input data more efficiently because it can ignore unreliable data which only introduces prediction error into the prediction.

## 1.1 Hypothesis and research objectives

This thesis aims to investigate the challenges of the development of a statistical regression approach for hourly $NO_2$ concentration level prediction for the urban area using advanced machine learning regression techniques and the hypothesis is defined as the following:

> *Through the appropriate ensembling of state of the art statistical regression methods, a more accurate, robust and scalable high-temporal environmental model can be created than the current state-of-the-art air pollution dispersion techniques*

### 1.1.1 Evaluation Criteria

The accuracy of the air pollution models is defined in multiple ways in the literature, however, the aim of the thesis is to increase the accuracy by every accuracy evaluation method presented in the literature for all the range of the observation spectrum. These accuracy evaluation methods include the mean absolute error, root mean squared error, normalised mean squared error, correlation coefficient, fractional bias, geometric mean bias, geometric variance, predictions are within a factor of two of observations.

### 1.1.2 Research objectives

To investigate the hypothesis, the thesis aims to carry out research investigating the following research objectives:

> Research Objective 1: *Establish an evaluation framework to investigate the feasibility of using a statistical regression approach for hourly $NO_2$ concentration level predictions*
>
> Research Objective 2: *Evaluate the accuracy of advanced statistical regression algorithms using the evaluation framework to compare predictions of the most accurate statistical regression and the state-of-the-art air pollution dispersion methods*
>
> Research Objective 3: *Evaluate the sensitivity to the input data of the statistical regression approach using the developed evaluation framework*

Research Objective 4: *Develop a model ensemble method to efficiently combine multiple Random Forest statistical regression models*

Research Objective 5: *Evaluate the scalability of the developed statistical regression methods (including the Random Forest Regression and the Random Forest ensemble methods)*

Research Objective 6: *Develop an efficient ensemble of the statistical regression approach for large-scale dataset*

### 1.1.3 Contribution

The work in this thesis contributes to the Environmental Science and Computer Science fields. The novel air pollution statistical regression model developed in this thesis contributes to the Environmental Science field as it provides an accurate model for hourly $NO_2$ concentration level predictions. The novel ensemble regression method contributes to the Computer Science field as it is general regression technique which can be used to solve any regression task.

## 1.2 Thesis structure

The rest of the thesis is organized as follows. Chapter 2 explains a comprehensive literature review on the field of air pollution modelling including the challenges of modelling $NO_2$ concentration levels and the state-of-the-art methods then the chapter introduces the recent advances in the machine learning field including the developed regression algorithms for solving regression problems more accurately.

Next, Chapter 3 presents the work for the Research Objective 1. It describes the development of the evaluation framework where one of the state-of-the-art air pollution dispersion model, then the existing Land Use Regression method are evaluated. The chapter then presents the work for the Research Objective 2 which includes the sensitivity analysis of the advanced statistical regression algorithms to the given regression task using the developed evaluation framework. The chapter summarizes the result and presents the most accurate statistical regression algorithm for the hourly $NO_2$ concentration level prediction.

Chapter 4 presents the work for the Research Objective 3. It introduces the accuracy sensitivity study of the applied data which provides an insight into the statistical regression prediction generation process and it helps to understand what data is important for the model to make accurate predictions. This analysis leads to the work for the Research Objective 4 as the analysis reveals how the different data sources providing

Chapter 5 introduces the work for the Research Objective 5. It presents the scalability and robustness analysis of the developed statistical regression methods by applying them on a large-scale high-temporal environmental modelling scenario. The finding of this analysis leads to the work of the Research Objective 6.

The thesis concludes in Chapter 6 with the summary of all contributions that the thesis presents and the list of limitations of the developed statistical regression models. Finally, the future work is presented in the last section of this chapter.

# CHAPTER 2

# Literature Review

The aim of this chapter to introduce the existing literature related to the work that will be presented in this thesis. In the first section (Section 2.1), a general introduction to the air pollution is presented highlighting the relevant knowledge to understand the air pollution and the problems introduced by the air pollution. The second section (Section 2.2) focuses on the air pollution modelling and the existing methods for predicting concentration levels for air pollution modelling. It also highlights the recent challenges in the field of air pollution modelling. The following section (Section 2.3) discusses the statistical regression algorithms in the machine learning field which can be utilized for a novel statistical regression approach. Finally, the Section 2.4 finalizes the chapter.

## 2.1 Air pollution in the urban area

The World Health Organisation (WHO) reported that more than 50% of the human population lives in cities from 2010 and the urbanization process is increasing. This urbanization process leads to the large development of cities and managing this development is getting more important than ever was before. The increased amount of population living in the urban area cause larger traffic inside the city, but the urbanization process comes with the increased amount of constructions and renovations to improve cities capacity for handling the increased amount of population. Increased traffic is generating more pollution and also the heavy urbanization process requires new factories which will also generate more pollution [WHO (2009)].

A very good illustration of the pollution issues in the urban area is the pollution emission levels in the United Kingdom. According to a report by the Department for Environment, Food

and Rural Affairs (Defra), between 1980 and 2007 car traffic in the United Kingdom increased from 215 billion to 404 billion vehicle kilometres and the number of cars per UK household from 0.76 to 1.11 [Faulkner & Russell (2010)]. In the past 10 years, the statistics show a very high, but constant level of traffic volumes on the roads, however, high traffic volumes simple mean high pollution emission levels on the roads. Fortunately, some technological inventions (e.g. the catalytic converter) and regulations to develop engines with less emission (e.g. EURO vehicle emission standards) can help to reduce the emission levels of one vehicle, however, the observed average pollution levels are still increasing in the urban area [Pilling et al. (2007)].

There are regulations to keep to pollution levels to a certain amount to avoid the health consequences of the exposure of the high pollution levels. These regulations are controlled by the environmental protection agencies around the world (e.g. the European Environmental Agency (EEA) is defining the accepted pollution levels for the countries in the European Union).

According to the European Environment Agency [Guerreiro et al. (2013)] and the World Health Organization [WHO (2009)], one of the most concerning pollutant in the urban area is the Nitrogen Dioxide ($NO_2$).

Nitrogen dioxide is a reactive gas generated mostly by high-temperature combustion processes (e.g. burning fuel in car engines and in power plants). Usually just a small fraction of the nitrogen oxides emission is $NO_2$, however, studies show that the usage of exhaust after treatment systems and the increased penetration of diesel vehicles increasing this fraction from 5-10 percent to 70 percent [Nova et al. (2007)]. This leads to serious problems in traffic hotspots due to the fact that public transport is using mostly diesel vehicles. $NO_2$ primarily affects the respiration system. Short-term exposure can result in changed lung function, long-term exposure can result in symptoms of bronchitis in asthmatic children, however, $NO_2$ is highly correlated with other pollutants, therefore, it is difficult to differentiate the single effect of the $NO_2$ [WHO (2003)].

There are regulations in place for the nitrogen dioxide concentration levels. In Europe, the annual average concentration level must be below 40 $\mu gm^{-3}$ and the 200 $\mu gm^{-3}$ hourly concentration level must not be exceeded 18 times a year. It is the only pollutant which has regulation to control the hourly concentration level as even short-term exposure to high concentration levels (200 $\mu gm^{-3}$) can result in adverse health effects [Guerreiro et al. (2013)].

The traffic and the industy are the two of the main sources of the $NO_2$ pollutant, however, the nitrogen dioxide also has a complex chemical lifecycle. Figure 2.1 shows the simplified version of the chemical lifecycle of the $NO_2$ gas. There are three major chemical processes that control the concentration of the $NO_2$ in the atmosphere:

$$NO_2 + hv \rightarrow NO + O^{\bullet} \tag{2.1}$$

and

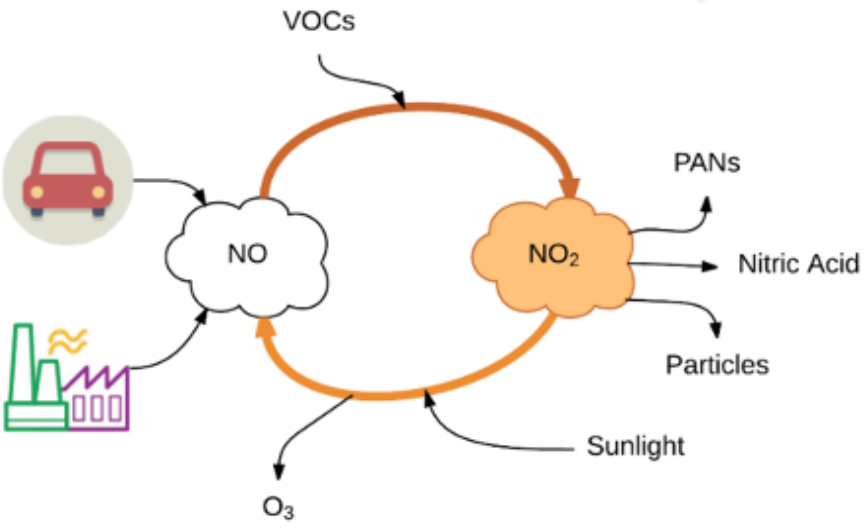



Figure 2.1: $NO_2$ chemical life cycle [Corbitt (1990)]

$$O^{\bullet} + O_2 + VOC \rightarrow O_3 + VOC \tag{2.2}$$

and

$$NO + O_3 \rightarrow NO_2 + O_2, \tag{2.3}$$

where $h\nu$, $O^{\bullet}$, $NO$, $O_2$ and $VOC$ represent sunlight, ionized oxygen, nitrogen monoxide, molecular oxygen and volatile organic compounds, respectively. Equation 2.1 represents the process when sunlight interacts with the $NO_2$ molecule and decomposes it to $NO$ and ionized oxygen. Equation 2.2 describes the process when the ionized oxygen and oxygen molecule forms an ozone molecule using VOC as the catalyst. Equation 2.3 presents the process when the nitrogen monoxide reacts with ozone and generates nitrogen dioxide and oxygen gas. Figure 2.1 also introduces peroxyacyl nitrates (PANs), nitric acid and other particles as the result of the nitrogen dioxide transformation process due to nature of the actual volatile organic compounds acting as catalyst in the process described in Equation 2.2. The complex chemical lifecycle of the $NO_2$ pollutant makes the prediction of the concentration level of the $NO_2$ challenging because it is not just emitted from the source and dispersed by the wind, but it is reacting with other gases in the air.

## 2.2 Air pollution modelling

Computational models of air pollution have been in existence for over 80 years [Daly & Zannetti (2007); Jerrett et al. (2004)]. Air pollution was modelled in different scales, with different approaches and using different data sources depending on the geographical and meteorological properties of the modelled area. Also, the development of the computational power of modern computers opened the way before new computational intense methods and the application of large-scale wireless sensor networks created the possibility of new data collection techniques [Kumar et al. (2015)]. To understand the principles of the air pollution modelling, first, the dimension scale of the method needs to be defined. Table 2.1 shows the five dimension categories defined by [Srivastava & Rao (2011)]. The average dimension of the urban scale is 100x100x5km with the resolution of 2 kilometres, but this depends on the population density of the given urban area. The work in this thesis will only consider the urban scale because the recent updates in the regulations are challenging the models on this spatial level.

| Model | Typical Domain Scale | Typical resolution |
|---|---|---|
| Micro scale | 200x200x100m | 5m |
| Urban scale | 100x100x5km | 2km |
| Regional scale | 1000x1000x10km | 36km |
| Continental scale | 3000x3000x20km | 80km |
| Global scale | 6500x6500x20km | 200km |

Table 2.1: Dimension scale of air pollution modelling [Srivastava & Rao (2011)]

A large number of different air pollution models were developed in the last couple of decades for many reasons: different geological locations and different climate conditions require different approaches, as well as the technology, allows the researchers to be able to run new, models with higher computational requirements and analyse the output of the models in more efficient ways. In terms of the urban scale air pollution modelling, the state-of-the-art methods follow the same principle [Srivastava & Rao (2011)]:

- the models require knowledge about the pollution emission levels and characteristics (e.g. point or line pollution source)
- the models require information about the weather around the modelling area
- the models use a mathematical model to estimate the concentration levels for the modelling area based on the emission levels of the pollution sources and the observed weather state
- these models are called air pollution dispersion models, because, the models calculate the concentration level by dispersing the pollution using these mathematical calculations

The existing air pollution dispersion models differ from the underlying mathematical model to calculate the pollution dispersion, but they all require the same input data groups: emission levels of the pollution sources and weather information for the dispersion [Srivastava & Rao (2011)].

Emission inventory databases are available for the scientists to provide the emission information for the air pollution dispersion models [Gurjar et al. (2008)].

Weather information includes data about the wind speed and direction, temperature and humidity and turbulent fluxes. Table 2.2 contains these factors with the ranking of the importance considering air quality, urban climatology and urban planning.

| Parameter | Air Quality | Urban Climatology | Urban Planning |
|---|---|---|---|
| Wind speed | Very important | Important | Very Important |
| Wind direction | Very important | Important | Very Important |
| Temperature, humidity | Important | Extremely Important | Very Important |
| Turbulent fluxes | Very important | Very important | Very important |

Table 2.2: Importance of weather parameter when modelling air pollution [Srivastava & Rao (2011)]

### 2.2.1 The state-of-the-art modelling approach

Air pollution dispersion has been studied for decades. One of the most studied technique called the Gaussian dispersion model [Hosker Jr (1975)]. It was one of the first models developed to model pollutant dispersion and the popularity of this model is still significant thanks to the simplicity of the underlying three-dimensional Gaussian distribution calculation. The model assumes that the pollution distribution is following a three-dimensional Gaussian distribution. This calculation does not require very complex equation systems or partial differential equations which means the model can generate output without heavy, computational intense calculations. Also, the model can handle a large number of pollution sources (as the sources are independent and can be calculated concurrently) and the model is able to pinpoint these sources [Hosker Jr (1975)]. Early implementation was only able to model static pollution conditions (for example average means), but later on, researchers have implemented time-dependent Gaussian dispersion models [Scire et al. (2000)].

Figure 2.2 shows the calculated air pollutant concentration distribution directed by the wind. The equation to calculate spatial concentration levels includes wind speed, wind direction, emission rate and effective stack height (the height of the actual source).

The underlying three dimensional Gaussian distribution equation that drives the model is defined as:

$$P(x,y,z) = \frac{Q}{2\pi\sigma_z(x)\sigma_y(x)} e^{-\frac{y^2}{2\sigma_y^2}} \left\{ e^{-\frac{(z-H_0)^2}{2\sigma_z^2}} + e^{-\frac{(z+H_0)^2}{2\sigma_z^2}} \right\}, \tag{2.4}$$

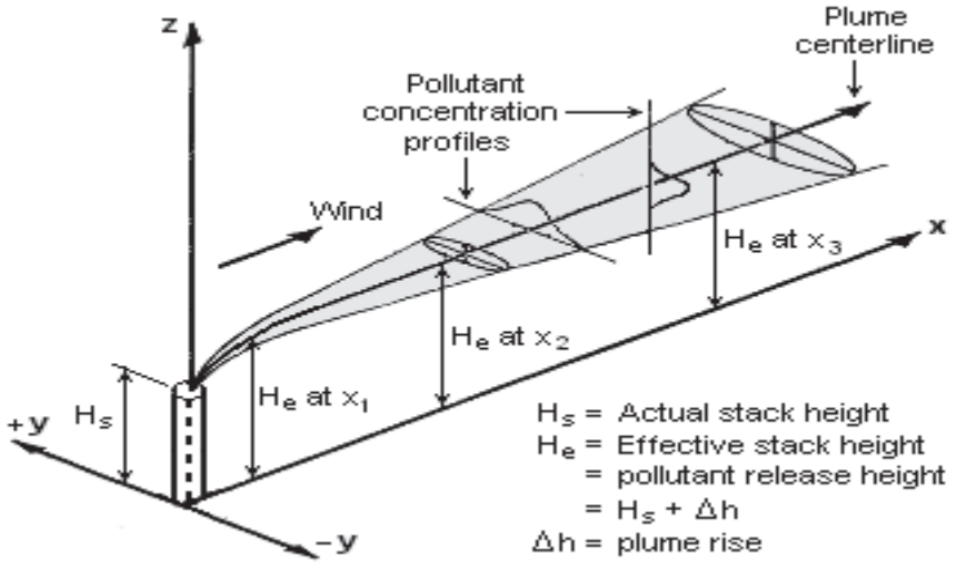



Figure 2.2: Gaussian air dispersion plume [Hosker Jr (1975)]

where $P(x, y, z)$ is the rate of pollution generated by the source, $Q$ is the emission rate, $\sigma_z(x)$ and $\sigma_y(x)$ are the horizontal and vertical standard deviation of the plume, $H_0$ is the height of the emission source and $x, y, z$ are the distances from the source along the three axis.

In practise, modellers are interested to calculate ground level pollution concentration levels which can be expressed by the simplifcation of Equation 2.4:

$$P(x, y, 0) = \frac{Q}{\pi\sigma_z(x)\sigma_y(x)} e^{-\frac{y^2}{2\sigma_y^2}} e^{-\frac{H_0^2}{2\sigma_z^2}}. \tag{2.5}$$

Both equations (Equation 2.4 and Equation 2.5) depends on vertical ($\sigma_z$) and horizontal ($\sigma_y$) standard deviation. According the empirical experiments by Pasquill [Pasquill (1961)], they can be calculated with the following equations [Martin (1976)]:

$$\sigma_y(x) = c_0 * x^{c_1}, \sigma_z(x) = c_2 * x^{c_3} + c_4, \tag{2.6}$$

where $c_0, c_1, c_2, c_3, c_4$ are constants and they are depending on weather stability classes described by Table 2.3.

To understand how the different stability classes affect the modelled concentration level spatially, the visualization of the concentration levels by the Gaussian air pollution dispersion model was generated. Figure 2.3 shows the visualization of a single point pollution source in different weather stability classes.

In theory, the Gaussian dispersion technique is able to generate accurate hourly pollution

TABLE 2.3. PARAMETERS FOR HORIZONTAL AND VERTICAL STANDARD DEVIATION CALCULATION [HOSKER JR (1975)]

| Stability class | | | x <1.0km | | | x >1.0km | | |
|---|---|---|---|---|---|---|---|---|
| | $c_0$ | $c_1$ | $c_2$ | $c_3$ | $c_4$ | $c_2$ | $c_3$ | $c_4$ |
| A: very unstable | 213 | 0.894 | 440.8 | 1.041 | 9.27 | 459.7 | 2.094 | -9.6 |
| B: unstable | 156 | 0.894 | 106.6 | 1.149 | 3.3 | 108.2 | 1.098 | 2.0 |
| C: slightly unstable | 104 | 0.894 | 61.0 | 0.911 | 0.0 | 61.0 | 0.911 | 0.0 |
| D: neutral | 68 | 0.894 | 33.2 | 0.725 | -1.7 | 44.5 | 0.516 | -13.0 |
| E: slightly stable | 50.5 | 0.894 | 22.8 | 0.675 | -1.3 | 55.4 | 0.305 | -34.0 |
| F: stable | 34 | 0.894 | 14.35 | 0.740 | -0.35 | 62.6 | 0.180 | -48.6 |

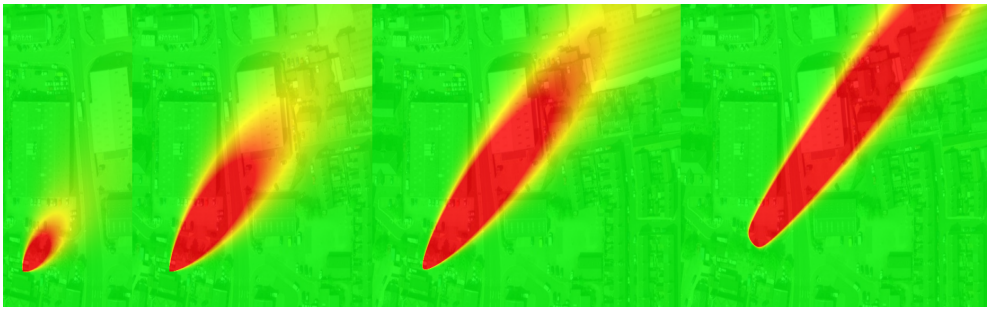


Figure 2.3. Point pollution source dispersion under different stability classes (A,B,C,D)

concentration level predictions, however, this requires close the perfect emission data as well as close to perfect weather condition data. It is practically impossible to collect close to perfect data for emission and weather, therefore, scientists use approximate data to feed the models [Hosker Jr (1975)].

Multiple implementations of the Gaussian dispersion model exist and studies were carried out to investigate to prediction accuracy and other properties of the implementations Carruthers et al. (1994); Scire et al. (2000). The Atmospheric Dispersion Modelling System (ADMS) method generates good prediction accuracy for multiple pollutants in the study of [Hanna et al. (2001)] which study includes the evaluation using multiple datasets. The CALPUFF method [Levy et al. (2002)] indicates good annual prediction accuracy for multiple pollutants (including sulphur dioxide, nitrogen oxides, and fine particles) produced by nine power plants. Carruthers conducted a study about the validation of the ADMS model in London in 2003 [Carruthers et al. (2003)]. According to the study, the model could reach very good annual concentration level accuracy for $NO_x$ which shows that the method is feasible for concentration level prediction as it has very low computational requirements. Kalhor at el. compared the predictions of AERMOD, ADMS and ISC3 models for annual $PM_{10}$ concentration levels in Mobarakeh steel complex, Iran. They reported good accuracy on the annual average concentration level, but the models are not sufficient to produce accurate concentration level predictions on higher temporal resolution

(all three models are overpredicting the measured maximum hourly concentration levels) [Kalhor & Bajoghli (2017)].

### 2.2.2 Other air pollution dispersion methods

There are other air pollution dispersion models in the literature which only differ in the underlying mathematical calculations. Gokhale and Khare defined 4 groups of air pollution models [Gokhale & Khare (2004)]:

- Deterministic models: these models are based on mathematical description of the atmospheric processes.
- Numerical models: these models are solving complex mathematical equation systems to generate concentration level predictions.
- Statistical models: these models are based on semi-empirical statistical relations between the available data (e.g. meteorological data and pollution concentration levels)
- Statistical distribution models: these models are mathematical models based on probability distribution functions.

The introduced state-of-the-art air pollution models belong to the deterministic category as they are drived by the Gaussian mathematical process. The rest of this section is dedicated to the introduction of the other categories.

#### 2.2.2.1 Numerical air pollution models

The Eulerian and Lagrangian dispersion models and the computational fluid dynamic models are the most often applied numerical air pollution models.

Eulerian and Lagrangian dispersion models are also well-established air pollution dispersion models as the first implementation originated in 1980's. The modelled area is divided into "small squares" (two-dimensional) or "small volumes" (three-dimensional) like grid cells. It is common to use equivalent sized cells during the modelling. Using these grid cells, it is possible to create a large set of mathematical expressions based on the position of each individual cell. These expressions include chemical transformations as well as the movement of different pollutant over the modelled area. Simulation is based on Eulerian method, where the model is assuming that pollution in one parcel is moving parallel to the wind direction with the velocity of speed. Simulation can be executed via forward and backward calculations in time. The main difference between the Eulerian and the Lagrangian models is that the Lagrangian model uses the Lagrangian method to calculate the transition between the cells which method supports the variable size of the cells (not just in terms of size but the shape by transforming the given coordinate space). Figure 2.4 shows an example visualization of the output of the applied Eulerian dispersion model [Reynolds et al. (1973)].

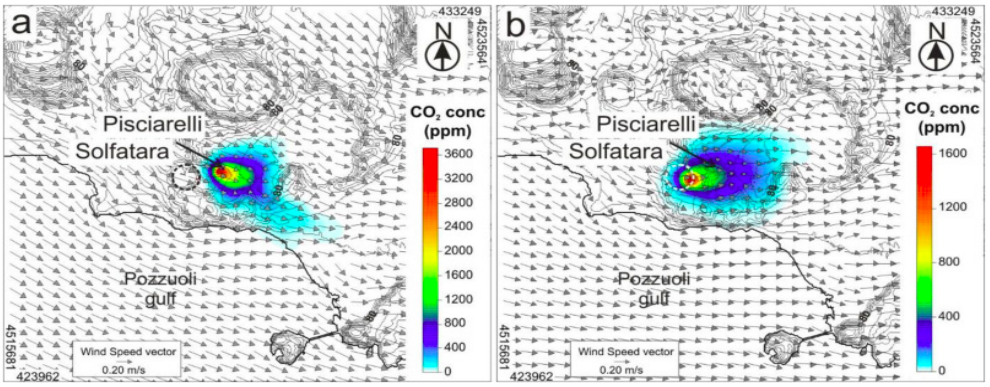



Figure 2.4. Visualization of the Eulerian dispersion model [Pedone et al. (2017)]

Multiple implementations of the models exist and there are studies evaluating the accuracy and other properties of these implementations [Yamartino et al. (1992); Christensen (1997)]. The CALGRID method [Yamartino et al. (1992)] provides a model with good prediction accuracy to predict daily ozone concentration level [O'Neill & Lamb (2005)] and the Danish Eulerian Hemispheric Model (DEHM) model [Christensen (1997)] indicates good prediction accuracy for sulphur and nitrogen compound concentration level predictions [Hole et al. (2009)]. Oettl conducted a study about Eulerian and Lagrangian dispersion models in 1995 [Oettl et al. (2001)]. This study utilizes the results of a previous measurement campaign near a major road at Elimaki in southern Finland, a campaign specifically designed for model evaluation purposes. He concluded that the models are predicting pollution levels with a small amount of error, but the calculation itself requires a huge amount of computational time.

Modelling the air pollution dispersion using computational fluid dynamic (CFD) models have been widely studied due to the fact that modelling the movement of particles in the air can be similar to the movement of particles in the fluids. It is possible to consider the air pollution modelling problem as a huge system where the air is flowing in the same sense of the fluids are flowing in those models except the air has slightly different physical properties [Craig et al. (1999)].

With the development of computational performance, researchers were able to produce very computational intense fluid dynamic models which turned out to be useful for modelling not just fluid dynamics, but air pollution as well. Two mayor representatives of CFD models are FLUENT [Riddle et al. (2004)] and RANS [Galmarini et al. (2009)]. While the first model is solving the three dimensional Raynolds averaged equitation, RANS is solving the Reynold Averaged Navier-Stokes equation. Both of them has an extremely high computational requirement and also it is hard to validate the output of the models.

Researchers conducted a deep analysis of the CFD method where they analysed the $NO_x$ concentration level in Stockholm using the CFD method. The model could achieve high accur-

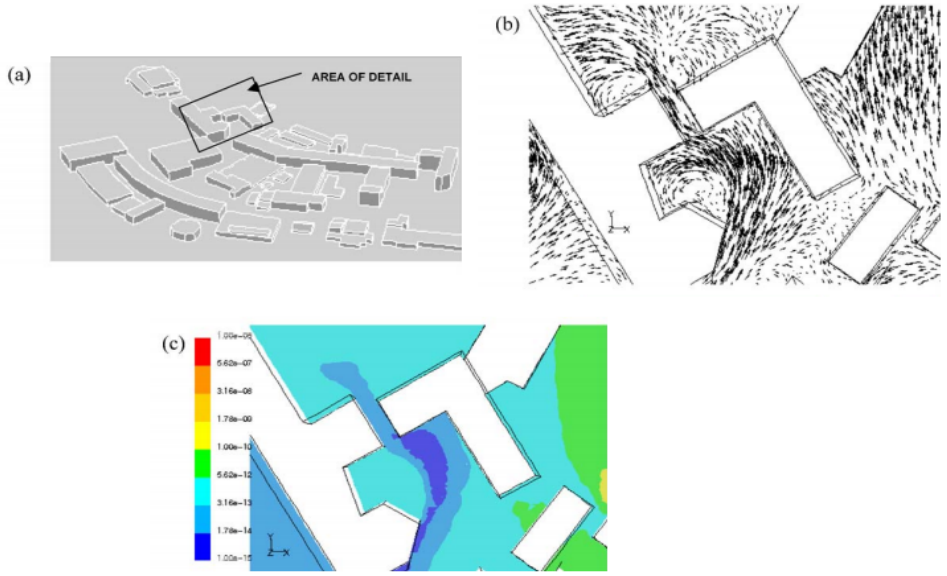



Figure 2.5. The FLUENT model showing a portion of the site layout (a) including the area where details of the predicted wind field (b) and predicted gas concentration for 1.5 m above the ground (c) are shown [Riddle et al. (2004)]

acy on the hourly level, however, calculation of the prediction requires an enormous amount of computational power [Gidhagen et al. (2004)].

#### 2.2.2.2 Statistical air pollution models

Statistical air pollution models based on computation models which exploits the semi-empirical statistical relationship between the available data (e.g. meteorological data) and the air pollution concentration levels. The models utilize historical observations to build the internal representation of the extracted knowledge. This internal data then can be used to make predictions based on any input data given to the model.

The models in this group differ in the way they extract the knowledge from the historical observations: many different statistical methods have been developed in the past and these methods generates different

Mueller et al reported good prediction accuracy for average $NO_2$ concentration level predictions on two weeks average time-scale for the Zurich, Switzerland modelling area. They used the Generalized Additive Models (GAM) to build their statistical model. Their input dataset contains 26 monitoring stations' observation data and they generated 26 independent model for each monitoring station data and used the GAM to merge the models into a single prediction model [Mueller et al. (2015)].

Pohata and Lungu reported good prediction accuracy for $NO_2$ and other pollutant daily averages for the Ploiesti, Romania modelling area. They have used the autoregressive integrated moving average (ARIMA) method to process the concentration level time-series data and build

the model. This method analyses the past concentration levels and build a regression model based on the past time-series observations. This model then can be applied for predicting future concentration levels [Pohoata & Lungu (2017)].

#### 2.2.2.3 Statistical distribution air pollution models

The statistical distribution based air pollution models utilze mathematical distribution functions to predict the air pollution conecntration levels. The models exploit the fact that frequency distribution of the pollution concentration levels shows strong relation with the frequency distribution of windspeed. This allows to model in this category to fit a mathematical distribution function to the air pollution concentraiton levels and generate concentration level predictions with the calculated functions.

Lu and Fang proposed a method to fit three theoretical distributions (log-normal, Weibull and type V Pearson distributions) to estimate the $PM_{10}$ and $PM_{2.5}$ pollutant daily average concentration levels in the Sha-Lu, Taiwan modelling area. They reported good accuracy for their method [Lu & Fang (2002)].

Giavis et al. developed a method to calculate the $PM_{10}$ hourly concentration levels using lognormal, gamma and Weibull theoretical distributions. They concluded that the lognormal distribution is the most appropriate method for this prediction task and the Weibull distribution is inapporopriate for this task [Giavis et al. (2008)].

### 2.2.3 Urban scale air pollution dispersion models

The importance of urban scale air pollution modelling resulted in a new set air pollution dispersion models specifically developed for the urban environment. These models are using the fact that the majority of the air pollution is generated by the traffic in the urban area and the models are using traffic data (such as volume, flow speed on roads, compound of the fleet inside the city) and vehicular emission standards to determine the pollution concentration generated on the roads. Also, the traditional dispersion methods do not work effectively as the urban geometry has its own effect on the pollution concentrations [Vardoulakis et al. (2003)]:

- houses and buildings along the roads are creating special turbulences which can lead many different situations depends on the weather (and mostly on the wind speed and the wind direction).

The process which has an effect of the concentration levels in the urban street environment is called as the urban canyons process (or street canyon process) as in some circumstances the buildings are forming canyons along the roads. One of the most important property for canyons is the geometry of the buildings and the length (L) of the road segment where aspect ratio (AR) means the height of the canyon (H) divided by the width (W) of the canyon. It is possible to classify them based on these properties into the following categories [Hunter et al. (1992)]:

- Wide canyon: AR is less than 0.3

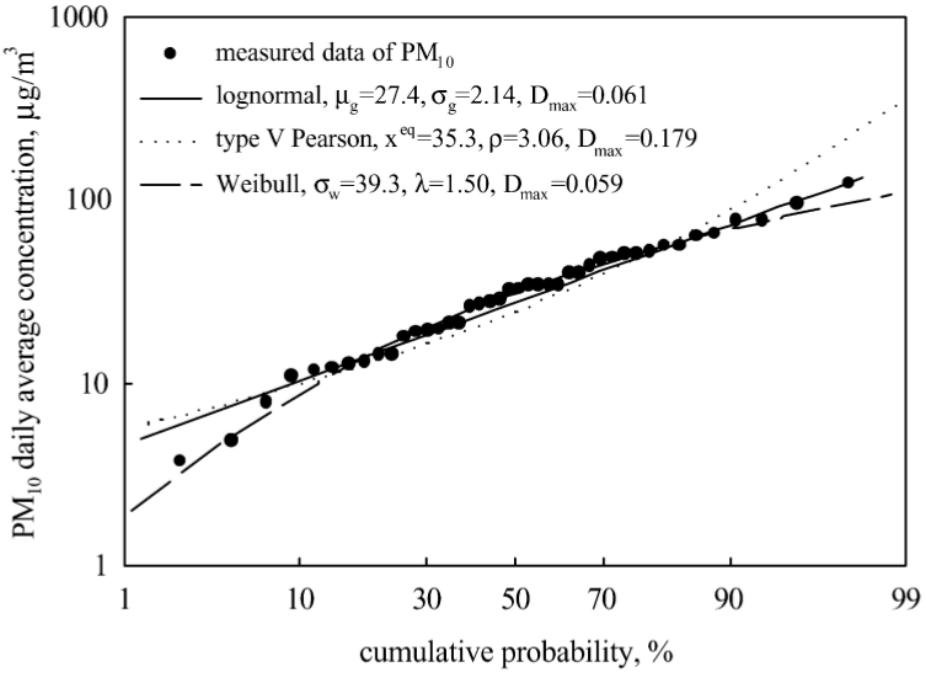



Figure 2.6. Daily average concentration level paired with the different distributions cumulative probability [Lu & Fang (2002)]

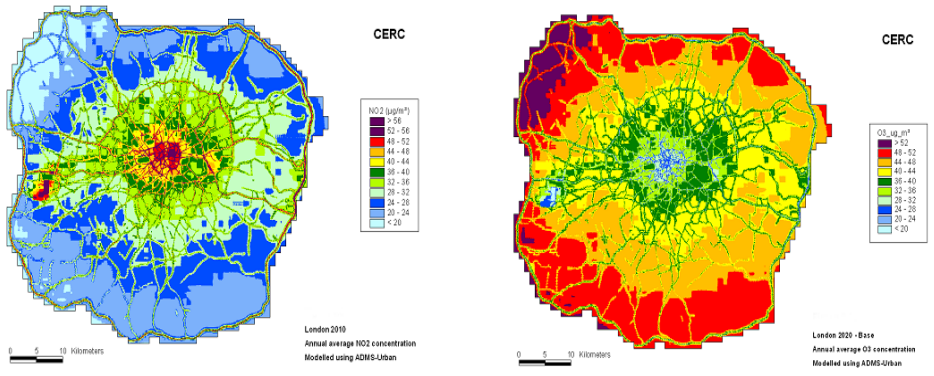



Figure 2.7. Contour plot of London showing the annual average $NO_2$ and $O_3$ concentrations predicted by ADMS-Urban for 2010 [McHugh et al. (1997)]

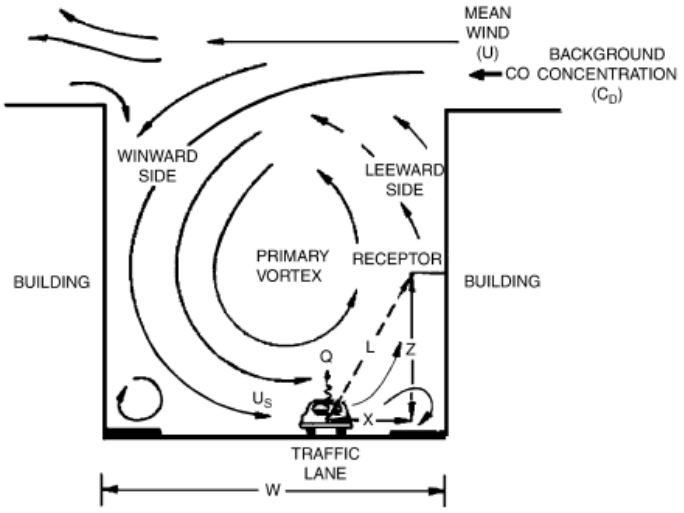



Figure 2.8. Pollutant dispersion in a regular street canyon [Dabberdt et al. (1973)]

- Regular canyon: AR is approximately 1.0
- Deep canyon: AR is greater than 2.0
- Short canyon: L/H is approximately 3.0
- Medium canyon: L/H is approximately 5.0
- Long canyon: L/H is greater than 7.0

In the terms of weather, the climate in the urban canyons is controlled by the wind, because the climate depends on the street geometry as the wind can cause alternated pollution dispersion. The wind can alternate the climate of the street if the perpendicular or near-perpendicular wind speed is larger than 1.5-2.0 m/s and the difference between the angle of the street and the wind direction is larger than 30 degree [Vardoulakis et al. (2003)].

Three main dispersion conditions were identified based on these factors [Hunter et al. (1992); Oke (1988)]:

- Isolated roughness flow: for wide canyons, the space between the buildings is enough for the wind to enter into this space and pick up and carry over the pollution from the ground level
- Wake interference flow: for those canyons which are between the wide canyons and regular canyons, clearing effect of the wind is breaking down because there is not enough space for the wind to enter and exit to and from the canyon

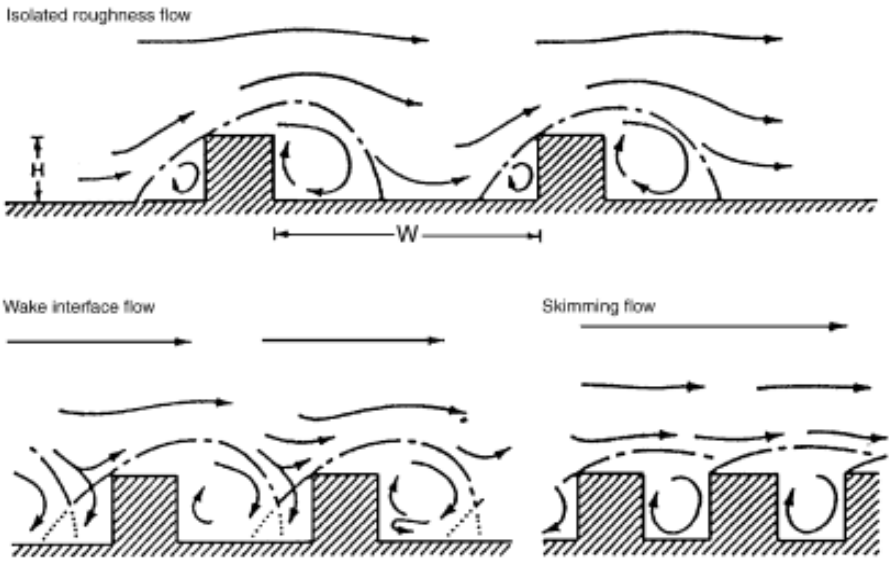



Figure 2.9. Perpendicular wind dependant turbulence conditions in canyons [Oke (1988)]

- Skimming flow: in some extreme cases the wind can cause a single vortex in the canyon which means the pollution is circulating back and cannot escape from the canyon.

The ADMS-Urban model is the extension of the ADMS model which specifically designed to have good prediction accuracy in the urban area McHugh et al. (1997).

Many studies utilized the ADMS-Urban model to predict the concentration level of different air pollutants. Righi et al. used the model to predict the concentration level of carbon monoxide for Ravenna, Italy. They concluded that the achieved accuracy of the model is very good on the low-temporal level (diurnal), however, the predictions generated by the model was underpredicting slightly the actual concentration level observations [Righi et al. (2009)].

There is also Gaussian air pollution dispersion models specifically developed for predicting the concentration levels in the urban area. The Operational Street Pollution Model (OSPM) was designed to cope well with the urban canyon effect. The model was utilized for in many modelling scenarios [Vardoulakis et al. (2007)].

Kukkonen et al. evaluated the OSPM model in one of the streets of Helsinki. He concluded that it is possible to utilise the street canyon dispersion model with reasonable accuracy using modelled urban background pollution and modelled meteorological data for carbon monoxide concentration level prediction [Kukkonen et al. (2003)].

Rzeszutek et al. evaluated the OSPM model in one of the streets in Krakow, Poland. They reported good prediction accuracy for $PM_{10}$ and $PM_{2.5}$ hourly conceration levels for the mod-

elling street canyon [Rzeszutek et al. (2018)].

Dezzutti et al evaluated multiple dispersion techniques (including STREET, OSPM, AEOLI-USF, STREET BOX and SEUS models) to predict hourly $NO_x$ concentration levels for one of the street canyons in Buenos Aires, Argentina. They reported good prediction accuracy using the SEUS model [Dezzutti et al. (2018)].

### 2.2.4 Land Use Regression approaches

Briggs et al. [Briggs et al. (1997)] developed an entirely new approach to air pollution modelling. Their approach was considering topographical, geographical and pollution-related (e.g. traffic emission information) information of the monitoring location and predicted pollution concentration levels based on these features using regression algorithm (which gives the name for these type of models, land use regression models).

The central idea behind the land use regression model is to extract the essential features of the monitoring station and the surrounding area (the buffer area) which include building numbers, road length, traffic volumes, buildings' height, land use and topographical information. Based on these features, a linear regression model can be trained where each feature have weights which describe how much contribution can be derived from that single feature (these weights are learned by the Linear Regression algorithm which configures these weight to reduce the prediction accuracy). This is quite an important property for the early models, because, with this method, researchers could rank the features and evaluate their importance related to the observed pollution concentration levels. This could help them identify the main pollution issues in the target area [Briggs et al. (1997)].

Land use regression models were developed in the past and evaluation of them was carried out:

- Cyrys et al. concluded that the land use regression model for Munich could achieve satisfying prediction for annual $NO_2$ concentration level predictions [Cyrys et al. (2005)]

- Marshall et al. developed a land use regression model in the Greater Vancouver area and their evaluation showed good correlation to monitoring data for annual $NO$ and $NO_2$ concentration level predictions [Marshall et al. (2008)]

- Gulliver et al. developed a land use regression model for London with good annual prediction accuracy and they concluded that it is prediction quality is equivalent to prediction of the existing air pollution dispersion models [Gulliver et al. (2011)]

- Liu at el. developed a land use regression model to predict the annual $PM_{10}$ and $NO_2$ concentration levels in the Shanghai, China modelling area. They reported good prediction accuracy by that model [Liu et al. (2016)]

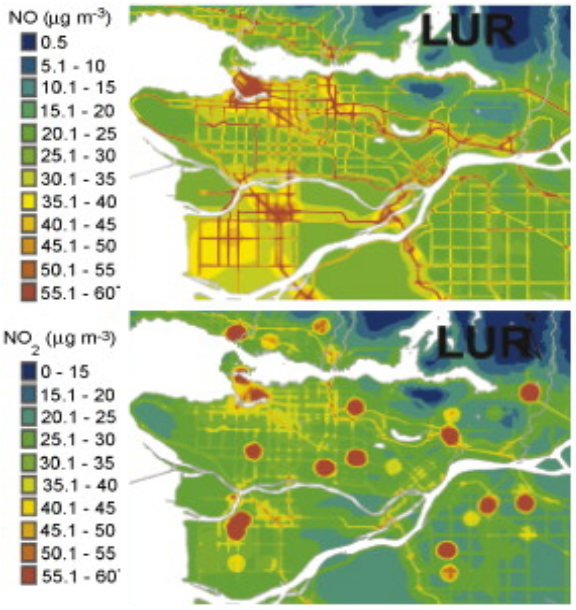



Figure 2.10. Visualization of the prediction of the Land Use Regression method for annual $NO$ and $NO_2$ concentration levels [Marshall et al. (2008)]

- De Hoogh et al. created a land use regression model for annual concentration level predictions for multiple pollutants (including $NO_2$, $PM_{2.5}$ and $O_3$ for multiple modelling area in Western Europe. They reported good accuracy for their approach [De Hoogh et al. (2018)]

- Naughton et al. developed a land use regression model exploiting wind sector based data for predicting $NO_2$ concentration levels across Ireland. They reported good correlation for this approach [Naughton et al. (2018)]

- Larkin et al. developed a land use regression model for annual $NO_2$ concentraiton levels using data from more than 5000 monitoring stations around the world. They validated the model and concluded that it produced predictions with good correlation to the observations [Larkin et al. (2017)]

Land use regression models are used only to predict annual and monthly averages because all the features are insufficient to be able to predict hourly changes in concentration levels.

Hoek et al. indicated that developing Land Use Regression model which can produce prediction with high temporal and spatial resolution is the interest of study [Hoek et al. (2008)].

Isakov et al. indicated that predicting hourly averages of pollutant concentration levels with the Land Use Regression approach is challenging. They stated that one fundamental problem for predicting hourly averages of concentration levels was to collect data with the necessary temporal resolution but they were not considering the regression algorithm prediction capability used for

the prediction. [Isakov et al. (2012)]

The existing Land Use Regression methods use the Linear Regression statistical regression algorithm to learn the relationship between the input data (the land use related data) and the observed concentration pollution levels. The next section introduces the Linear Regression algorithm to understand how it learns from the input data and how the algorithm does generate the predictions.

Linear regression is a method to create prediction based on the following equation:

$$\hat{y}(w, x) = w_0 + w_1 x_1 + ... + w_m x_m, \tag{2.7}$$

where $\hat{y}$ is the prediction for the input feature vector $x = \{x_1, x_2, ..., x_m\}$, $x_i$ are the features, $w_0$ called the intercept and $w_i$ are the coefficients [Weisberg (2005)].

There are multiple ways to calculate the internal weights, but the most often used method uses the Ordinary Least Squares optimization where it solves the following mathematical equation

$$argmin(\sum_{\forall x_i \in X} (\hat{y}_i(w, x_i) - y_i)^2), \tag{2.8}$$

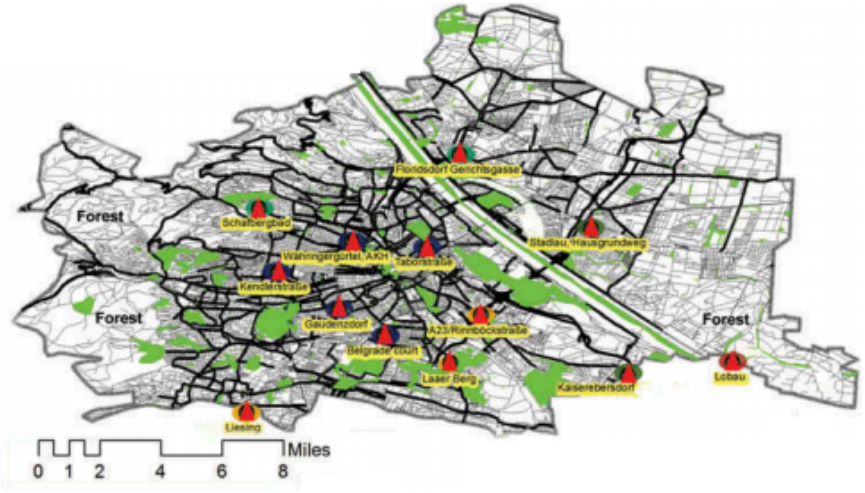
where $X = \{x_1, x_2, ..., x_n\}$ is the set of the feature vectors of the observations and $y_i$ are the target value for each observation.

Linear regression can only discover linear relations between the target value of the observation and the features, however, these relations (represented by the coefficients) can be ranked and described very well if the input data is normalized. This property of the algorithm established its popularity because researchers could understand the main factors of predictions [Weisberg (2005)].

To understand how the Land Use Regression method utilizes the Linear Regression algorithm for the concentration level prediction, Figure 2.11 shows modelling area and the generated concentration level prediction equations for a study which applied the model for Vienna [Alam & McNabola (2015)].

The variables of the equations are described in the following list:

- $V_1$: Major road length in the buffer
- $V_2$: Open space area
- $V_3$: Population density
- $V_4$: Temperature (Celsius)
- $V_5$: Rainfall/precipitation (mm)
- $V_6$: Maximum sustained wind speed (km/h)




*Vienna 1* $\text{Ln(PM)} = 5.041 + 2.543 \times 10^{-2} V_1 - 8.970 \times 10^{-2} V_2 + 8.233 \times 10^{-6} V_3 - 1.935 \times 10^{-2} V_4 - 3.475 \times 10^{-2} V_5 - 6.869 \times 10^{-1} V_6$

*Vienna 2* $\text{Ln(PM)} = 4.906 + 3.242 \times 10^{-2} V_1 - 9.933 \times 10^{-2} V_2 + 6.304 \times 10^{-6} V_3 - 2.098 \times 10^{-2} V_4 - 3.925 \times 10^{-2} V_5 - 5.936 \times 10^{-1} V_6$

Figure 2.11. The modelling area and the developed Linear Regression equations for daily fine particulate concentration level predictions [Alam & McNabola (2015)]

The two equations cover different time periods, therefore, the observed concentration levels and the input data are different.

### 2.2.5 High-temporal pollution modelling in the urban area

Recent updates in the air quality directives have generated interest in understanding the hourly concentration level changes of the $NO_2$ air pollution [WHO (2000)].

Air pollution dispersion models and Land Use Regression models have been used to predict annual concentration level of many pollutants, however, there are only a small number of studies investigating the high-temporal predictions of these models.

The air pollution dispersion models struggle to make accurate hourly concentration level predictions because of the uncertainty in the input data [Berkowicz et al. (2008); Owen et al. (2000); Vardoulakis et al. (2007); Morgenstern et al. (2007)]:

- uncertainty in the vehicle emission inventory data
- uncertainty in the fleet composition data
- uncertainty in the traffic estimation data

One approach to overcome these uncertainties is to try different modelling scenarios with increased and decreased numbers in different input data (such as increased emission per one vehicle type or increased number of vehicles) to tune the prediction to get more accurate predictions to the actual observations [Westmoreland et al. (2007)]. This approach, however, needs expert knowledge to carefully tune the input parameters which makes it hard to implement for city-wide modelling application.

The Land Use Regression method gives accurate concentration level predictions at low-temporal level (e.g. annual and monthly) [Brauer et al. (2003); Briggs et al. (1997); Stedman et al. (1997); Hochadel et al. (2006)] similarly to the air pollution dispersion models. The method, however, struggles to make accurate predictions on the high-temporal level because:

- the input data only contains low-temporal data (e.g. the number of buildings within the buffer area doesn't change hourly)[Briggs et al. (2000)]
- the Linear Regression algorithm fails to provide an accurate statistical regression model for the hourly concentration level predictions [Champendal et al. (2014); Sánchez et al. (2011)]

### 2.2.6 Evaluation methods

The chapter introduced many air pollution prediction models including application case studies, however, the chapter only described the models' accuracy in general. This section of the thesis is dedicated to introducing the accuracy metrics applied in these studies because reducing the prediction error (in other words, improving the accuracy) is the main focus of this thesis.

The air pollution models can generate numeric concentration level predictions. These predictions can be then compared to the concentration level observations. This process defines the accuracy of the given method. The literature has multiple methods to define the way of calculating the accuracy, but the following ones are the most frequently applied ones: mean absolute error ($MAE$), root mean squared error ($RMSE$), normalised mean squared error ($NMSE$), Pearson correlation coefficient ($r$), fractional bias ($FB$), geometric mean bias ($MG$), geometric variance ($VG$), predictions are within a factor of two of observations ($FAC2$).

#### 2.2.6.1 Mean absolute error

$MAE$ is defined by the following equation:

$$MAE = \frac{1}{n}\sum |y_i - \hat{y_i}| \tag{2.9}$$

where $n$ is the number of the observations, $y_i$ is the observed target value, $\hat{y_i}$ is the prediction produced by the model. Mean absolute error ($MAE$) indicates the expected average magnitude of error for the prediction based on the validation process. It only describes the magnitude of the error and not the direction. The perfect model (which model would produce exactly the same

concentration level predictions as the observations) would produce $MAE = 0$. The zero $MAE$ accuracy level can be achieved only if the predictions are identical to the observations.

#### 2.2.6.2 Root mean squared error

Root mean squared error ($RMSE$) is defined by the following equation:

$$RMSE = \sqrt{\frac{1}{n}\sum (y_i - \hat{y_i})^2} \tag{2.10}$$

.

$RMSE$ tells us the average magnitude of error, but it has a special property because it gives more penalty for larger errors. Analysing $MAE$ and $RMSE$ gives more information about the variation of the error. If the difference of $RMSE$ and $MAE$ is low then the variation of the error is low and the predictions have the same magnitude of the error. The perfect model would produce $RMSE = 0$. The zero $RMSE$ accuracy level can be achieved only if the predictions are identical to the observations.

#### 2.2.6.3 Normalised mean squared error

Normalised mean squared error ($NMSE$) is defined by the following equation:

$$NMSE = \frac{\frac{1}{n}\sum (y_i - \hat{y_i})^2}{\frac{1}{n}\sum y_i * \frac{1}{n}\sum \hat{y_i}} \tag{2.11}$$

.

Normalised mean squared error helps to understand the normalized prediction error. The $RMSE$ and $MAE$ error levels are insensitive to the absolute value of the observations, therefore the error contribution of a single prediction-observation pair is independent of the observation value itself. The prediction of 101 $\mu gm^{-3}$ when the observation is 100 $\mu gm^{-3}$ contributes with the same level as the prediction of 11 $\mu gm^{-3}$ in the case of observation of 10 $\mu gm^{-3}$ for the $RMSE$ and $MAE$ accuracy levels. The first case only has 1 percent error, but the second case has 10 percent error. To overcome this issue, $NMSE$ uses the normalized error to the observation and summarizes it. When two models need to be compared, combining the $NMSE$ with $MAE$ or $RMSE$ gives a nice understanding of how to two models introduce errors on the different scale of the observation range (e.g. if the $RMSE$ levels are the same, but the $NMSE$ shows lower value, then that implies that we have less error in the lower end of the observation range). The perfect model would give the $NMSE = 1$ level and only the matching prediction-observation pairs can achieve this level.

#### 2.2.6.4 Correlation coefficient

Pearson correlation coefficient is defined by the following equation:

$$r = \frac{cov(X, Y)}{\sigma_X \sigma_Y}, \tag{2.12}$$

where $cov(X, Y)$ is the covariance of the $X$ and $Y$ and the $\sigma$ is the standard deviation.

It expresses the strength of the linear correlation between the two variables. If it is close to $+1$ or $-1$ then there is a strong linear relationship between them. In general, prediction with higher r value has the better approximation of the observations.

#### 2.2.6.5 Fractional bias

Fractional bias ($FB$) is defined by the following equation:

$$FB = \frac{\frac{1}{n}\sum y_i - \frac{1}{n}\sum \hat{y}_i}{2 * (\frac{1}{n}\sum y_i + \frac{1}{n}\sum \hat{y}_i)} \tag{2.13}$$

.

Fractional bias expresses the average direction of the predictions against the observations. The model is overpredicting if the $FB$ is greater less than zero, otherwise, the model is underpredicting. Fractional bias helps to understand the general prediction quality and gives a clear explanation of the predictions relative to the observations. The perfect model would give $FB = 0$, however, this value can be achieved by having non-matching prediction-observation pairs as the errors can cancel out and result in zero $FB$ value.

#### 2.2.6.6 Geometric mean bias

Geometric mean bias ($MG$) is defined by the following equation:

$$MG = exp(\frac{1}{n}\sum ln(y_i) - \frac{1}{n}\sum ln(\hat{y}_i)) \tag{2.14}$$

.

The geometric mean bias represents the bias of the prediction to the observations, similarly to the fractional bias ($FB$). It is less sensitive to the outliers compare to the fractional bias due to the fact of the geometric nature. The perfect model would give $MG = 1$, however, this can be achieved by having non-matching prediction-observation pairs as the errors can cancel out and result in $MG = 1$ level.

#### 2.2.6.7 Geometric variance

Geometric variance ($VG$) is defined by the following equation:

$$VG = exp[\frac{1}{n}\sum (ln(y_i) - ln(\hat{y}_i))^2] \tag{2.15}$$

.

Geometric variance helps to understand the error level variance. Low values represent predictions that have consistent error levels and high values correspond to the large variation of the error levels. The perfect model would give $VG = 1$ and only the matching prediction-observation pairs can achieve this level.

#### 2.2.6.8 Predictions are within a factor of two of observations

Predictions are within a factor of two of observations ($FAC2$) is defined by the following equation:

$$FAC2 = \frac{1}{n} \sum \begin{cases} 1, & \text{if } 0.5 \leq \frac{\hat{y}_i}{y_i} \leq 2.0 \\ 0, & \text{otherwise} \end{cases} \tag{2.16}$$

.

This method gives an easily interpretable result of the prediction-observation data. The perfect model would give $FAC2 = 1$, however, this level doesn't represent the perfect observation-prediction match as the predictions only require to be a certain range of the observations.

#### 2.2.6.9 Definition of the good air pollution model

Hanna and Chang have reported the definition of the good air pollution model in 2004 [Chang & Hanna (2004)]. They defined the good model using the following criteria:

- The fraction of model predictions within a factor of two of observations is about 50 percent ($FAC2 > 0.5$)
- The mean bias is within $\pm$30 percent of the mean ($0.7 \leq MG \leq 1.3$)
- The random scatter is about a factor of two of the mean ($VG < 2$)

The study analysed multiple case studies (e.g. [Allwine et al. (2002); Britter & Hanna (2003); Hanna et al. (2003)]) and selected these criteria based on the models produced the most accurate predictions by the accuracy metrics they have in common. The authors also stated that these criteria levels need to be revised for new model evaluation exercises.

They have revised their first paper in [Hanna & Chang (2012)], however, the update contains weaker criterias for urban air pollution models, because the investigated field studies [Allwine et al. (2002); Allwine & Flaherty (2006); Watson et al. (2005); Allwine & Flaherty (2007)] demonstrated less accurate prediction results:

- The fraction of model predictions within a factor of two of observations is about 50 percent ($FAC2 > 0.3$)
- The mean bias is within $\pm$67 percent of the mean ($0.33 \leq MG \leq 1.67$)
- The random scatter is about a factor of three of the mean ($VG < 3$)

#### 2.2.6.10 Summary of the evaluation methods

The introduced air pollution modelling studies have used various accuracy evaluation metrics. Table 2.4 shows the summary of the applied accuracy evaluation method including the air pollution model category developed in the studies.

| Literature | Model | Temporal level | RMSE | MAE | NMSE | R | FB | MG | VG | FAC2 |
|---|---|---|---|---|---|---|---|---|---|---|
| Hanna et al. (2001) | Deterministic | Daily | | | X | | X | X | X | X |
| Levy et al. (2002) | Deterministic | Annual | | | | X | | | | |
| Carruthers et al. (2003) | Deterministic | Annual | | | X | X | X | | | X |
| Kalhor & Bajoghli (2017) | Deterministic | Annual | | | | | | X | X | X |
| Righi et al. (2009) | Deterministic | Annual | | | X | X | X | | | X |
| Vardoulakis et al. (2007) | Deterministic | Daily | | | X | X | X | | | X |
| Kukkonen et al. (2003) | Deterministic | Hourly | | | | X | X | | | |
| Rzeszutek et al. (2018) | Deterministic | Annual | | | X | X | X | | | X |
| Dezzutti et al. (2018) | Deterministic | Hourly | | | X | | X | | | X |
| Berkowicz et al. (2008) | Deterministic | Monthly | | | | X | | | | |
| Owen et al. (2000) | Determinitic | Hourly | | | | X | | | | |
| Christensen (1997) | Numeric | Weekly | | | | X | | | | |
| O'Neill & Lamb (2005) | Numeric | Hourly | | X | | X | | | | |
| Oettl et al. (2001) | Numeric | Hourly | | | X | X | X | | | X |
| Gidhagen et al. (2004) | Numeric | Hourly | | | | X | | | | |
| Mueller et al. (2015) | Statistical | Annual | X | | | X | | | | |
| Pohoata & Lungu (2017) | Statistical | Daily | | | | X | | | | |
| Briggs et al. (1997) | Statistical | Annual | | X | | X | | | | |
| Cyrys et al. (2005) | Statistical | Annual | X | | | X | | | | |
| Marshall et al. (2008) | Statistical | Annual | | | | X | X | | | |
| Gulliver et al. (2011) | Statistical | Annual | X | | | X | | | | |
| Liu et al. (2016) | Statistical | Annaul | | | | X | | | | |
| De Hoogh et al. (2018) | Statistical | Annual | X | | | X | X | | | |
| Naughton et al. (2018) | Statistical | Annaul | | | | X | | | | |
| Larkin et al. (2017) | Statistical | Annual | X | X | | X | | | | |
| Lu & Fang (2002) | Statistical distribution | Hourly | | | | X | | | | |
| Giavis et al. (2008) | Statistical distribution | Hourly | X | X | | | | | | |

Table 2.4. Summary of the applied accuracy evaluation techniques in the literature

## 2.3 Advanced statistical regression algorithms

The last section of the chapter introduces statistical regression algorithms from the machine learning field which algorithms can be the surrogate statistical regression algorithm for the Linear Regression to increase the overall prediction accuracy for the Land Use Regression approach for hourly concentration level predictions. The decision to which algorithm to include to this list was based on previous studies where algorithms were solving similar environmental prediction problems with better prediction accuracy (e.g. Neural Network Regression or Decision Tree Regression methods, Random Forest Regression [Champendal et al. (2014); Sánchez et al. (2011)]) or algorithms were successfully applied to non-linear regression tasks (e.g. Nearest Neighbour Regression, Support Vector Regression [Gardner & Dorling (1999); Tso & Yau (2007)]).

The section will also demonstrate that these algorithms can solve the challenging non-linear concentration level prediction task by applying the algorithms to a very simple prediction example. This example is based on [Sánchez et al. (2011)], where the authors discussed the intra-day variation of the $NO_2$ pollution concentration level. The example uses the simplified version of this data (Figure 2.12), which only contains one independent variable (hour of the day) and one dependent variable (pollution level concentration) only. This simple data helps to demonstrate to the problem of the non-linear regression prediction task. This data will be feed into the algorithms and the algorithms will be applied to the same data to see how the algorithms can solve this simplified problem. Figure 2.12 also shows that the Linear Regression algorithm struggles to make accurate predictions even in this simplified example because it can only fit a single line to the observation and it is not sufficient for non-linear regression problem such as pollution concentration level predictions.

### 2.3.1 Nearest neighbour regression

Nearest neighbour regression is a simple algorithm which uses the whole train dataset to find the $k$ closest observations to the record which needs a prediction. The parameter $k$ defines the number of closest neighbours for the method (e.g. $k = 1$ means that the method will consider the closest neighbour, while $k = 3$ means that the method will find the three closest neighbours and use them to make the prediction). The prediction $\hat{y}$ is calculated based on the closest neighbours observation $y$ values by averaging them. The distance is defined by an equation and which distance is expressed by the Minkowski distance function:

$$(d(x_i, x_j) = (\sum_{k=1}^{m} |x_{i,k} - x_{j,k}|^p)^{p^{-1}}, \tag{2.17}$$

where $x_i, x_j$ are feature vectors) which can be used to express other distance functions (Manhattan ($p = 1$), Euclidean ($p = 2$)) by simply changing the $p$ parameter. The $k$ parameter defines the number of neighbours for the model [Altman (1992)].

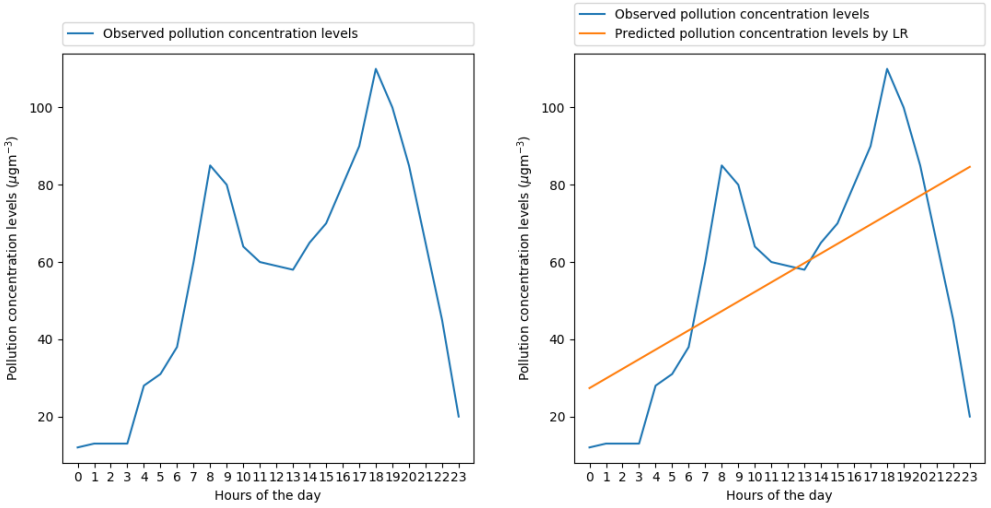



Figure 2.12. Simplified example data for the non-linear regression task (left) and the predictions on this example by the Linear Regression algorithm (right)

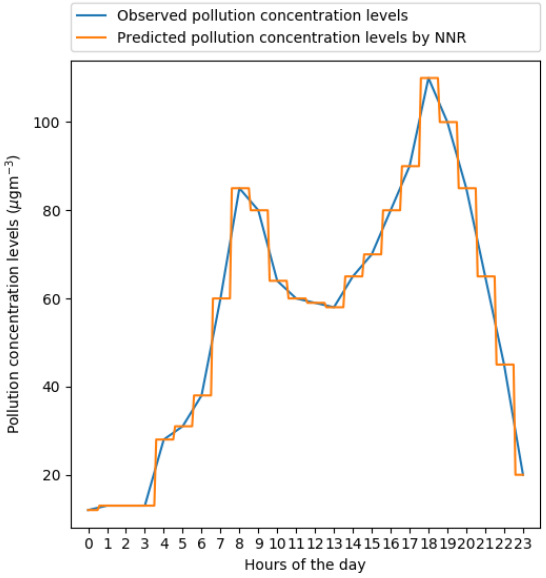



Figure 2.13. Predictions by the nearest neighbour regression algorithm on the example dataset

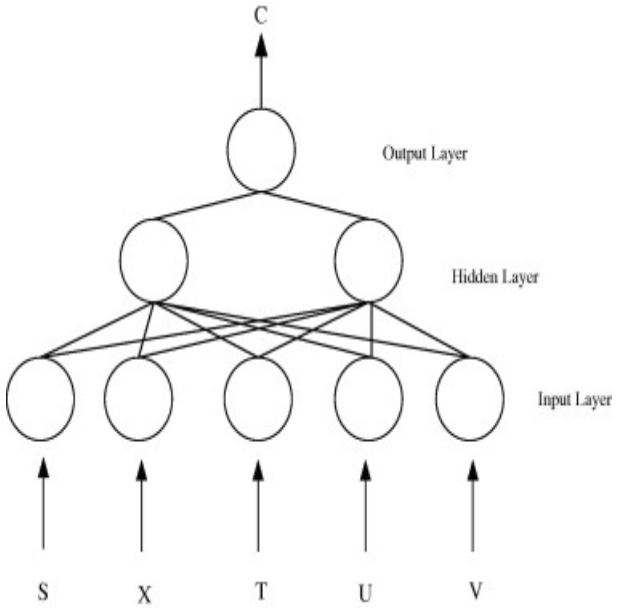



Figure 2.14. Visualization of an example neural network neuron structur [Wang et al. (2011)]

Figure 2.13 shows the predictions by the algorithm on the example dataset. It shows steps in the predictions as the input dataset only contains discrete measurements for every hour. The figure shows that the algorithm does not struggle to solve the non-linear regression task as the prediction is generated by using the closest neighbour (the data point itself in this case) from the input dataset.

### 2.3.2 Artificial neural network regression

Artificial neural network regression follows the idea of a cell located in brains called the neurones. This cell has one output and many input connections and it creates an output signal (called activation) if the signals from the inputs are strong enough. In theory, the brain is just a huge network of neurons therefore, it is possible to create an artificial brain having a large number of artificial neurons connected through as a weighted graph [Rumelhart et al. (1986)].

The artificial representation of the neuron is a node which has weights for each input and simulates the activation process by having an activation function ($\phi(\sum w_i a_i)$ where $a_i$ is the activation output of a node from the previous layer and $w_i$ is the corresponding weight). Figure 2.14 shows the connected layers of neurons. Neurons often have two types of different implementations depending their activation functions: linear ($\phi(x) = x$) and sigmoid ($\phi(x) = \frac{1}{1+exp(-x)}$).

Figure 2.15 shows the predictions by the algorithm on the example dataset. The predictions are correlating well with the observations as the internal structure of the neural network contains weights which allow the algorithm to produce non-linear predictions.

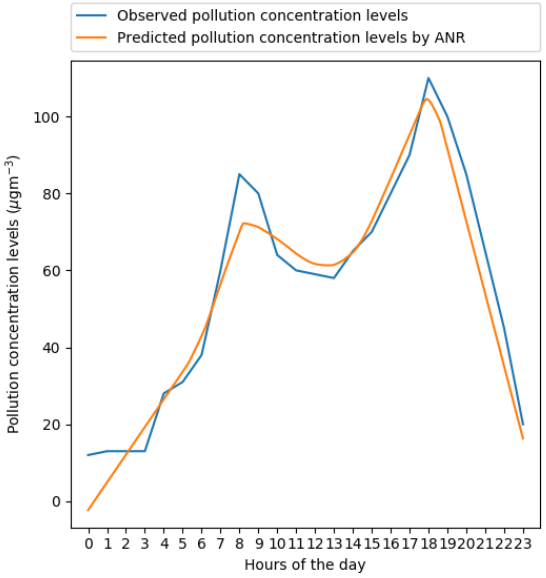



Figure 2.15. Predictions by the artificial neural network regression algorithm on the example dataset

### 2.3.3 Support machine vector regression

The support vector machine regression creates a hyperplane or a set of hyperplanes to separate the multi-dimensional input space. This hyperplane is calculated to have the largest margin between the target feature's minimum and maximum value since in general, the mathematical function (which describes the hyperplane) which has the largest margin will have the best approximation for the prediction target feature. A quadratic mathematical problem can be formulated to find the best function which problem has an interesting property: it uses a kernel function to distort the input features value space (for example the kernel can be linear, polynomial, gaussian, radial basis function (rbf), etc.). This quadratic mathematical problem contains the penalty parameter ($C$) for the wrong predictions and the problem maximizes the margin ($\epsilon$) for the hyperplanes (Figure 2.16). With custom kernel functions, non-linear problems can be predicted well with the support vector machine regression [Smola & Schölkopf (2004)].

Figure 2.17 shows the predictions by the algorithm on the example dataset. The internal kernel used by the algorithm is able to distort the feature input space (in this example, the hour of the day data) to generate a mathematical function which fits the observed concentration level through the day.

### 2.3.4 Decision tree regression

Decision tree regression is a decision tree induction based regression technique where tree induction algorithms create a decision tree and every leaf of this tree contains a prediction value and every other internal node has decision criteria (for example $x_4 < 0.5$). The decision tree is built

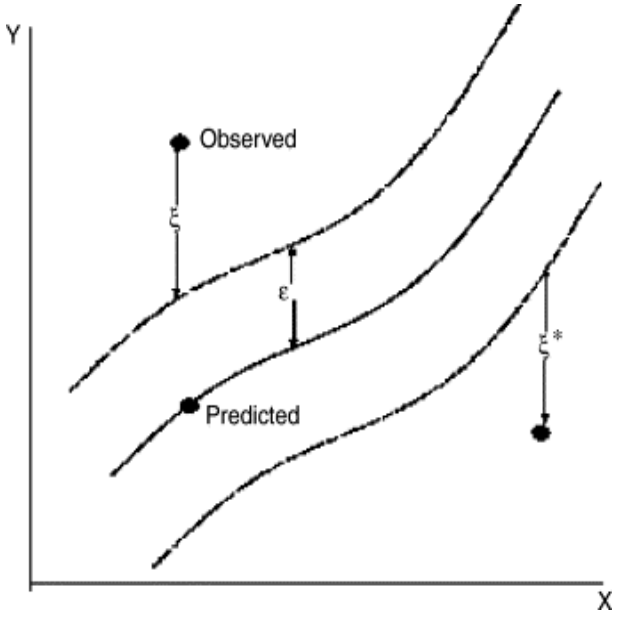



Figure 2.16. Example of the input space transformation for the SVR method to minimalise the margin [Vapnik (2013)]

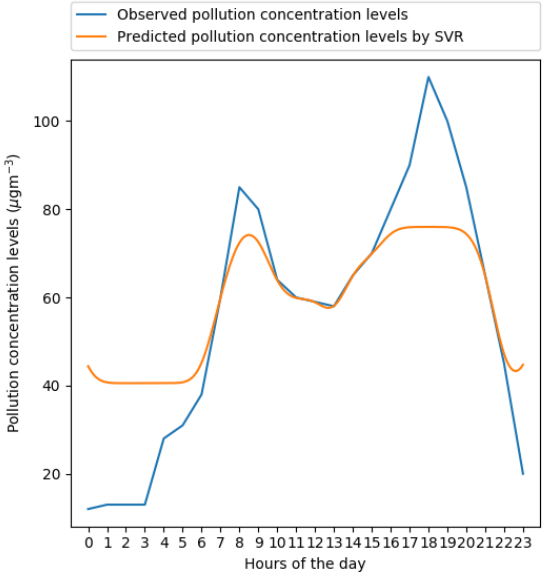



Figure 2.17. Predictions by the support vector machine regression algorithm on the example dataset

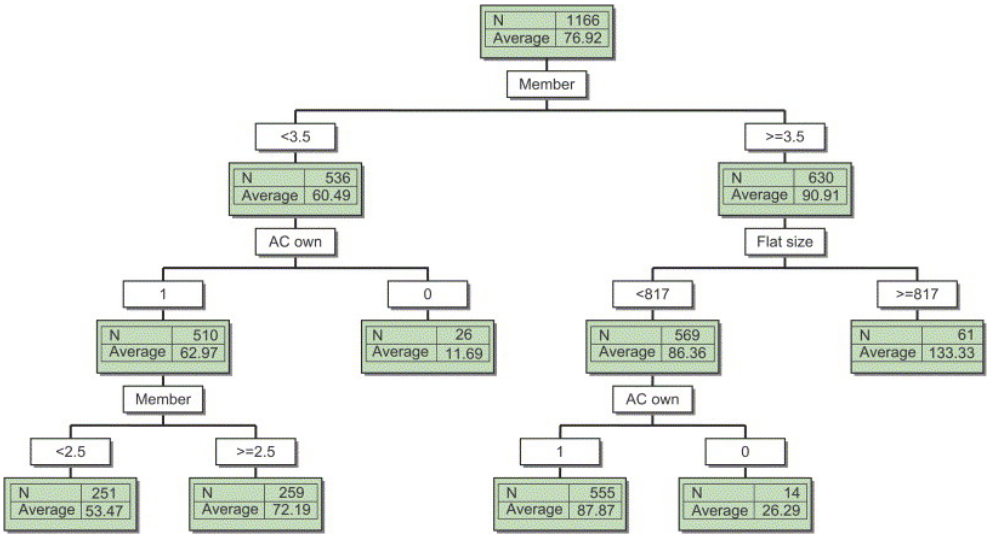



Figure 2.18. Example decision tree for statistical regression prediction [Tso & Yau (2007)]

to have the best fit for the training dataset and every prediction starts at the root, evaluates it, then decides to take the left or right children (if it is a binary decision tree) then evaluate all the internal node until it ends at a leaf node where there is a prediction value. Figure 2.18 shows an example of the decision tree regression model. There are many different tree induction algorithms in the literature where the algorithms terminate the tree induction process based on different criteria (e.g. depth of the tree or number of the observations in each node). Early termination of the tree induction process helps to avoid the overfitting to the given data and it helps to increase the generalization of the generated statistical regression model [Quinlan (2014)].

Figure 2.19 shows the predictions by the algorithm on the example dataset. The internal tree structure is able to predict the concentration levels with good correlation, however, it is only able to predict in steps as the input dataset only contains observations for discrete hours.

### 2.3.5 Random forest regression

Random forest regression is an ensemble method based on the decision tree regression. Instead of training one large decision tree for the regression, it follows the idea of the ensemble methods where the algorithms train models on the different random subsets of the train data (in terms of observations as well as features) and rank the created sub-models on the efficiency based on the other part of the training data (Figure 2.20). With this procedure, the method can randomly pick up an interesting part of the data and have a large number of efficient sub-models. The prediction is based on a voting procedure, where each sub-model has a vote (basically generates a prediction) and based on the average of the individual predictions, the final prediction is calculated [Breiman (2001)].

Figure 2.21 shows the predictions by the algorithm on the example dataset. The algorithm

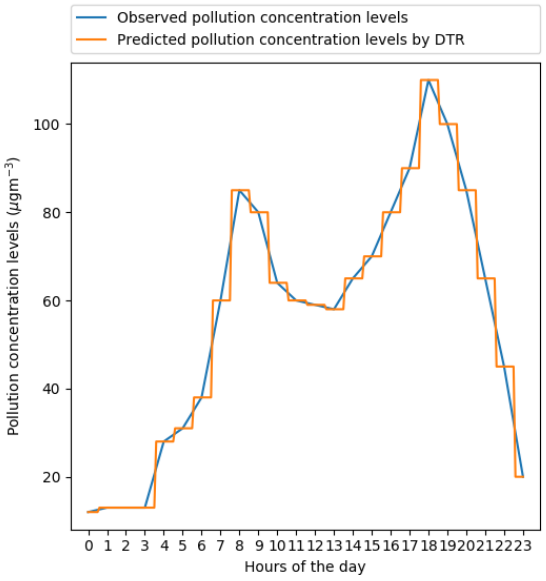



Figure 2.19. Predictions by the decision tree regression algorithm on the example dataset

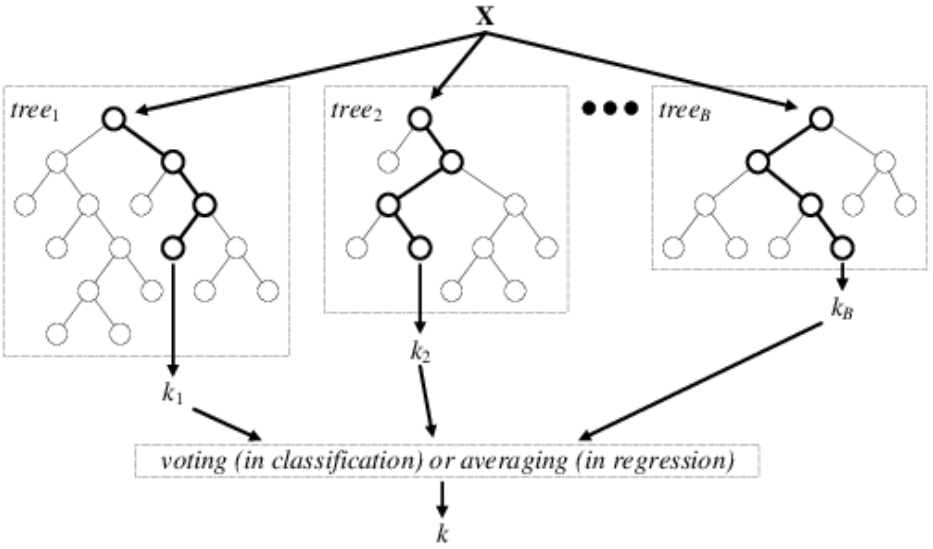



Figure 2.20. Example of the Random Forest Regression method [Verikas et al. (2016)]

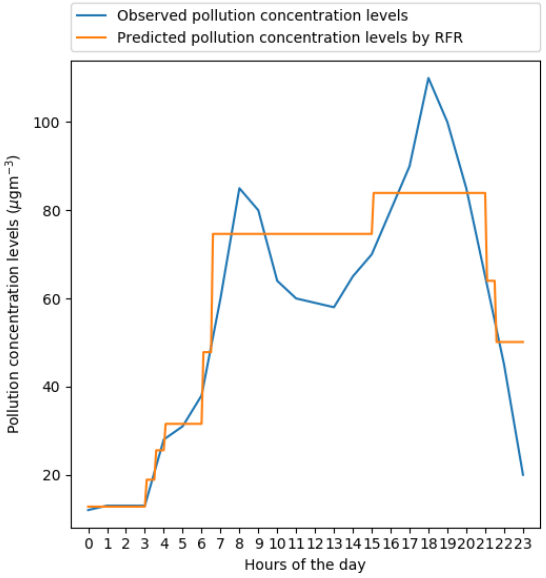



Figure 2.21. Predictions by the random forest regression algorithm on the example dataset

generates predictions with good correlation, however, it struggles to make very accurate predictions on this simple example because the available data is very small and the algorithm requires larger dataset to be able to exploit the prediction power of multiple decision trees.

## 2.4 Summary

In this chapter, the literature survey of the air pollution modelling has been presented. The survey introduced the problem caused by air pollution in the urban area including the description of the most concerning air pollutant, the nitrogen dioxide. This section includes important information for the rest of the thesis because the focus of this thesis is the high-temporal $NO_2$ pollutant concentration level modelling. Next, the chapter reviewed the current methods in the field of air pollution modelling including the state-of-the-art air pollution dispersion methods. This section introduced the existing methods for the pollution modelling which will provide the baseline predictions for the later comparisons. Then, the review of the Land Use Regression approach has been presented. The challenges of the high-temporal air pollution modelling are described in terms of the state-of-the-art air pollution dispersion and existing Land Use Regression approaches. This helps to form the experiments for the next chapter.

The literature reviewed in this chapter does not make it clear what is the prediction accuracy of the existing methods for high-temporal air pollution modelling. The next chapter will look into the investigation of the prediction accuracy of the existing methods and also the development of an accurate Land Use Regression model for hourly concentration level prediction.

# CHAPTER 3

## Statistical Regression approach for high-temporal environmental predictions

This chapter presents the empirical study to develop a statistical regression approach for hourly $NO_2$ concentration level predictions with comparable high-level accuracy rate to the current state-of-the-art air pollution dispersion models. The first step of this development is to establish an evaluation framework which supports the comparison of the different approaches. Using this evaluation framework, it is possible to compare the high-temporal prediction accuracy of state-of-the-art air pollution dispersion model and the existing Land Use Regression approach and experiment with the advanced machine learning regression techniques. Moreover, it allows determining the most accurate advanced machine learning technique for this given regression task. This information is a contribution to the Environmental Science field, because it gives a guideline on which existing model to use for the urban scale hourly $NO_2$ concentration level predictions to get the most accurate predictions.

In the first section (Section 3.1), the motivation of this work is explained which introduces the problem domain and reviews the relevant literature for the work described in this chapter. The application of a state-of-the-art air pollution dispersion model is then described along with the necessary dataset to generate the concentration level predictions (Section 3.2). This prediction output provides the necessary baseline for the models introduced in the rest of the chapter. The following section (Section 3.3) introduces the application of the existing standard Land Use Regression technique to the same area and discusses the difficulties of such a regression task. The fourth section of this chapter (Section 3.4) covers the application of different computationally intense regression methods including the hyperparameter tuning of these techniques to achieve

the best possible high-level prediction accuracy level. This section also compares the prediction output of the most accurate method with the prediction output of the state-of-the-art air pollution dispersion model. Finally, the Section 3.6 finalizes the chapter.

## 3.1 Motivation

The current state-of-the-art air pollution dispersion technique was developed to predict concentration levels (not just $NO_2$ but every type of air pollutants such as Particulate Matter (PM), Sulphur Dioxide ($SO_2$), etc.) on low-temporal resolution (e.g. annual or monthly) to understand the average exposure of a particular pollutant in the modelling area. Studies [Stocker et al. (2012); Namdeo et al. (2002); Berkowicz (2000); Cimorelli et al. (2005)] show that the implementations of this approach (e.g. ADMS [Carruthers et al. (1994)], OSPM [Vardoulakis et al. (2007)] predict annual and monthly concentration levels sufficiently accurately to carry out the required exposure analysis.

The air pollution dispersion models, however, struggle to make accurate hourly concentration level predictions because of the uncertainty in the input data [Berkowicz et al. (2008); Owen et al. (2000); Vardoulakis et al. (2007); Morgenstern et al. (2007)]:

- uncertainty in the vehicle emission inventory data
- uncertainty in the fleet composition data
- uncertainty in the traffic estimation data

One approach to overcome these uncertainties is to try different modelling scenarios with increased and decreased numbers in different input data (such as increased emission per one vehicle type or increased number of vehicles) to tune the prediction to get more accurate predictions to the actual observations [Westmoreland et al. (2007)]. This approach, however, needs expert knowledge to carefully tune the input parameters which makes it hard to implement for city-wide modelling application.

There is an orthogonal modelling approach to the air pollution dispersion models for environmental concentration level predictions. The approach uses historical observations to build a statistical regression model and applies this model to generate concentration level predictions. The core idea of this statistical regression approaches [Briggs et al. (2000)] is to extract information around the monitoring station (a rectangular shaped area called the buffer area) and use this data to predict the concentration levels as a regression task. The data extracted from the buffer area doesn't include the uncertain data (e.g. vehicle emission inventory data or fleet composition data) used by the air pollution dispersion models which gives the advantage to avoid using these input data. The Land Use Regression (LUR) method is the most popular implementation of this approach [Brauer et al. (2003); Briggs et al. (1997); Stedman et al. (1997); Hochadel et al. (2006)] where only land use related data used to train a Linear Regression algorithm. Using the

Linear Regression algorithm is beneficial as it produces a statistical regression model which can be interpreted easily (as the weights of each input feature explain the importance of the feature to the concentration level).

The LUR method gives accurate concentration level predictions at low-temporal level (e.g. annual and monthly) [Brauer et al. (2003); Briggs et al. (1997); Stedman et al. (1997); Hochadel et al. (2006)] similarly to the air pollution dispersion models. The method, however, struggles to make accurate predictions on the high-temporal level because:

- the input data only contains low-temporal data (e.g. the number of buildings within the buffer area doesn't change hourly)[Briggs et al. (2000)]

- the Linear Regression algorithm fails to provide an accurate statistical regression model for the hourly concentration level predictions [Champendal et al. (2014); Sánchez et al. (2011)]

The first point can be solved by adding high-temporal input data (e.g. weather data) to the existing land use input data. This helps the underlying statistical regression algorithm to have the required input data to discover the hidden relationship of the input data and the observer hourly $NO_2$ concentration levels. This addition, however, makes the regression problem complex as the input data now has a mix of low-temporal and high-temporal input data. Studies [Champendal et al. (2014); Sánchez et al. (2011)] indicate that the Linear Regression algorithm struggles to make accurate hourly concentration level prediction using this complex input data. There are, however, other statistical regression methods (e.g. Neural Network Regression or Support Vector Regression) as the advances in the machine learning field produced many different statistical regression algorithms recently [Gardner & Dorling (1999); Sánchez et al. (2011); Tso & Yau (2007); Champendal et al. (2014)]. Complex regression problems can be solved with these methods as they can extract the hidden relationship of the input data and the regression prediction target using their computationally intense internal structure. These methods differ from the Linear Regression algorithm as they require a certain level of tuning the make predictions sensibly as well as these methods require more computation to be able to make predictions.

The main goal of this chapter is to develop a statistical regression model capable of accurately predicting the hourly $NO_2$ concentration levels. Such a model would not rely on the uncertain data (e.g. vehicle emission inventory data) used by the state-of-the-art air pollution dispersion models and it would provide at least the same prediction accuracy level as the air pollution dispersion models (which is not possible with the existing LUR models). To achieve this goal, the following tasks have to be carried out:

- apply the air pollution dispersion model to generate prediction result. This result provides a baseline for further accuracy comparison.

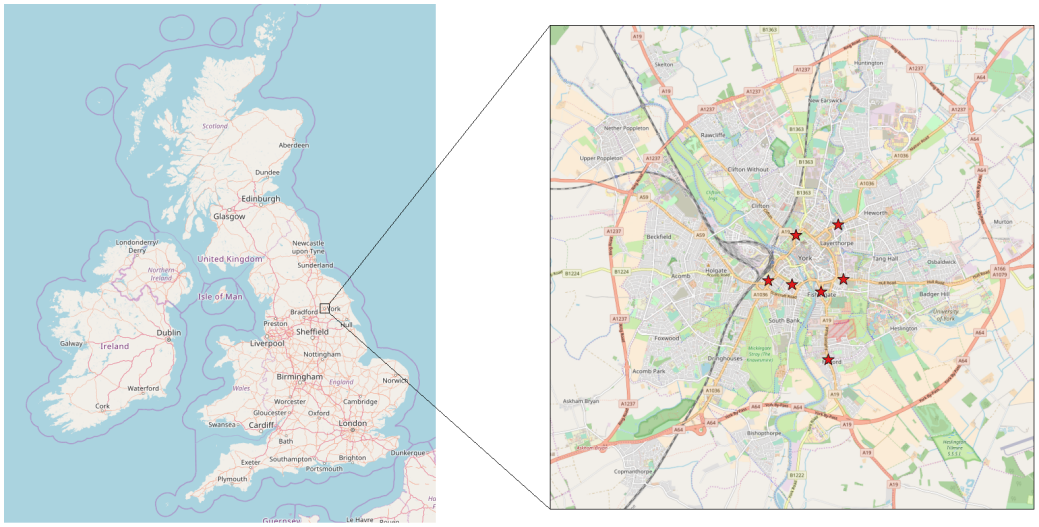



Figure 3.1. Geographical map of York with the monitoring station locations (red stars)

- apply the existing LUR approach. This application provides predictions to validate the outcome of the existing studies
- tune and apply other statistical regression approaches and compare the accuracy with previous model applications. This step provides an understanding of which algorithm provides the most accurate predictions on the hourly $NO_2$ concentration levels

## 3.2 Application of the Operational Street Pollution Model

The application of an air pollution dispersion model creates the baseline for further model comparison. Operational Street Pollution Model (OSPM) air pollution dispersion model was selected as it produces hourly $NO_2$ predictions with the same accuracy as the state-of-the-art Atmospheric Dispersion Modelling System (ADMS) air pollution dispersion model, but it is free to use for research purposes [Vardoulakis et al. (2007)]. Using the OSPM method helps to generate research materials which are reproducible and can be verified by other researchers easily. The WinOSPM 5.1.90 software contains the OSPM model including tools to convert the required data to the correct format.

To carry out the model application, York has been selected for the modelling scenario (Figure 3.1).

### 3.2.1 Input data requirement

The software requires the following data to make hourly $NO_2$ predictions and evaluate the prediction accuracy:

- Geographical information for the monitoring stations

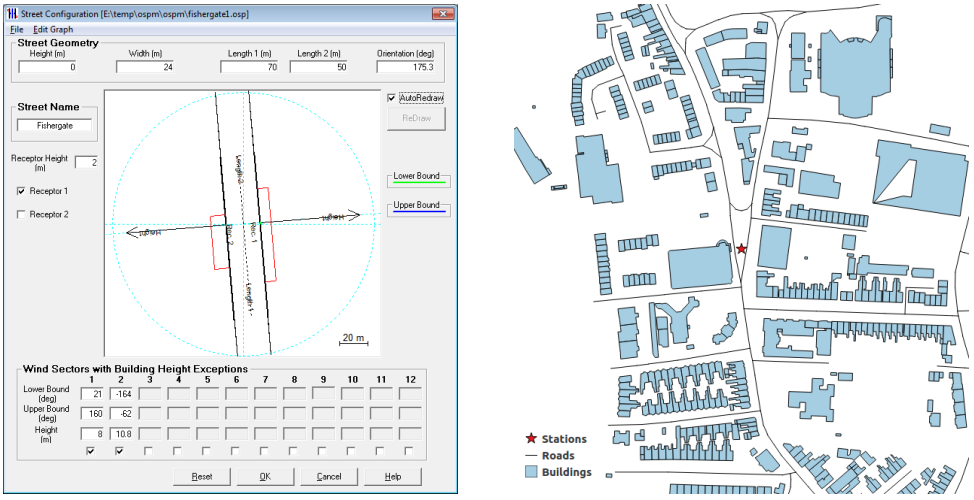



Figure 3.2. The WinOSPM representation (left) and the map (right) of the Fishergate monitoring station

- Traffic volume data
- Emission inventory database
- Meteorological data
- Background pollution data
- Monitoring (observation) data (required to carry out the evaluation of the generated predictions)

**Geographical information for the monitoring stations** The WinOSPM software requires the user to input the geographical data of the surroundings of the modelled receptor position which receptor position defines the prediction target location for the dispersion model (therefore the model is going to generate concentration level prediction at this specific location). This surrounding data includes the width and the orientation of the street canyon and the height and position of the buildings alongside the street. To calculate these data, building data from the Ordnance Survey's 2009 version of Mastermap$^{TM}$ Topography layer was acquired. This layer gives spatial information (e.g. geometry, surface area, etc.) about buildings within the area of interest. Figure 3.2 shows the WinOSPM representation of the surrounding of the Fishergate monitoring station.

**Traffic volume data** Traffic data was provided by the City of York Council's Transportation Managment Group where they developed a complete city scale traffic model. This model contains predicted average traffic volumes for each road including car, light goods vehicle (LGV) and heavy goods vehicle (HGV) counts. The dataset contains three time periods (morning peak period (from 7 AM to 9 AM), inter-period (from 10 AM to 4 PM), afternoon peak period (from 5

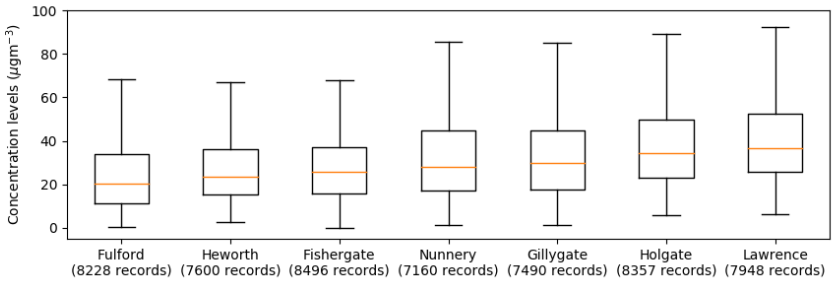



Figure 3.3. Hourly $NO_2$ observation data in York from its 7 monitoring stations that covers the time period between 1st January 2013 and 31st December 2013. The red line in the figure represents the median value of the available observations.

PM to 8 PM)) and it contains an hourly average traffic volume for each time periods. This dataset provides the geographical layout of the road network in York including the lane numbers and lane directions. The closest roads for each monitoring station have been selected and converted the traffic data into the right format.

**Emission inventory database** The National Atmospheric Emission Inventory group (`http://naei.beis.gov.uk/`) maintains the UK Vehicle Emission Inventory database which contains the required emission information for the air pollution dispersion model (e.g. petrol and diesel cars average emission data calculated for multiple years).

**Meteorological data** Meteorological data from the Weather Underground database (`https://www.wunderground.com/weather/api/`) has been acquired by using its API to download data. This database contains hourly average observations for cities and includes temperature, relative humidity, wind speed, wind direction, and pressure measurements. The relevant York dataset has been collected using this API. Unfortunately, this dataset does not contain solar radiation data.

**Monitoring (observation) data** The City of York Council (CYC) operates a network of high precision (chemiluminescence-based) instruments in York to monitor the air quality. Monitoring data from 7 roadside stations and 1 background station have been acquired which covers the time period between 1st January 2013 and 31st December 2013. Figure 3.3 shows a boxplot of the observations produced by each station. These readings are considered to be as low pollution levels as the higher percentile of observation data is below 55 ug/m3. Also, the observations at each station do not differ very much as the pollution levels are low in the most cases.

### 3.2.2 Accuracy evaluation of the OSPM model

Figure 3.4 shows the prediction output of the applied OSPM model. It contains 55859 $NO_2$ hourly concentration predictions resulting in 18.49 $\mu gm^{-3}$ RMSE and 13.93 $\mu gm^{-3}$ MAE high-

level accuracy levels. The predictions also have 0.53 NMSE, 0.69 R, 0.46 FB, 1.71 MG, 1.99 VG and 0.61 FAC2 levels. According to [Chang & Hanna (2004)], this model does not achieve the good model classification, because the MG level is exceeding the 30 percent acceptance limit, however the model meets the other criterias. The model fails to make accurate predictions at the time of high concentration levels which is in line with other studies findings such as [Berkowicz et al. (2008); Owen et al. (2000); Vardoulakis et al. (2007); Morgenstern et al. (2007)]. According to these studies, the main reason is the uncertainty in the underlying datasets (e.g. vehicle emission inventory database, estimated fleet composition, estimated traffic volumes). To further validate the result of this model application, this OSPM predictions result was compared at the Gillygate station with the result of [Westmoreland et al. (2007)] study. The comparison indicates 19.32 $\mu gm^{-3}$ RMSE error level for the OSPM model which is similar to the 18.5 $\mu gm^{-3}$ (9.6 ppb) reported RMSE level in the paper. They have not used the definition of the good model ([Chang & Hanna (2004)]) to classify their model, however, they have done an extensive sensitivity analysis on the input dataset to understand how to change the input dataset to get more accurate predictions.

## 3.3 Application of the standard Land Use Regression approaches

In the literature, there is an orthogonal approach to the air pollution dispersion modelling technique to generate concentration level predictions where a statistical regression model is trained based on the historical observations. The core idea of the statistical regression approaches [Briggs et al. (2000)] is to extract information around the monitoring station (a rectangular shaped area called the buffer area) and use this data to predict the concentration levels as a regression task. This regression task can discover the relationship between the input and the target data (in this case the concentration levels) and it does not need to use uncertain datasets (e.g. emission inventory data). The standard implementation of the Land Use Regression technique described in [Brauer et al. (2003); Briggs et al. (1997); Stedman et al. (1997); Hochadel et al. (2006)] was developed. This implementation includes the application of the hyperparameter-free Linear Regression regression algorithm and the usage of the following data sources:

- Monitoring data: using the hourly $NO_2$ concentration levels is essential for any prediction model as this provides readings of the pollution levels and this data provides the target data for the regression models

- Land use data: an example of this category is the area of green space within the specific area. The high proportion of the green area indicates low pollution level (clean air) in general as there is not much built-up area within the given area

- Building data: an example of this data category is the number of buildings which corresponds to the number of people living in the specific area. If we have high number of

buildings in one area then that can cause increased pollution levels (e.g. they commute every day by cars or they visit businesses in the area by car)

- Road data: an example of this category is the length of the road within the area. If there is large number of roads presented in one area that give the chance for heavy traffic during the commute hours, therefore, the pollution level can be high in this area

- Traffic data: one of the primary sources of the $NO_2$ pollutant is the vehicle emission, therefore, the data describes the amount of the cars and lorries within the given area would be an important information for any pollution model

- Meteorological data: distribution of the $NO_2$ pollutant is highly depending on the current weather circumstances. Strong wind can flush out all the pollution from the streets quickly if the direction is optimal (for a given street geometry) as well as strong wind can close down a street blocking the pollution to escape and allowing the pollution to slowly build up. Also, the pollution concentration level can decrease if it is raining as it will clear out the air from the pollutants as well as clouds during the rain can decrease the solar radiation which decreases the formation of $NO_2$ from other compounds in the air. Also, rain helps to decrease the $NO_2$ concentration levels because it flushes the pollution out of the air.

- Time related data: the regression model can have benefit having time related data for training such as hour of the day or month of the year as it can discover certain high-level processes purely from the data (e.g. summer months where schools runs do not happen therefore the pollution level can be lower in general comapred to the school periods)

To carry out the model application, York has been selected for the modelling scenario (Figure 3.1).

All the input data needs to be first converted to tabular format. The converted data then can be feed into the statistical regression algorithm to generate a regression model. This model contains all the internal information to generate the concentration level predictions.

### 3.3.1 Input data

The same data sources has been used as for the air pollution dispersion model application, however different data preprocessing was necessary to extract the data for the regression task. Also, further data sources similarly to [Hochadel et al. (2006); Stedman et al. (1997); Briggs et al. (1997)] were introduced as these studies provide a guideline on data used in previous studies. The standard 100 meter rectangular shape buffer area was selected similarly to [Gilbert et al. (2005); Morgenstern et al. (2007)].

**Monitoring data** The target of the regression task is to predict the hourly $NO_2$ concentration levels. The same dataset as in Section 3.2 was used which dataset is maintained by the City

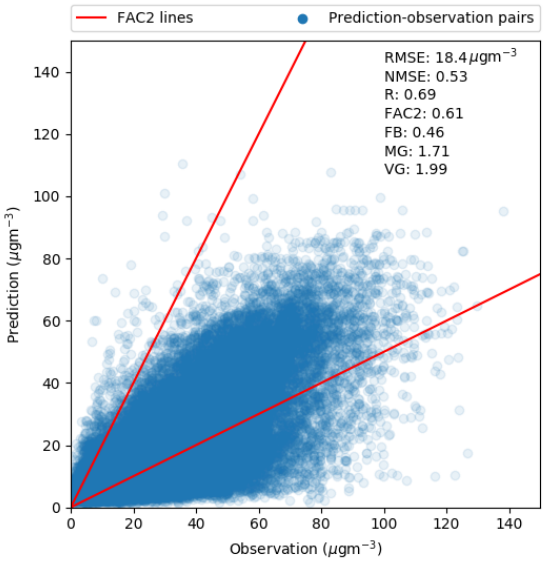



Figure 3.4. Hourly prediction and observation scatter graph for the OSPM model

of York Council's Air Management Group. This guarantees the fairness for model prediction accuracy comparison.

**Land use data** One of the most often used data source is the land use data for the Land Use Regression models [Briggs et al. (1997); Stedman et al. (1997); Sahsuvaroglu et al. (2006)]. Land use data has been collected using the Open Street Map database. The available data describes the areas (in polygons format) usage scenarios (e.g. leisure, green areas, farm, etc.). The following data for each buffer area (around the monitoring stations) were extracted: "landuse_area" and "leisure_area" which are proportional area measurements of the specific subcategory of the polygons to the buffer area in the database.

**Building data** The Ordnance Survey's 2009 version of Mastermap$^{TM}$ Topography layer data was used (similarly to the previous air pollution dispersion model application) to obtain building information for buffer area of each station. This layer gives spatial information (e.g. geometry, surface area, etc.) about buildings within the area of interest. This database has been processed and the number of the buildings and area of the buildings covered by each buffer area generated the "buildings" and "buildings_area" features.

**Road and traffic data** The same traffic data has been used as in the previous model application. However, two different types of data from this data source have been extracted. The first one only covers static (in time) information such as the length of all roads within the buffer area ("road_length") as well as the calculated "road_length" scaled to the roads' lane number ("road_lane_length"). The second type is the representation of the traffic amount appears within the buffer area. The roads within the buffer area were selected, then the traffic volume informa-

| Feature | Unit | Source | Data group |
|---|---|---|---|
| $no_2$ level | $\mu gm^{-3}$ | CYC | - |
| road_length | meter | Open Street Map | R |
| road_lane_length | meter | Open Street Map | R |
| buildings | - | OS Mastermap | B |
| buildings_area | area | OS Mastermap | B |
| landuse_area | area | Open Street Map | L |
| leisure_area | area | Open Street Map | L |
| traffic_car | vehicle*meter/hour | CYC | V |
| traffic_lgv | vehicle*meter/hour | CYC | V |
| traffic_hgv | vehicle*meter/hour | CYC | V |
| wind_direction | degree (angle) | Weather Underground | W |
| wind_speed | m/s | Weather Underground | W |
| temperature | celsius degree | Weather Underground | W |
| rain | indicator | Weather Underground | W |
| pressure | hPa | Weather Underground | W |
| hour | - | Generated | T |
| day_of_week | - | Generated | T |
| month | - | Generated | T |
| bank_holiday | indicator | Generated | T |
| race_day | indicator | Generated | T |

Table 3.1. Summary of the collected data

tion from the traffic model was calculated to generate the "traffic_car", "traffic_lgv", "traffic_hgv" information for each time periods (morning peak period (from 7 AM to 9 AM), inter-period (from 10 AM to 4 PM), afternoon peak period (from 5 PM to 8 PM)) available in the traffic model.

**Meteorological data** The same weather information data was used as in the previous model application. This data includes temperature, relative humidity, wind speed, wind direction, and pressure measurements.

**Time related data** Time-related indicators (e.g. hour of the day, day of the week, bank holiday, etc.) for the statistical regression model are important because the regression models can use this information to discover temporal patterns in the input data. Some York specific event indicator was included such as event (e.g. York horse races when tens of thousands of visitors come to the city leading to significantly higher traffic volumes than the normal at the certain time of day) indicator which affects the traffic pattern in the whole city.

Figure 3.5 shows the surroundings and the buffer area of the Fishergate station. This buffer area is a 100-meter wide rectangular area. This buffer area contains 31 buildings which are covering 50.11% of the buffer area. Also, the area contains 248 meters of road (464 single lane


Figure 3.5. Buffer area of the Fishergate monitoring station

meters). This buffer area does not contain any leisure nor landuse polygons. Table 3.1 contains the summary of the data prepared for the land use regression task.

### 3.3.2 Evaluation methodology of the statistical regression methods

A validation framework was implemented to determine the general accuracy of the standard LUR model. This framework consists the state-of-the-art location based leave one out cross validation (LOOCV) similarly to [Briggs et al. (2000); Cyrys et al. (2005); Marshall et al. (2008)]. This framework is an iteration based validation technique where one station data was left out from the regression training phase to build the model and the model is applied to that station data to generate predictions. Evaluation of the predictions and the observations is possible with this framework by calculating the error levels for each iteration. Using this approach helps to understand the average error level of the application of the method to a wider area as the framework provides an understanding of the error level of applying the model to an unknown (at least to an unknown area to the model) area. This validation framework is implemented using the scikit learn library [Pedregosa et al. (2011)] which contains extensively tested implementation of a large set of machine learning algorithm including regression algorithms as well as others.

### 3.3.3 Accuracy evaluation of the standard LUR model

The standard Land Use Regression model only uses land use data to train a Linear Regression model [Briggs et al. (2000)]. The method gives good accuracy level on the prediction of low temporal resolution (e.g. annual and monthly level) however studies [Briggs et al. (2000); Champendal et al. (2014); Sánchez et al. (2011)] suggest that this method struggle to make accurate hourly $NO_2$ concentration level predictions. Land use related data (building (B), land use (L), road (R) and traffic(V)) was selected from the preprocessed dataset to generate a dataset

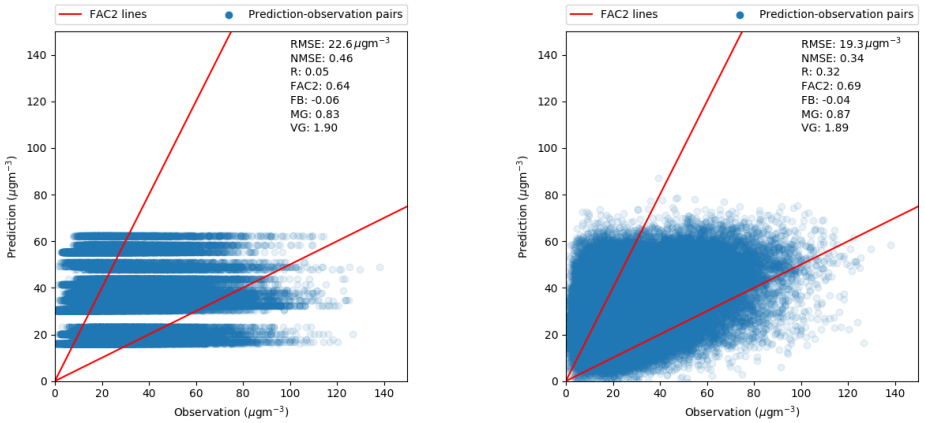



Figure 3.6. Hourly predictions and observations for the standard Land Use Regression (left) and the Linear Regression (right) models

which necessary for the standard Land Use Regression model and the Linear Regression method from the scikit-learn machine learning library was used to train the underlying regression model. The validation framework produced an overall 22.65 $\mu gm^{-3}$ RMSE and 18.02 $\mu gm^{-3}$ MAE error levels which indicate higher error levels than the state-of-the-art OSPM model's prediction accuracy levels. The model achieved 0.46NMSE, 0.05 R, -0.06 FB, 0.83 MG, 1.90 VG and 0.64 FAC2 accuracy levels. Figure 3.6 shows the predictions generated by the implemented standard Land Use Regression model. This figure shows that using the standard approach for hourly predictions struggle to make accurate predictions because the input data only contains low-temporal knowledge (e.g. number of buildings) which confirms the outcome of the previous studies [Champendal et al. (2014); Sánchez et al. (2011)]. This also explains why the correlation coefficient is very low. To understand how this approach is generating the concentration level predictions, the visualization of the observation-prediction pairs has been generated. Figure 3.6 shows this plot. The model generates an average concentration level prediction for each station and generates that only concentration level value for every hour for a given station (that explains the constant line-shaped prediction levels).

The Linear Regression method can be also trained using all the available preprocessed data (which is all the data used for the standard Land Use Regression approach plus the high-temporal time and weather-related data). The evaluation (using the cross-validation evaluation framework) of this method shows that this approach can achieve 19.39 $\mu gm^{-3}$ RMSE and 15.39 $\mu gm^{-3}$ MAE error levels. These predictions also have 0.34 NMSE, 0.32 R, -0.04 FB, 0.87 MG, 1.89 VG, 0.69 FAC2 accuracy levels. Figure 3.6 shows the generated prediction for the Linear Regression method. The result indicates that the approach creates more accurate hourly predictions than the standard Land Use Regression method however it still produces less accurate model than the state-of-the-art OSPM air pollution dispersion model. The prediction-observation (Figure

3.6) graph shows a very wide cloud shape which indicates that the linear regression algorithm struggles to learn the non-linear (e.g. concentration level and hours of the day) relationship between the input and prediction target data. This result is in line with the previous studies findings [Champendal et al. (2014); Sánchez et al. (2011)] as researchers reported that this algorithm fails to make accurate predictions on the hourly levels due to its weak ability to learn non-linear nature of the given regression problem.

## 3.4 Advanced statistical regression approaches

The Linear Regression algorithm provides a simple and elegant solution to discover the linear relationship between the input and the regression target data. Extending the traditionally used land use data with the high-temporal time and weather-related data generates a dataset which contains non-linear relations to the concentration levels, therefore the Linear Regression method struggles to make accurate hourly concentration level predictions (as suggested by [Briggs et al. (2000)] and discovered in the previous section as Figure 3.6 indicates poor predictions quality and the models produce high RMSE error levels). Having established that the existing Land Use Regression approaches fail to generate accurate hourly $NO_2$ concentration level predictions, the application of advanced machine learning regression algorithms is investigated further in the rest of this chapter. Studies [Champendal et al. (2014); Sánchez et al. (2011)] are suggesting that similar environmental problems can be solved with better prediction accuracy using other methods (e.g. Neural Network Regression or Decision Tree Regression methods, Random Forest Regression) than the standard Linear Regression method. Other methods (e.g. Nearest Neighbour Regression, Support Vector Regression) were successfully applied to non-linear regression tasks in the past [Gardner & Dorling (1999); Sánchez et al. (2011); Tso & Yau (2007); Champendal et al. (2014)]. This section investigates the potential prediction accuracy level of the advanced regression techniques including the Nearest Neighbour Regression, Neural Network Regression, Support Vector Regression, Decision Tree Regression and the Random Forest Regression method. All of these methods require a certain level of tuning depending on their hyperparameter requirements (which makes them harder to use compared to the Linear Regression method which is a hyperparameter-free method). The methods are highly sensitive to their hyperparameters which parameters control the process to build their inner structure in the phase of training (generating) the regression model. These methods are only capable of extracting the relationship between the input data and prediction target effectively if they are used with their tuned hyperparameters. One method to find out these optimal configurations is the execute a grid-type hyperparameter search which helps to understand the achievable accuracy level for each hyperparameter settings within a range. Using this method, a wide grid-type hyperparameter search was executed to understand the sensitivity of the algorithms and the best hyperparameter for each individual methods was selected to tune these algorithms to achieve the best possible prediction accuracy level. This optimization helps to determine the prediction accuracy level boundary for

each method by analysing the high-level RMSE error level trends according to the hyperparameter search. Optimizing the algorithms to produce the lowest RMSE error level helps to reduce the concentration level prediction errors in the situations of high observed concentration levels. The RMSE penalizes the large differences of the observation and prediction concentration levels, therefore, optimizing the methods to the lowest RMSE levels results in accurate predictions in the case of high observation levels which helps to identify the interesting pollution episodes as these episodes are the interest for urban planners and researchers. This optimization, however, gives hyperparameters only valid for the applied input dataset (e.g. the dataset which feeds the leave one out cross validation framework). Using a different dataset from the same domain (e.g. applying these techniques to a different modelling area) or from a different domain (e.g. a completely different regression task) require another execution of this optimization. This other execution gives valid hyperparameters to those other problems.

### 3.4.1 Nearest Neighbour Regression

The scikit-learn implementation of the Nearest Neighbour Regression has two hyperparameters:

- the number of the nearest neighbours to calculate the prediction
- the power (p) parameter for the Minkowski distance calculation

The generation of the regression model is sensitive to these hyperparameters and it is not clear what is the optimal configuration to use for this specific regression task. The grid hyperparameter search was configured to find the RMSE prediction accuracy for the neighbour and the p parameters between 1 and 100 and 1 and 5, respectively. Figure 3.7 shows the result of this search. Each p parameter reaches its prediction accuracy minimum at a certain point within the given neighbour parameter range, therefore, it is not possible to reach more accurate predictions using any other parameter combination. The method is depending on its hyperparameter as the different models generated by different hyperparameter configuration produces predictions with the accuracy range of 21.5 and 20.2 $\mu gm^{-3}$ RMSE error. The method gives its best RMSE prediction accuracy using the neighbour=23 and p=2 providing 20.2 $\mu gm^{-3}$ RMSE and 15.67 $\mu gm^{-3}$ MAE, 0.41 NMSE, 0.26 R, 0.04 FB, 0.94 MG, 1.79 VG, 0.68 FAC2 error levels. This is indicating that the method cannot provide more accurate model than the state-of-the-art OSPM air pollution dispersion nor the Linear Regression statistical regression method.

### 3.4.2 Neural Network Regression

The scikit-learn implementation of the Neural Network Regression algorithm has a flexible way to construct and train the internal neural network structure by providing the following hyperparameters:

- number of hidden layers and the neurons in each hidden layer
- train iteration

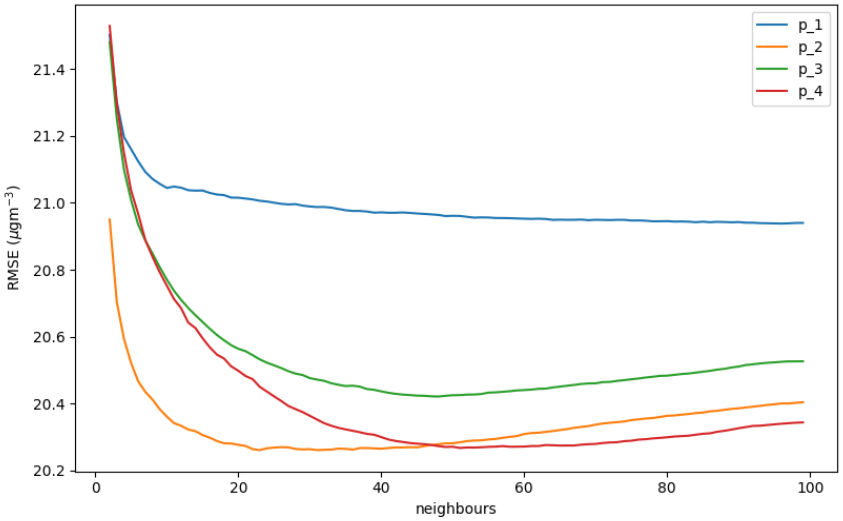



Figure 3.7. Hyperparameter investigation for Nearest Neighbour Regression method

- learning rate (alpha)

To train the Neural Network Regression model, the data has been normalized as suggested by [Pedregosa et al. (2011)]. This transformation of the data helps the algorithm to avoid numerical instability during the training phase.

It is clear that this algorithm also depends on the listed hyperparameters but it is not known what hyperparameter configuration gives the best model (considering the model's prediction accuracy) to this prediction task. The grid hyperparameter search was configured to investigate the high-level accuracy of the model using a different number of hidden layer configurations (from 1 to 5 hidden layers using sigmoid type neurons) with different neurons in each layer (from 5 to 500 neurons) with different train iterations (from 5 to 15) and different learning rates (from 0.00001 to 0.01).

Training the neural network regression model was able to produce numerically stable result using the 0.00001 learning rate as setting the learning rate greater than this value made the training phase unstable and training the input weights of the neurons high ending up a model predicts extremely high concentration levels independently from the input data. Also, applying more than 1 hidden layer generated the same numerical instability. Figure 3.8 shows high-level RMSE error level depending on the number of train iterations and the number of the neurons in the hidden layer. This indicates that increasing the number of train iterations helps to increase the prediction accuracy, however, this rate is minor. Furthermore, increasing the number of neurons in the hidden layer increases the prediction accuracy up to the 200 neurons where the model reaches its

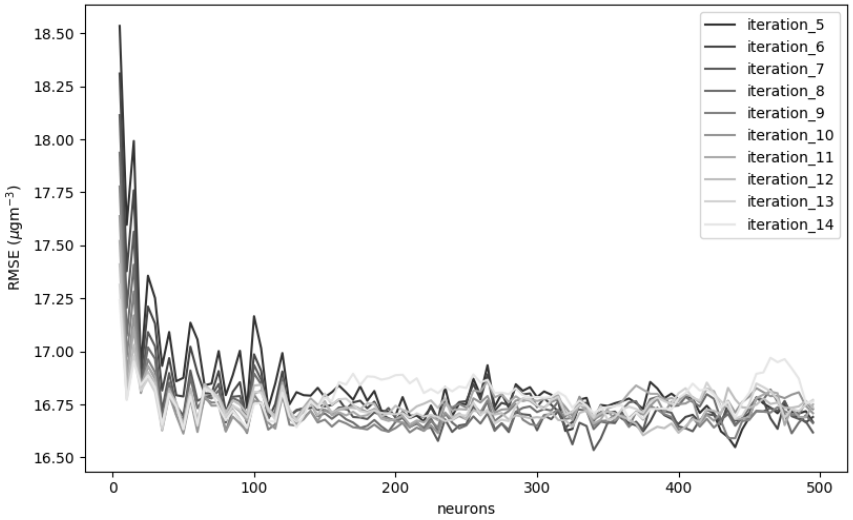



Figure 3.8. Hyperparameter investigation for Neural Network Regression method

most accurate state.

In summary, the algorithm is depending on the hyperparameters as the figure shows that the RMSE high-level accuracy varies between 18.5 and 16.57 $\mu gm^{-3}$ RMSE levels. Using the 230 neurons and 7 iterations hyperparameters generated a neural network regression model with 16.57 $\mu gm^{-3}$ RMSE and 12.95 $\mu gm^{-3}$ MAE, 0.26 NMSE, 0.49 R, 0.00 FB, 0.86 MG, 1.52 VG, 0.78 FAC2 high-level prediction accuracy.

### 3.4.3 Support Vector Regression

The scikit-learn library implements the epsilon Support Vector Regression algorithm which has the epsilon ($\epsilon$) and the error penalty ($C$) hyperparameters. The algorithm has very high computational requirements, therefore, the suggested method to apply this algorithm to a large regression task (such as the hourly concentration level prediction) is to use bagging where the training data is sampled $n$ times and $n$ models are built (then the average of the output of the $n$ models is used to generate the combined prediction). The bagging method, therefore, requires two additional hyperparameters: the number of the models and the sample rate. The sensitivity of this algorithm depends on these hyperparameters as they control the generation the underlying model. It is not clear that what hyperparameter configuration produces the most accurate Support Vector Regression model for this regression task, therefore, a grid-type search was executed to find the optimal configuration for the algorithms hyperparameters.

Before executing the search, the input and target data have been transformed into the required normalized form as the method requires normalized input data ([Pedregosa et al. (2011)]).

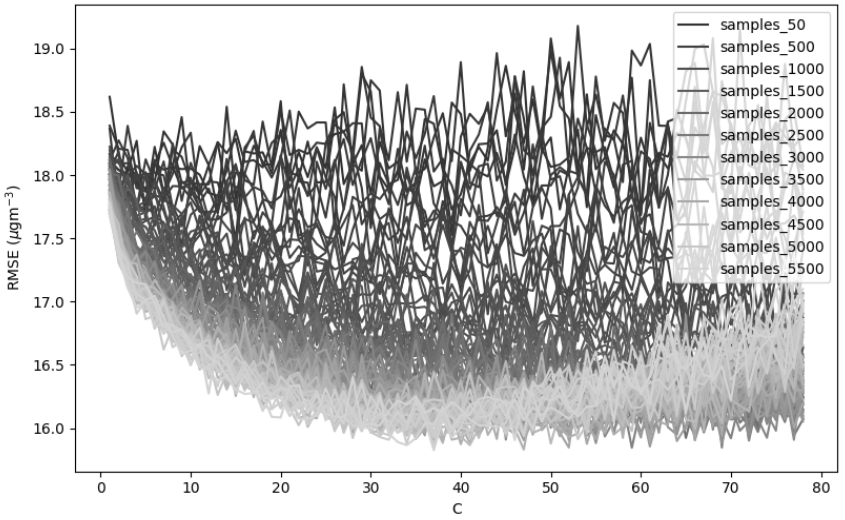



Figure 3.9. Hyperparameter investigation for Support Vector Regression method

The hyperparameter search was configured to calculate the accuracy level for the following hyperparameter ranges:

- Epsilon from 0.0001 to 1.0
- C from 1 to 100
- Number of models from 5 to 100
- Sample rate from 50 to 5000

The hyperparameter search provided sufficient understanding of the algorithm's prediction behaviour. Overall, the epsilon hyperparameter had very little effect on the prediction accuracy, therefore, the default 0.1 value was selected. Also varying the number of models had very little effect on the high-level RMSE error levels, therefore, the default value of 10 was selected. Figure 3.9 shows the hyperparameter tuning for the $C$ and sample_rate. It confirms that the algorithm is sensitive to its hyperparameters as the RMSE level varies between 19.1 and 15.9 $\mu gm^{-3}$ RMSE. The figure shows that the algorithm is unstable as changing the hyperparameters doesn't cause discrete increase or decrease in the high-level RMSE error level. There are two clear trends from this figure:

- increasing the sample size results in more accurate predictions as this statistical regression method has the chance to discover knowledge from more data

- the algorithm reaches its prediction optima at C=42 which suggest that this is the best accuracy level that the algorithm can reach

Increasing the sample size increases the prediction accuracy, however, this has an exponential computational cost as the Support Vector Regression model doesn't scale well with the input data size (the algorithm time complexity is quadratic to the number of input observation data points). The sample size hyperparameter search was limited using the 5000 upper boundary value as increasing this parameter caused the cross-validation framework to finish the 7 iterations in 4 hours. Minor improvement can be achieved by increasing the sample rate further however it produces a computationally expensive statistical regression model.

In summary, this method gives its best prediction using n=10, epsilon=0.1, C=40 sample_rate=4200 providing 15.93 $\mu gm^{-3}$ RMSE and 12.25 $\mu gm^{-3}$ MAE, 0.24 NMSE, 0.55 R, 0.00 FB, 0.88 MG, 1.52 VG, 0.80 FAC2 error levels on this regression task.

### 3.4.4 Decision Tree Regression

The scikit-learn framework implementation of the Decision Tree Regression algorithm provides multiple tree-induction termination methods:

- *depth* method which only grows a tree to a certain depth
- *minleaf* method which grows the tree's branches until the leaf node has at least the given min_leaf number of observations
- *maxleaf* method which grows the tree until the number of leaf nodes in the tree reaches the given max_leaf parameter

The scikit-learn version of the Decision Tree Regression method optimizes the mean squared error achievable by the decision tree regression model on the training data during the search for the split in each iteration.

These methods generate different decision trees as they terminate the induction process differently. This termination process helps the model to avoid overfitting and increases the prediction accuracy achievable by the model itself. It is not clear however which method can produce the best decision tree in the terms of this regression task.

The method depends on its hyperparameters as they define how to build the internal decision tree and when to terminate the induction of this tree, therefore, the method is sensitive to its hyperparameters. A grid hyperparameter search was executed to find the optimal configuration of these hyperparameters. The search was configured for each method to investigate the parameters from value 2 to value 100.

Figure 3.10 shows the result of this investigation. It confirms that the method is sensitive to its hyperparameters as the RMSE high-level accuracy varies between 18.4 and 16.18 $\mu gm^{-3}$. The accuracy of the *depth* method flats out around depth=42 as the decision tree reaches its maximum

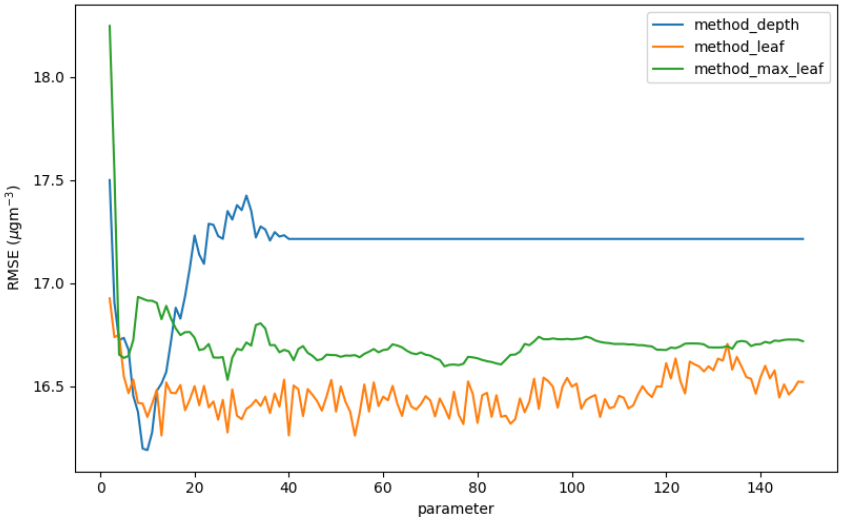



Figure 3.10. Hyperparameter investigation for the Decision Tree Regression method using its three (*depth*, *minleaf*, *maxleaf*) tree induction techniques

depth at this level (each leaf has 1 observation in this case). The other two methods (*minleaf* and *maxleaf*) show a flat RMSE accuracy level after parameter=30 configuration. The figure shows that the best RMSE accuracy level can be achieved by using the *depth* method configured to depth=12 which provides 16.18 $\mu gm^{-3}$ RMSE and 12.30 $\mu gm^{-3}$ MAE, 0.25 NMSE, 0.58 R, 0.01 FB, 0.95 MG, 1.49 VG, 0.79 FAC2 high-level errors.

### 3.4.5 Random Forest Regression

The scikit-learn framework implementation of the Random Forest Regression method provides one additional parameter to the underlying Decision Tree Regression method's hyperparameters: the number of the decision tree models to train based on the random sampling of the input data (this parameter called the "estimator" in the framework).

The method depends on its hyperparameters as they define the technique to build the internal tree structures for the trees. It is not clear what hyperparameter configuration produces the most accurate Random Forest Regression model on this regression task. Grid hyperparameter searches were executed to find the optimal configurations for each method.

For the *depth* method, the depth parameter search range was set from 2 to 50 and the estimator parameter range from 5 to 200. Figure 3.11 shows the result of the investigation. The *depth* tree induction method of the Random Forest Regression algorithm shows similar behaviour than the Decision Tree Regression's one as the accuracy level flats out after the depth parameter of 35. However, increasing the number of estimators helps to build more accurate regression model as

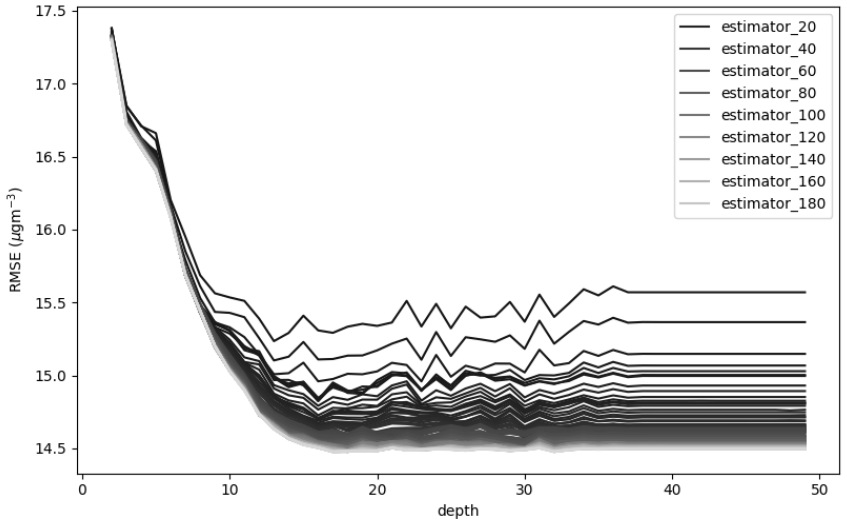



Figure 3.11. Hyperparameter investigation for the Random Forest Regression method using the *depth* tree induction technique

the algorithm has more chance to extract knowledge from the randomly presented sampled input data. From this run, it is not clear however that what trend this increase does follow. To analyse this trend, a second investigation was executed to find out that what is the trend of the accuracy level if we increase the number of estimators for the Random Forest Regression's *depth* method. The search was configured to only investigate depth levels 10,15,20,25,30, but with increased estimator range (from 5 to 500).

Figure 3.12 shows the result of the second investigation run for the *depth* method which shows that increasing the number estimators indeed improves the high-level RMSE accuracy level, however after 200 estimators the accuracy level flats out again.

In summary, the *depth* method gives its most accurate prediction using the depth=25 and estimators=400 which produces 14.47 $\mu gm^{-3}$ RMSE high-level error. It is possible to further increase this accuracy level with some minor improvement however the computational cost of this improvement makes it non-practical.

To find out the best hyperparameters for the *maxleaf* method, the hyperparameter search was configured to max_leaf parameter range from 5 to 7000 and the number of estimators parameter range from 5 to 20. Figure 3.13 shows the result of this parameter search run. The high-level RMSE accuracy flats out at parameter max_leaf 5000, however, the increasing number of estimators provides more accurate overall models (similarly to the previous *depth* method).

To understand the high-level accuracy trend of the estimators parameter for the *maxleaf* method, the grid hyperparameter search was configured for a second run using only the 5000,

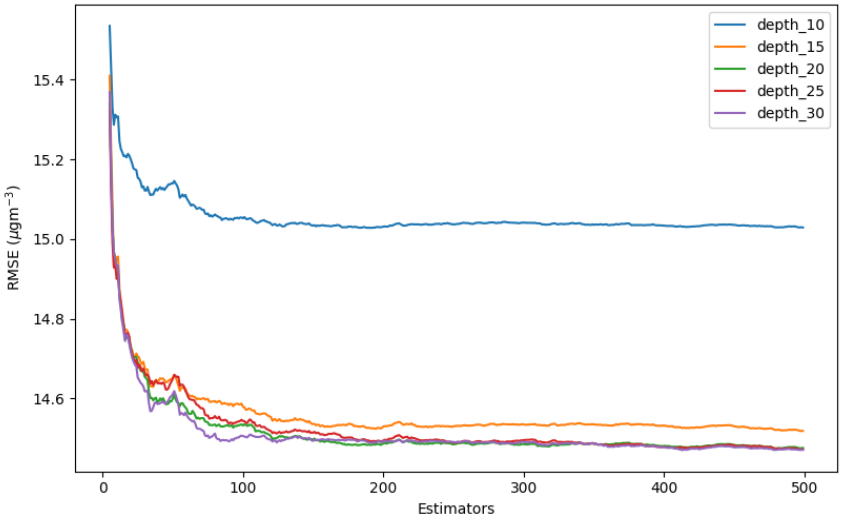



Figure 3.12. Hyperparameter investigation for the Random Forest Regression method using the *depth* tree induction technique

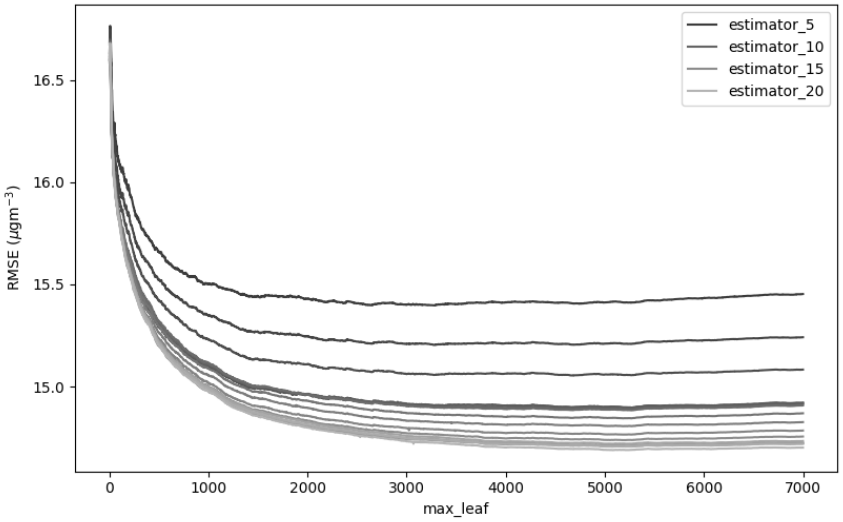



Figure 3.13. Hyperparameter investigation for the Random Forest Regression method using the *maxleaf* tree induction technique

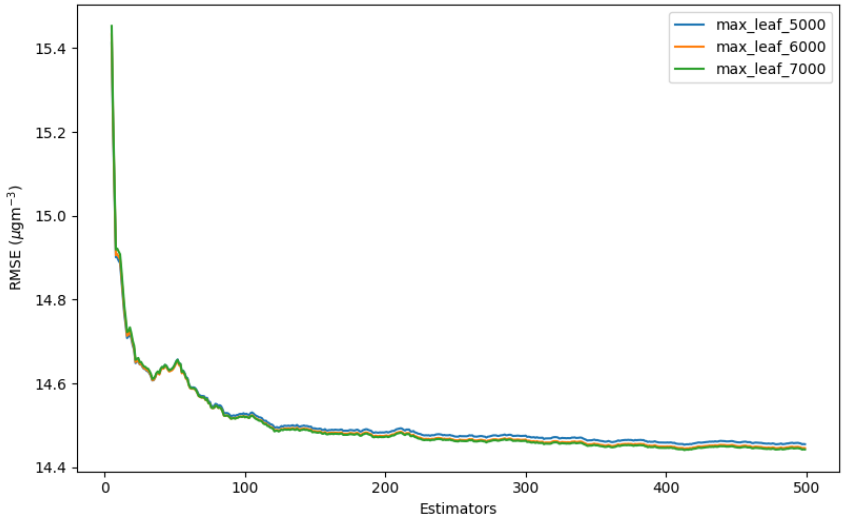



Figure 3.14. Hyperparameter investigation for the Random Forest Regression method using the *maxleaf* tree induction technique

6000, 7000 max_leaf parameter and the range from 5 to 500 for the estimators parameter. Figure 3.14 shows the result of the second hyperparameter search run. There is not much high-level RMSE accuracy difference in terms of using the 5000,6000,7000 max_leaf parameters, however increasing the number of estimators increases the accuracy up until the 400 estimators where the accuracy level flats out.

In summary, the *maxleaf* method gives the most accurate using the maxleaf=7000 and estimators=400 which model generates 14.47 $\mu gm^{-3}$ RMSE high-level error. Again, this accuracy level can be further improved by increasing the estimators however the improvement will imply very high computational cost.

Lastly, the parameter search was configured to find out the best hyperparameters to achieve to best high-level RMSE accuracy level for the *minleaf*. The minleaf parameter was set to range from 2 to 200 and the number of estimators from 5 to 200. Figure 3.13 shows the result of this hyperparameter search run. This result shows that using the *minleaf* method generates the most accurate (in term of the high-level RMSE accuracy) model at the minleaf=2 parameter (independently from the number of estimators). This means that it doesn't stop to generate the tree nodes until each leaf node only has 2 remaining observations. This allows the decison tree induction method to generate large trees capable of prediction the concentration levels accurately. Moreover, increasing the number of estimators increases the high-level RMSE accuracy however it is not clear that what is the optimal number of estimators to use for the *minleaf* method.

Another hyperparameter search was executed to find out this number using only the 2, 3, 4

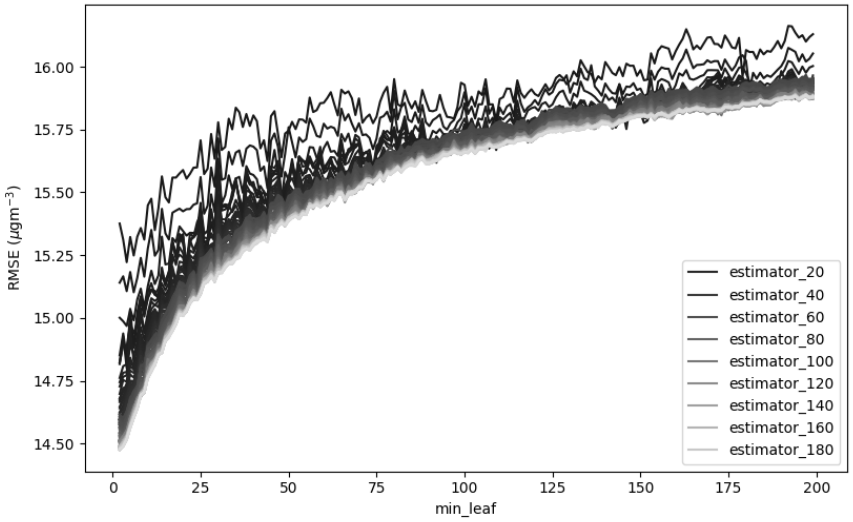



Figure 3.15. Hyperparameter investigation for the Random Forest Regression method using the *minleaf* tree induction technique

values for min_leaf parameter and setting the range from 5 to 1000 for the number of estimators parameter. Figure 3.16 shows the result of this run. It confirms that increasing the number of estimators increases the high-level accuracy, however, the accuracy flats out at 500 estimators as using more than 500 estimators does not give further improvement in the prediction accuracy.

In summary, the *minleaf* method gives its most accurate predictions using the minleaf=2 and estimators=600 parameters which model generates the prediction with 14.45 $\mu gm^{-3}$ RMSE high-level error.

The hyperparameter searches confirm that the method is sensitive to its hyperparameters as the RMSE high-level accuracy varies between 17.5 and 14.45 $\mu gm^{-3}$

The best hyperparameters for each decision tree induction method was selected based on the introduced hyperparameter search runs:

- the *depth* method using depth=25 and estimators=400 gives 14.47 $\mu gm^{-3}$ RMSE accuracy level
- the *maxleaf* method using maxleaf=7000 and estimators=400 gives 14.47 $\mu gm^{-3}$ RMSE accuracy level
- the *minleaf* method using minleaf=2 and estimators=600 gives 14.45 $\mu gm^{-3}$ RMSE accuracy level

This result suggests that *minleaf* method provides the most accurate hourly concentration level predictions within the many tree induction methods of the Random Forest Regression al-

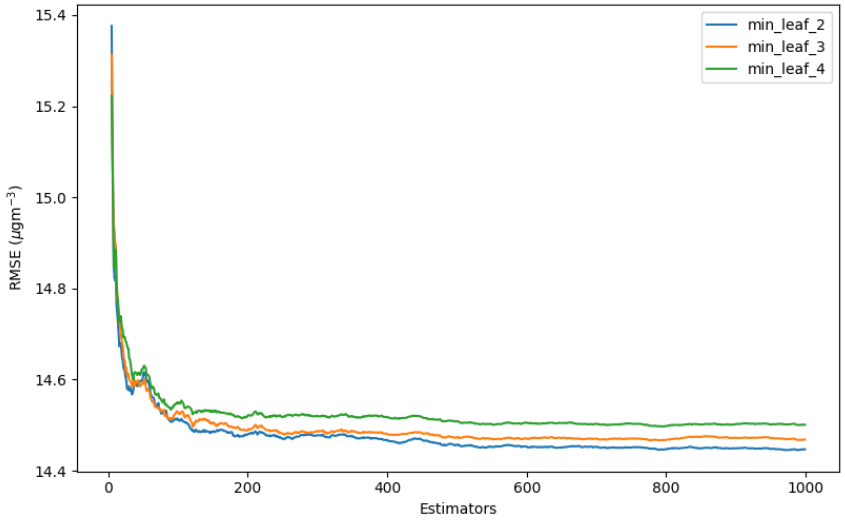


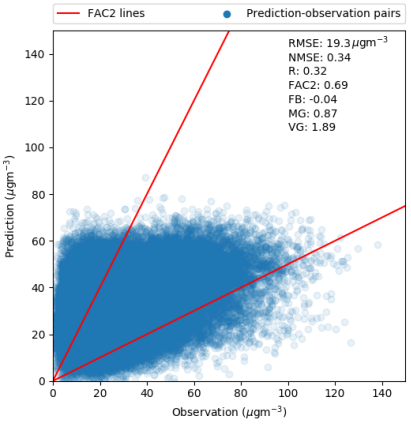

Figure 3.16. Hyperparameter investigation for the Random Forest Regression method using the *minleaf* tree induction technique

gorithm. This method could exploit the prediction power of more estimators (as it has its peak accuracy using 600 estimators instead of the other two methods 400 estimators) and the accuracy flats out at that level. Therefore, the *minleaf* tree induction method was selected for the Random Forest Regression algorithm which could provide a regression model with 10.75 $\mu gm^{-3}$ MAE and 14.45 $\mu gm^{-3}$ RMSE, 0.20 NMSE, 0.67 R, 0.03 FB, 0.97 MG, 1.40 VG, 0.83 FAC2 high-level errors.

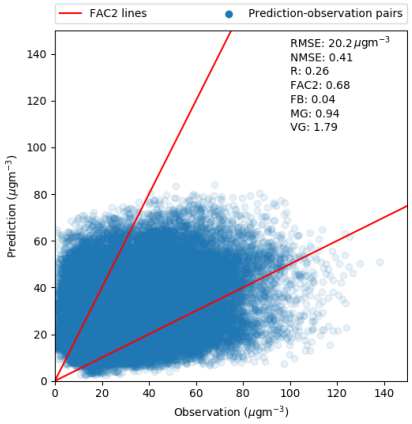
## 3.5 Evaluation and discussion

Finding the best hyperparameters for each statistical regression method gives us well-tuned algorithms to generate hourly concentration level prediction with the minimum achievable RMSE high-level accuracy levels. All the used methods have different sensitivity to the hyperparameters, therefore, the most accurate (lowest RMSE level) hyperparameter settings were selected for each algorithm. This also implies that MAE levels are close to the minimum (however it might happen that there is a very slight hyperparameter difference in the models which have the minimum achievable RMSE and MAE levels, but the overall prediction levels are going to be very close therefore it does not have any effect on this evaluation). The RMSE high-level accuracy level does not provide information about the quality of the predictions. Figure 3.17 shows all the observation-prediction pairs for each method which gives us more understanding of the individual algorithms.

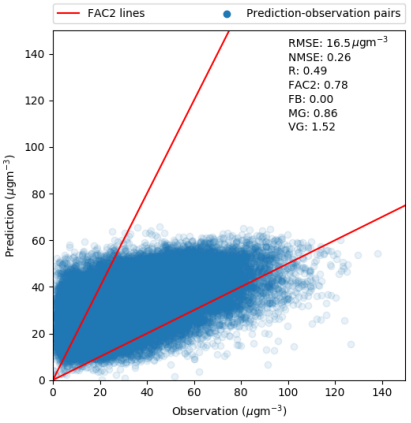
The Linear Regression algorithm struggles to make accurate concentration level predictions



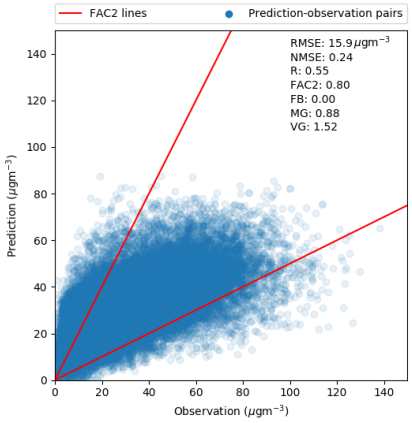

(a) Linear Regression model



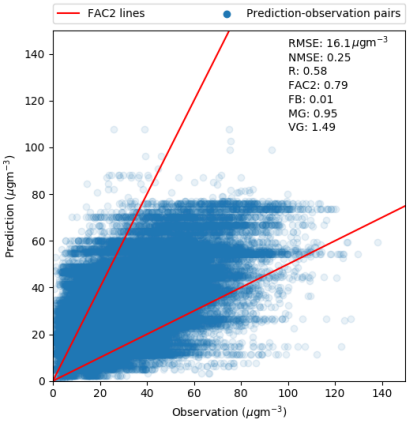

(b) Nearest Neighbour Regression model



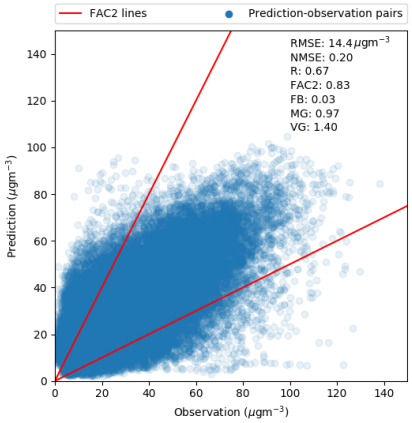

(c) Neural Network Regression model




(d) Support Vector Regression model




(e) Decision Tree Regression model




(f) Random Forest Regression model

Figure 3.17. Hourly prediction and observation scatter graphs for the statistical regression methods

greater than 70 $\mu gm^{-3}$. The main reason for this behaviour is that the algorithm itself fails to identify the non-linear relationship between the input and the prediction target data. Even though the method produces predictions with low accuracy, it is classified as good model according the [Chang & Hanna (2004)] because the FAC2, MG and VG levels are within the criteria range.

The Nearest Neighbour Regression algorithm shows even worse prediction-observation diagram as the shape of the prediction-observation pairs covers a wider area. The driving reason for this behaviour is that it is hard to make accurate predictions based on similarity of the historical observations as very similar observations can have very different observation concentration levels (e.g. concentration levels can accumulate at the observation stations prior to the observation hour depending on the weather circumstances of the prior hours). All the high-level accuracy measures show weaker prediction quality compared to the Linear Regression model, however, this result still classified as good model according to [Chang & Hanna (2004)].

The Neural Network Regression algorithm shows similar behaviour to the Linear Regression model as it fails to make accurate hourly concentration level predictions at high concentration level observations. This indicates that the algorithm fails to identify the non-linear relationship in the data even though it has a much more complex internal structure (which structure gives this algorithm the capability to discover complex relationship between the input and target data). The high-level error levels show that the method can provide more accurate predictions than the OSPM air pollution dispersion model, however, the low linear correlation coefficient value shows that the generated predictions are weakly correlating with the actual observations (OSPM: 0.69, Neural Network Regression: 0.49), however the NMSE level is better (OSPM: 0.53, Neural Network Regression: 0.26) which indicates that the method managed to decrease the normalized prediction error.

The Support Vector Regression algorithm provided the most accurate result non-tree based regression techniques. The method provides even lower high-level RMSE and MAE error levels (compared to the previous methods) which is in line with its observation-prediction plot where we can see that the model generates more accurate predictions at higher observation levels. It, however, struggles to make predictions with the same correlation level as the state-of-the-art air pollution dispersion pollution model (OSPM: 0.69, Support Vector Regression: 0.55).

The Decision Tree Regression algorithm provides concentration level predictions with high accuracy as it produces 16.18 $\mu gm^{-3}$ RMSE and 12.3 $\mu gm^{-3}$ MAE levels. The observation-prediction reveals the nature of the algorithm's predictions. The observation-prediction pairs are showing smaller pollution dispersion, however, it shows some flat prediction values for certain observations which indicates that the regression decision tree reached its limitation and cannot provide more detailed predictions in these cases. This also effects the linear correlation as it has even lower level than the Support Vector Regression's level (OSPM: 0.69, Decision Tree Regression: 0.58).

The Random Forest Regression algorithm provided the most accurate model from all the

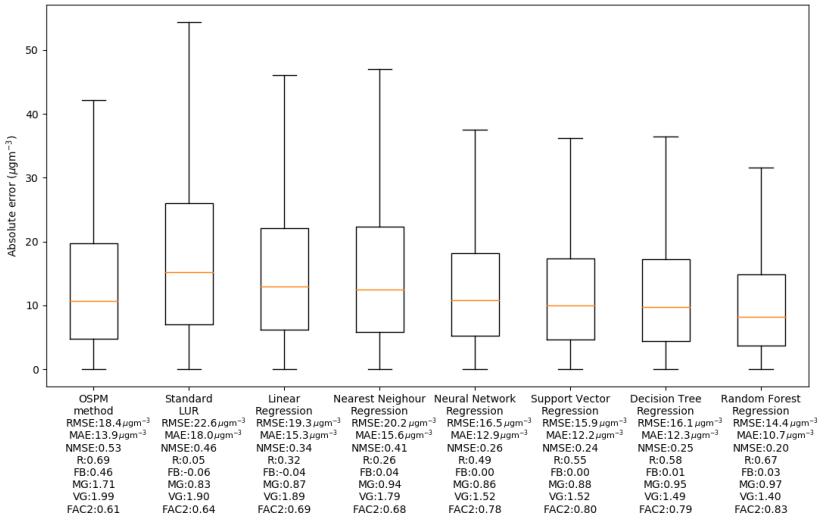



Figure 3.18. Absolute error of the hourly concentration level predictions for all the investigated methods (red line shows the median of the absolute prediction errors)

investigated methods as it produced predictions with 14.45 $\mu gm^{-3}$ RMSE and 10.75 $\mu gm^{-3}$ MAE levels (also including the most accurate NMSE, FB, MG, VG and FAC2 levels). It has the smallest dispersion in the observation-prediction plot and it does not show the Decision Tree Regression methods limitations as the high number of trees could produce very detailed concentration level predictions at all observation levels. The model provided predictions with almost the same linear correlation level as the OSPM air pollution dispersion model (OSPM: 0.69, Random Forest Regression: 0.67) which makes this statistical regression model as good as the current state-of-the-art air pollution dispersion model in terms of hourly $NO_2$ concentration level predictions.

Looking at the observation-prediction chart helped to understand the statistical regression models prediction behaviour however it did not provide a well-structured comparison between prediction accuracy of the methods. To do that, the absolute error of the observation-prediction pairs for each method ware plotted. Figure 3.18 shows the comparison of the absolute error box plot of the predictions for each method. This result of this graph is in line with the Figure 3.17 as it shows that the most accurate statistical regression model (the Random Forest Regression statistical regression method) produces more accurate hourly $NO_2$ concentration level predictions than the OSPM state-of-the-art air pollution dispersion model.

## 3.6 Summary

The aim of this chapter is to develop a statistical regression approach for hourly $NO_2$ concentration level prediction providing the same high-level accuracy as the state-of-the-art air pollution dispersion models.

As the baseline model application evaluation indicates, the OSPM air pollution dispersion model produces 18.49 $\mu gm^{-3}$ RMSE high-level accuracy using York as the modelling area. The chapter discusses the application of this model including the validation of the prediction results. Orthogonally to the state-of-the-art method, the existing statistical approaches provide 22.65 $\mu gm^{-3}$ (standard LUR technique) and 19.39 $\mu gm^{-3}$ (Linear Regression method using the combination of low- and high-temporal input data) RMSE high-level accuracy. The chapter investigates the result in details which result is in line with the outcome of previous studies of the relevant literature [Briggs et al. (2000); Champendal et al. (2014); Sánchez et al. (2011)].

Using the sufficiently tuned Random Forest Regression technique, however, provides 14.45 $\mu gm^{-3}$ RMSE accuracy which indicates that this statistical regression approach can reach even more accurate prediction level than the current state-of-the-art method without using uncertain data (e.g. emission inventory database). The chapter describes the hyperparameter tuning details for this and many other methods which indicate that it is required to analyse the hyperparameters for these methods as the accuracy is sensitive to the configured hyperparameters. This analysis, however, is only valid for the York modelling area (which is represented by the York dataset). In the case of a different modelling area (e.g. a dataset covers a different area), the hyperparameter search needs to be re-executed to find out the hyperparameter configuration which sets the models to generate the most accurate predictions. The result of this chapter is a contribution to the Environmental Science field as it provides details of the application of the existing Random Forest Regression technique to the urban-scale hourly $NO_2$ concentration level predictions which model is able to generate predictions with the same high-level accuracy as the current state-of-the-art. The result indicates that it is possible to generate more accurate hourly $NO_2$ concentration levels using the Land Use Regression approach by applying the Random Forest Regression algorithm.

The Random Forest Regression technique, however, builds the underlying statistical regression model based on historical observations (both concentration level and other input data) which raises the question how the actual algorithm uses the input data to make the hourly $NO_2$ concentration level predictions and what data is introducing what type of error during the generation of the predictions. The next chapter will investigate the different errors that introduced by the different input data to give more understanding of the model's prediction and in theory to allow to develop even more accurate statistical regression model.

# CHAPTER 4

# Analysis and optimization of the Statistical Regression approach

This chapter presents the detailed analysis of the application of the Random Forest statistical regression method for hourly $NO_2$ concentration level predictions and introduces a novel approach to exploit the knowledge extracted from the analysis to improve the accuracy of the statistical regression approach. The chapter begins with analysing the accuracy sensitivity of the applied data for building the Random Forest Regression method to understand what data (or data source) introduces error to the concentration level predictions. The gained knowledge from this analysis contributes to the Environmental Science field as the analysis provides a guideline for data collection for applying the Random Forest Regression method for future applications. The second part of this chapter analysis the prediction outcome of different Random Forest Regression models trained on different subsets of the available features of the original input data. Based on the insight gained from this analysis, a novel ensemble method is proposed which ensemble method contributes to the Computer Science field as the algorithm forms a general ensemble method which can be used in any other regression task.

In the first section (Section 4.1), the motivation of this work is explained which introduces the aim of the initial analysis. Section 4.2 describes the analysis and discusses the results. Based on the findings of the second section, Section 4.3 carries out a new application using a new traffic dataset and the results of the evaluation will be described there. The findings of this experiment open the possibility of model ensembling to combine the prediction of the different Random Forest Regression algorithms; this will be described in Section 4.4. Finally, the Section 4.5 finalizes the chapter.

## 4.1 Motivation

The previous chapter provided the details of the efficient application of a statistical regression approach for hourly $NO_2$ concentration level predictions. The evaluation of the application showed that the proposed method can achieve more accurate predictions than the current state-of-the-art air pollution dispersion model. The underlying model, the Random Forest Regression algorithm needs to be trained on historical observations which were covered by data collected and extracted for the modelling area, York. The quality of the statistical regression model therefore highly depends on the input data itself. The data collection was based on the input data appeared in the literature and the model training used all the available data, however, this data itself can contain errors and uncertainties (e.g. the digital map source used to extract land use features can contain old information or the acquired meteorological data describes average weather conditions in York which is the same as the conditions at the location of each monitoring station). It is not known how these errors and uncertainties affect the quality of the statistical regression model (and affect the accuracy of the prediction generated by the model).

The aims of the work presented in this chapter are

- to understand the effect of using data from different data sources to the prediction accuracy generated by the Random Forest Regression algorithm
- exploit the knowledge extracted during the analysis to develop a model generating predictions with higher accuracy

The first aim plans to give an understanding of the consequence of using data from different data sources and extract knowledge on the Random Forest Regression sensitivity to the different data sources (what data is important to make accurate predictions and what data is useful to this statistical regression approach).

The second aim is targeting to exploit the knowledge gathered during this analysis to create a statistical regression model which generates more accurate prediction than the already developed statistical regression approach.

## 4.2 Input data analysis for the statistical regression method

The statistical regression approach can provide a similar accuracy level to that obtained by the current state-of-the-art air pollution dispersion model for the hourly $NO_2$ concentration level predictions using the combination of high-temporal input data and the Random Forest Regression algorithm. This method, however, requires historical observations to learn the hidden relationship in the data. This implies that the underlying statistical regression model depends on the given input data. It is unknown that what data source is the most beneficial to the hourly $NO_2$ prediction task given the Random Forest algorithm. One way to evaluate the achievable prediction accuracy of using data from the different data sources is to execute a feature analysis using

groups of coherent features.

### 4.2.1 Feature analysis of the Random Forest method

The following data was used during the application of the statistical regression methods (including the Random Forest Regression algorithm):

- Land use data (group code: *L*): this data source provided the landuse_area and the leisure_area features
- Building data (group code: *B*): this covers the buildings and building_area features
- Road data (group code: *R*): this contains the road_length and road_lane_length features
- Traffic data (group code: *V*): this includes the traffic_car, traffic_lgv, traffic_hgv features
- Time-related data (group code: *T*): this data source provided the hour, day_of_week, month, bank_holiday, race_day features
- Weather data (group code: *W*): this covers the wind_direction, wind_speed, temperature, rain (indicator), pressure features

It is not clear that how these data used by the Random Forest algorithm to generate the internal decision trees during the training phase. It is possible to analyse the generated tree models inside the Random Forest model to analyse the prediction process of this algorithm, however, this analysis is practically unfeasible, because the model has 600 independent decision trees over 19 input features. Another way to evaluate the benefit of using data from different data sources to evaluate the accuracy of the Random Forest algorithm using all the subsets of the available data features. The overall input data has 6 data groups which give 63 possible data subsets. The high-level RMSE accuracy can be calculated by using the evaluation framework for every 63 combinations. This evaluation helps to understand

- what are the data sources to use to train the Random Forest Regression to achieve the most accurate hourly $NO_2$ prediction level (e.g. RMSE)
- what are the data sources that introducing errors into the prediction by having uncertain data causing less accurate predictions

Figure 4.1 shows the result of this experiment, where each data point has a label which label explains the selected data sources by indicating a 0 (data source has not been selected) or a 1 (data source has been selected) after the code of the data source. An example of this label is $L0B0R0V0W1T1$ where weather and time-related data was used to train the model. The most accurate predictions can be achieved by using only time and weather-related input data. The visualization of all the data subsets also shows a trend:

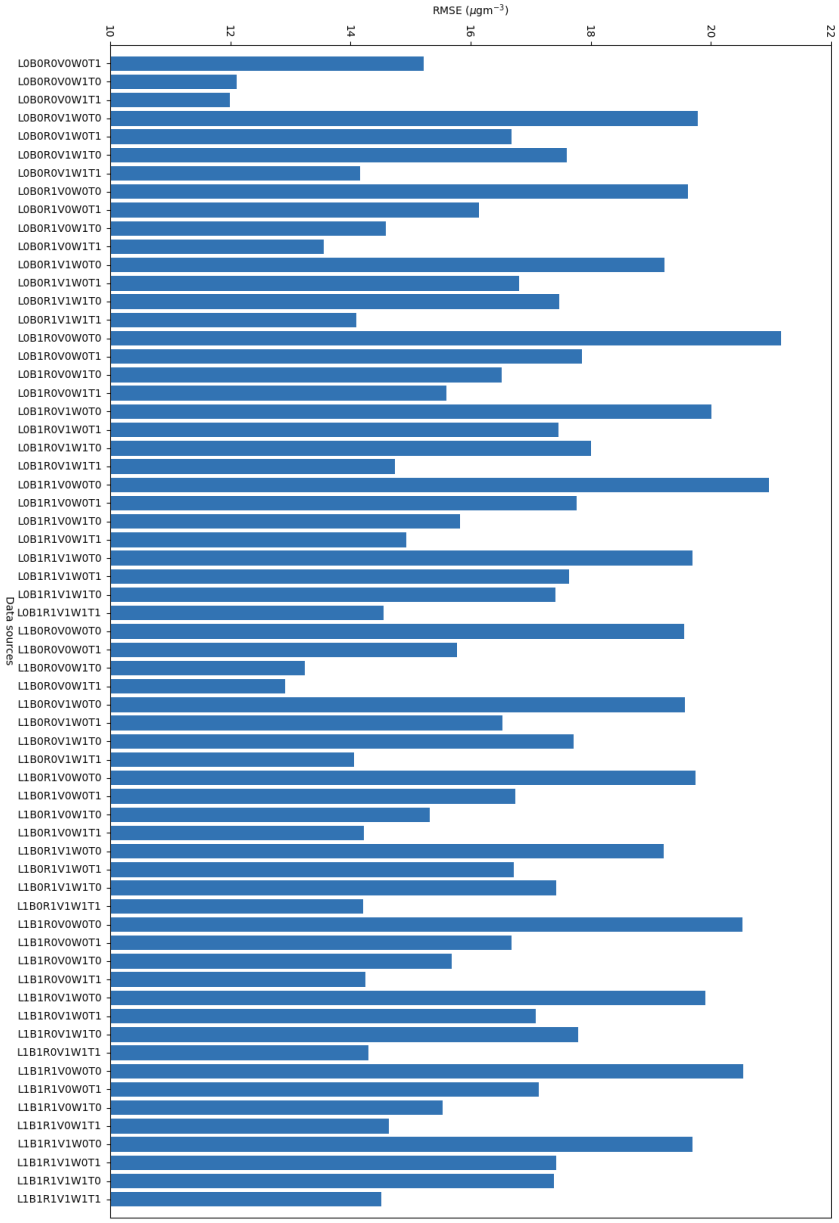



Figure 4.1. Accuracy investigation of the different input data subsets

- not using time and weather-related data always gives a model which produces less accurate predictions than
- using only time-related data, or
- using only weather-related data, or
- using both time and weather-related data

There is a periodic form in Figure 4.1 which also indicates this trend as the analysis of the shape of the figure using groups of four reveals the trend.

This result is important as it suggests that using only the high-temporal data (the time and weather-related data) will give an accurate statistical regression model. If similar model application is required (e.g. developing a similar $NO_2$ concentration level prediction model for another urban area) then collecting only time and weather-related data, as well as air quality data (the $NO_2$ concentration level observations), would be sufficient to develop an accurate statistical regression model which data sources are very easy to utilize, therefore, the model can be developed very quickly. This gives natural usability to this approach for developing initial models very quickly compared to the air pollution dispersion models where users have to collect data from various sources and investigate the uncertainty in these datasets (e.g. data in the emission inventory database related to the specific model application area). Unfortunately, using weather and time related data as the only input data also has the disadvantage of ignoring all the other important input data, therefore, urban planners are not able to investigate the effect of changing the urban environment. For example, if the urban planners want to investigate the effect of building a new school (which causes increased traffic on the surrounding roads) to the pollution concentration levels, the model would ignore the increased traffic data information and would produce the same pollution concentration level predictions to the base scenario.

The result makes sense in terms of the given regression task as these data are important for the actual $NO_2$ concentration levels:

- Weather data provides information about the wind and temperature conditions of an hour which have direct effect on the $NO_2$ concentration levels because the wind and rain can flush out the pollution from an area and certain temperature levels allow to form $NO_2$ from other gases
- Time-related data provide crucial information for the statistical model as certain time of the day has always higher concentration levels (e.g. school runs, afternoon traffic peak period, ) what patterns can be learned from the this data

To investigate this trend even further, the Figure 4.2 shows the box plot of the observed high-level RMSE accuracy levels of the evaluation of the Random Forest Regression method trained firstly on data subsets without the time and weather-related data (e.g. land use, building,

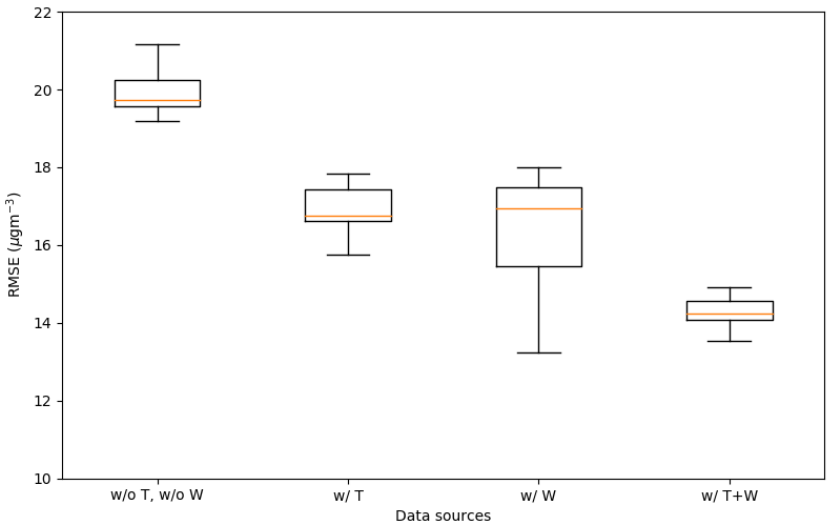



Figure 4.2. Prediction accuracy using the RFR method without Time and Weather data (w/o T, w/o W), using the Time data (w/ T), using the Weather data (w/ W) and using both the Time and Weather data (w/ T+W)

etc. data), then on data subsets containing only the time-related data (and other non time and weather-related data), thirdly on data subsets containing only the weather-related data and lastly on data containing both time and weather-related data (as well as all the other data sources in any combination). This plot indicates that the trend indeed exists and using time, weather and both time and weather-related data introduces more accurate statistical regression models.

The evaluation suggests that time and weather-related data are important for the given regression task, however, it is not known what error is introduced if we use further data from other data sources. Figure 4.3 shows the RMSE high-level accuracy for each combination of adding data from data sources excluding the time and weather data sources relative to the case of the Random Forest model using only the time and weather-related data. Again, the figure uses the same label to encode the additional data sources as earlier where the label $L1B1R0V0$ represents the input data which contains data from the land use and the building data sources additionally to the time and weather-related data. The figure shows that all the combination achieves greater than 1.0 relative RMSE level which suggests that using complete data groups does not provide more value to the Random Forest Regression model as it produces less accurate predictions using these additional data.

This evaluation was based on features from complete data sources (e.g. the land use data source provided two features which are the landuse_area and the leisure_area) which evaluation is good because it is possible to understand what data sources are important, however, it does not give a clear understanding of what individual features are important to the given regression

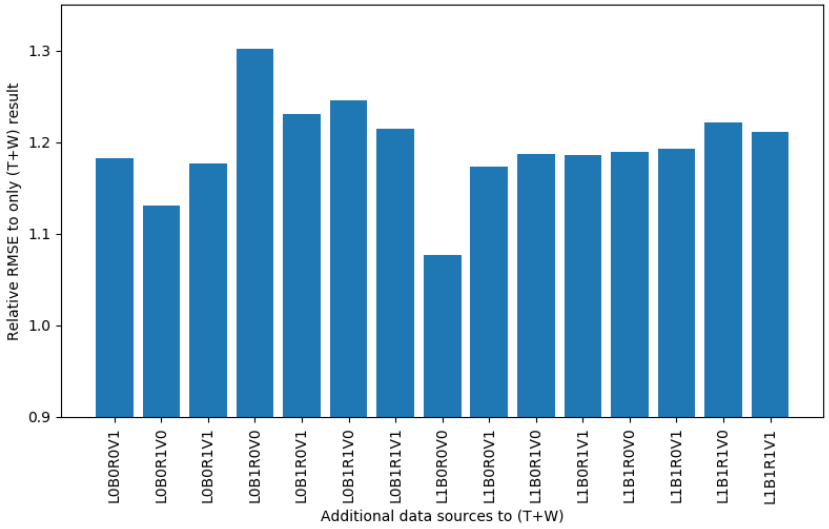



Figure 4.3. Relative RMSE accuracy using datasets compared to RFR method using only the Time and Weather data

problem. It is not clear that weather and time-related features are important and others only introduce errors, however, they introduce errors if they are given to the Random Forest Regression algorithm in groups. There is a possibility that these additional data feature groups does not help, but single individual features do (e.g. using the buildings_area with the time and weather-related data helps but because the evaluation used the building data source, the model had not just the buildings_area but the buildings feature which two features resulted in a less accurate model).

The number of possible combination of the data subsets using the 6 data sources gives 63 combinations. The number of possible combination of data subsets using the 19 individual features gives 524287 possible combinations. There are two problems to evaluate all of these 524287 possible combinations:

- the first problem is that the computational requirement for evaluating all the 524287 possible combinations would take unfeasibly long time as the evaluation of one combination takes approximately 10 minutes (which mean running all the evaluation on one machine would take approximately 10 years). Of course, this large-scale evaluation can be executed in a distributed computer network where multiple machines can execute the evaluation, but it would still need significant resources to do that

- the second problem is processing the result as interpreting (e.g. visualzing) the result of the large-scale evaluation is challenging as well as understanding the patterns from this result (for example Figure 4.1 shows only 63 data points and it is difficult to understand the patterns even in this small example)

The large-scale evaluation of the available features is challenging, however, only part of the result of the complete large-scale evaluation is important to understand what individual features can be used (in addition to the time and weather-related features) to further improve the accuracy of the Random Forest Regression statistical regression approach. This information can be collected by executing a stepwise feature optimization method which method is an iteration based algorithm including the following steps:

- the method calculates the accuracy of the model using all the available features and it starts the first iteration from this state
- in the beginning of each iteration, the method creates the list of possible next states which states include the addition of one currently not used (if it possible) single feature and the substraction of one currently used single feature
- the method then evaluates the accuracy of all the possible next states and selects the most accurate model
- finally, the method selects the most accurate state as its current state and it carries on with the next iteration
- in the case of local minima (where the current possible states does not offer improvement in the accuracy), it follows a simulated annealing approach and carries on with the non-optimal next step
- after a given number of local minima, the method randomly makes steps to step out from the local minima circle

The stepwise feature optimization method, therefore, produces the list of individual features to use to train the Random Forest Regression algorithm to achieve the most accurate statistical regression model. Figure 4.4 shows the result of this method on the current regression task. The method selected the time (hour, month, day_of_week, bank_holiday, race_day) and weather-related (wind_direction, wind_speed, rain, temperature, pressure) features which result is in line with the findings of the previous analysis. The previous analysis was investigating the subset of the input features based on their data source and it shows that using all the features of the time and weather data source generates the most accurate Random Forest Regression model. The current stepwise feature optimization analysis found the same features as the most optimal subset of the features from the all available input features (but this method has the advantage of cherry-picking any individual feature from the available features while the previous could only use groups of features based on their data source). The figure also shows that the method stuck in local minima in the first 75 iterations and find the global minima afterwards.

The Random Forest Regression algorithm using only the time and weather-related data produces a statistical regression model with 11.97 $\mu$gm$^{-3}$ RMSE and 8.85 $\mu gm^{-3}$ MAE accuracy

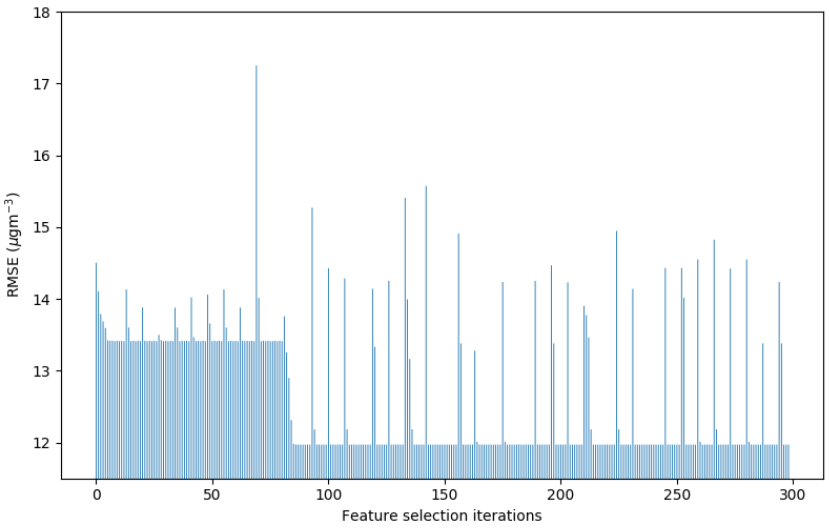


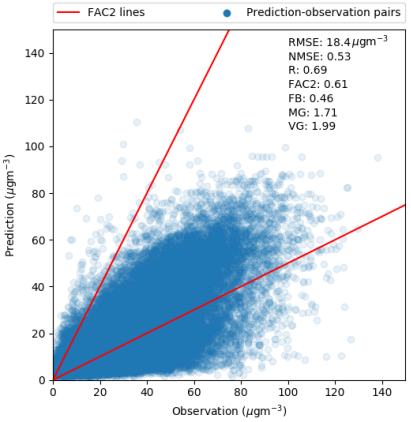

Figure 4.4. RMSE error levels during the feature optimization technique

according to the developed evaluation framework. These predictions also indicate 0.13 NMSE, 0.78 R, 0.00 FB, 0.93 MG, 1.25 VG, 0.90 FAC2 high-level accuracy levels. These values indicate that the model is more accurate than the previous RFR+ALL model as well as the state-of-the-art OSPM air pollution dispersion model. This approach will be referred to as RFR+TW in the rest of this chapter. Figure 4.5 shows the observation and prediction plot for the OSPM model, the Random Forest Regression and the RFR+TW approaches to understand the high-level RMSE accuracy difference in the terms of prediction-observation pairs. The plot shows that the predictions of the RFR+TW approach are more accurate as the shape of the point cloud is thinner than the shape of the cloud of the Random Forest Regression approach which result is in line with the high-level accuracy differences.

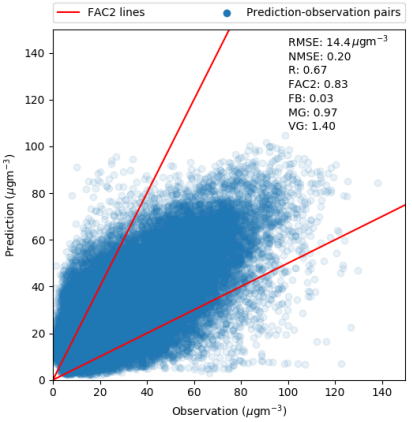
Figure 4.6 shows the box plot of the absolute errors of the predictions by the OSPM, Random Forest Regression and the RFR+TW approaches. The plot indicates that the RFR+TW model generates hourly $NO_2$ predictions more accurately than the state-of-the-art air pollution dispersion model (OSPM) having the same properties as the Random Forest Regression approach (e.g. avoid the usage of uncertain data sources such as the vehicle emission inventory dataset)

The stepwise feature optimization technique gave the list of features to use to maximize the achievable high-level accuracy by reducing the input data and keeping the features only matters to the given regression task for the underlying Random Forest Regression technique. It only selected the time and weather-related data which raises a question about the regression task:

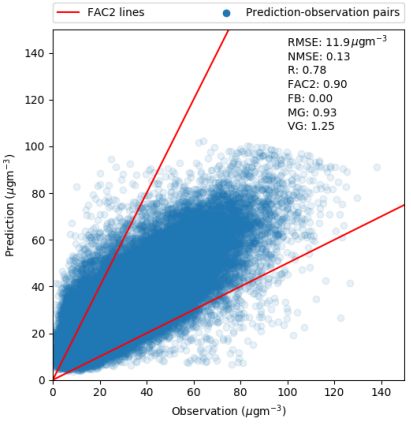
- if the traffic is the primary source of the $NO_2$ pollutant in the urban area, why does not the




(a) OSPM




(b) Random Forest Regression




(c) RFR+TW

Figure 4.5. Observation and prediction plot comparison for the OSPM, RFR and RFR+TW models

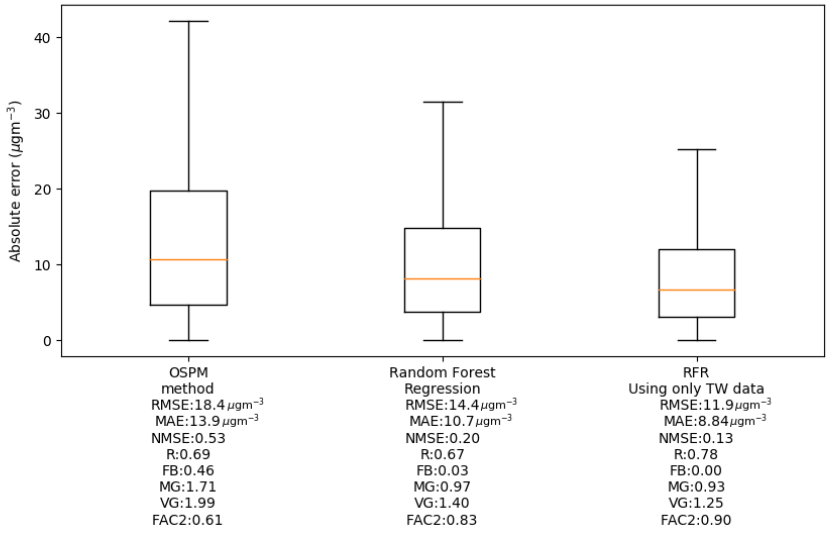



Figure 4.6. Absolute error plot of the predictions of the OSPM, the Random Forest Regression and the RFR+TW models

traffic information help to the statistical regression model (how is it possible that using the traffic data only introduces more error to the regression)

The next section is investigating this question by analysing the connection between the input data and the hourly $NO_2$ concentration levels.

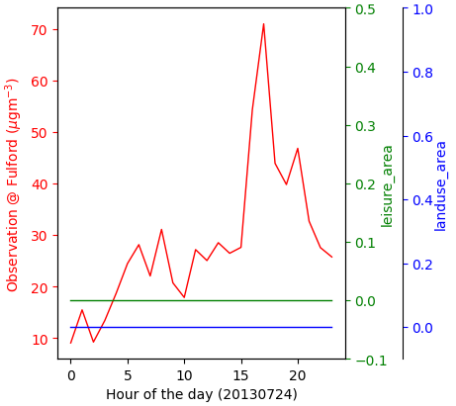
### 4.2.2 Input data analysis

The Random Forest Regression approach gives its most accurate predictions for the hourly $NO_2$ concentration levels if it uses only the time and weather-related input data. It is not clear, however, that why the usage of traffic data (or data from other data sources) introduces more error for the given regression task. The visualization of the input data and the $NO_2$ concentration levels might help to gather insight of the given regression problem.

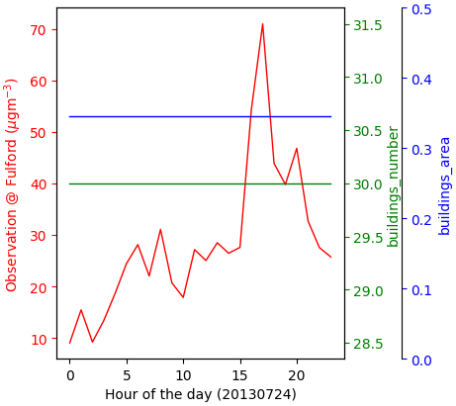
Figure 4.7 shows the input data features grouped by their data source and the concentration levels at the Fulford station for 24 hours of the day 24/07/2013.

**Monitoring data** The hourly $NO_2$ concentration levels provided by the monitoring stations are the prediction target for the regression algorithms. All of the plots include the concentration levels to understand the correlation between the input data and the concentration levels. The given example (24 hours of the day 24/07/2013 at Fulford station) shows low concentration levels in the morning, then it peaks in the afternoon.

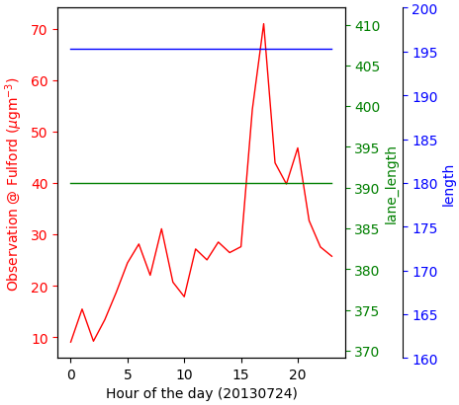
**Land use data** Land use data is a low temporal data source and the plot shows that the buffer area around the Fulford station has neither any land-use area nor leisure area as both features have the value of 0.0 across.



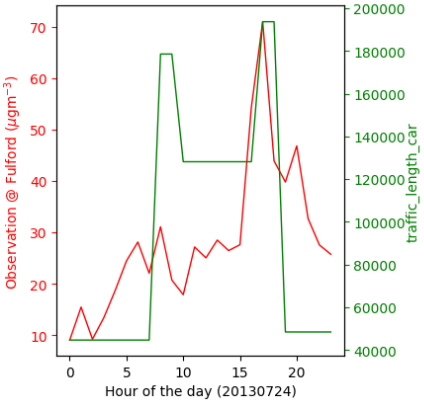

(a) Land use data



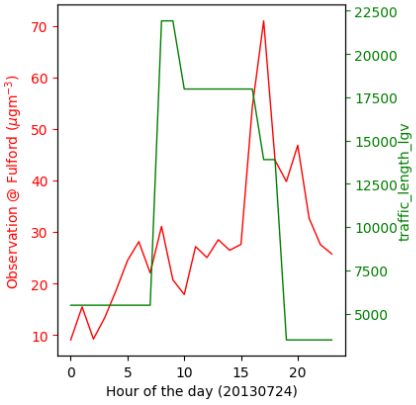

(b) Building data



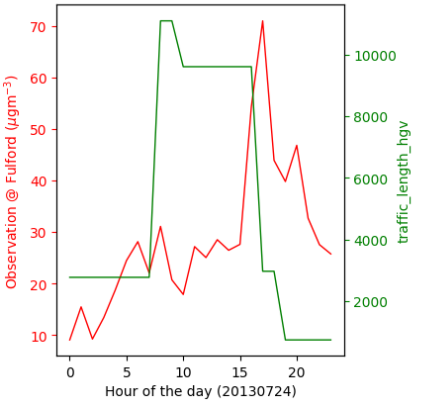

(c) Road data




(d) Car traffic data

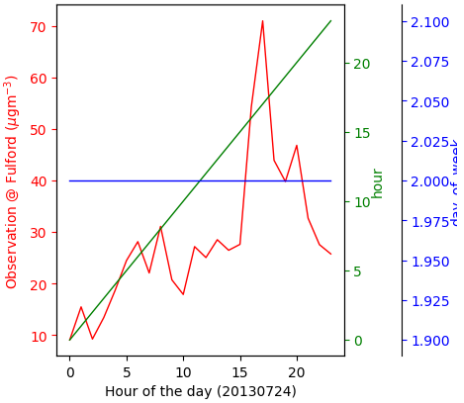



(e) LGV traffic data

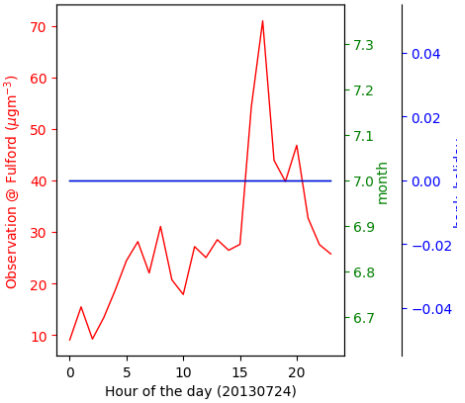



(f) HGV traffic data

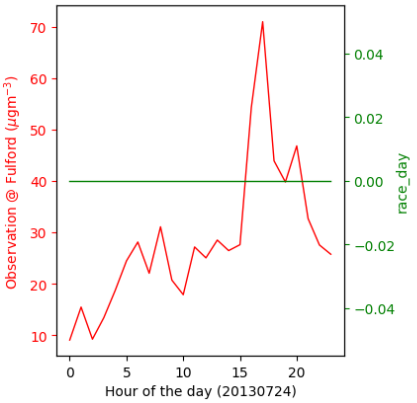
Figure 4.7. Concentration observation levels and different input data visualization



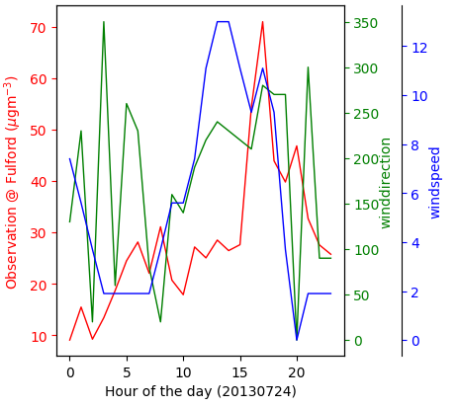

(a) Time related data first part



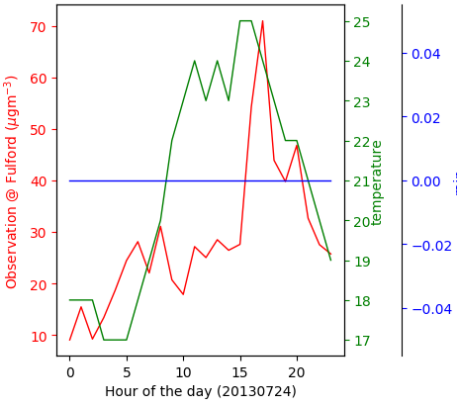

(b) Time related data second part



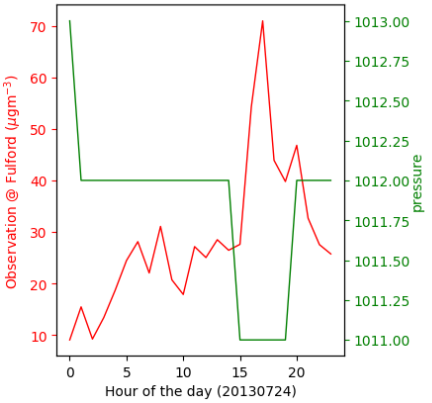

(c) Time related data third part




(d) Weather related data first part




(e) Weather related data second part




(f) Weather related data third part

Figure 4.8. Concentration observation levels and different input data visualization second part

**Building data** Similarly to the land use data, the building data is a low temporal data source and the plot shows that the buffer area has 30 buildings which cover approximately 35 percent of the buffer area and these values do not change over the day

**Road data** The last low temporal data is the road data and the plot shows that the buffer area has 195 meters of roads which roads are typical two-lane roads (showing the lane_length value of 380) and these features do not vary over the day

**Time-related data** The features in the time-related data group shows low variation in the visualized 24-hour time period as most of the features are low-temporal data (e.g. month is only changing once per month). These features, however, give information about the time and the statistical regression approach can learn time-dependent knowledge purely from the available observations (e.g. it can find out when the traffic is peaking at an average work day based on the observable high pollution levels during these hours)

**Weather-related data** Weather related data is a high temporal data source which covers the properties of the environment of the modelling area. The plot shows that features of this data group varies highly depending on the environmental circumstances of the given hour.

**Traffic data** The original traffic data source provided traffic volumes for each road within the modelling area and this dataset has been transformed into specific data for the buffer area by extracting the traffic related to the roads within the buffer area and weighted by the length of the roads (again roads only in the buffer area). These traffic volumes are artificially generated by a traffic model developed and maintained by the City of York Council's Transportation Management Group. This model only contains volumes for three vehicle categories (car, LGV, HGV) and only three time periods (and these time periods are extended to generate data for every hour of a day). The plot shows these different time periods for each category.

The low temporal data sources (land use data, building data, road data) provided features which are indicators of certain processes in the urban area which processes might cause an increased amount of $NO_2$ concentration levels in general. These indicators, therefore, do not have sufficient accuracy for the Random Forest Regression algorithm to use during its prediction process. The high-temporal data sources (traffic, time and weather-related data), however, provide important hourly information for the statistical regression model. The time and weather-related data are selected by the stepwise feature optimization method, however, the traffic data was not.

The plot helps to understand why the Random Forest Regression introduces more error in the case of using the traffic data. The plot shows that in the morning time period of the given day, the $NO_2$ concentration levels are low, however, the traffic data shows a significant amount of traffic for the same time period. If the $NO_2$ concentration levels have a general build-up period during the day, the statistical regression nature of the Random Forest Regression algorithm would have allowed the algorithm to learn this process and the algorithm could apply this knowledge during the prediction. This example is an edge-case scenario where to model fails to predict the concentration level accurately. Comparing the information contents of the observation data and

the traffic data reveals another important fact:

- the observation data actually represents the $NO_2$ concentration level at a given hour
- the traffic data is just an estimate of traffic volumes on the street calculated based on certain assumptions

The traffic data in its current form does not represent real-world (observed) traffic data as it is only an estimate, therefore, it does not give detailed information on the actual traffic around the monitoring station (or traffic in the buffer area) in the given hour. The real-time traffic data is an important information because it can identify traffic jams and traffic jams often cause high $NO_2$ concentration levels as many vehicles on the road are emitting pollution into the air. It is important to understand that an estimated traffic volume is a good indication of the average $NO_2$ concentration level for a given hour, however, it leads uncertainty in the case of predicting actual $NO_2$ hourly concentration levels. This crucial information is not covered by the current traffic data, however, it is possible to change this data source into another source which can provide the right data.

## 4.3 Changing the traffic data source

The original traffic data provided by the City of York Council's Transportation Management Group only contains estimates of traffic volumes for the roads in the York area. This data does not capture fine granularity of the actually observed traffic volumes which is required to give real-world information for the statistical regression method to be able to incorporate this data and exploit the information to make more accurate predictions. This data was originally selected because many previous studies included similar datasets to predict low-temporal pollution concentration levels. The Transportation Management Group also maintains a passive sensor network to count traffic volumes for roads in York. This simple traffic data count provides data for their traffic model which model also uses other assumptions about the vehicle movements in York.

### 4.3.1 Automated Traffic Count data

The Automated Traffic Count (ATC) data contains simple traffic count data because the passive sensor network contains automated traffic count instruments. This data only contains one single count value (compared to the three vehicle categories of the traffic model data) at the sensor location. The locations of the sensors are also limited in York (compared to the data provided by the traffic model which gives estimates for every road in York) as these sensors are real instruments and they need to be maintained by the Transportation Management Group. Most of the monitoring stations, however, have been co-located with an ATC instruments, therefore, the ATC data can be extracted for certain monitoring stations. The data itself contains data gaps (as expected from real-world data). The data availability of the ATC data, therefore, creates a different

regression task:

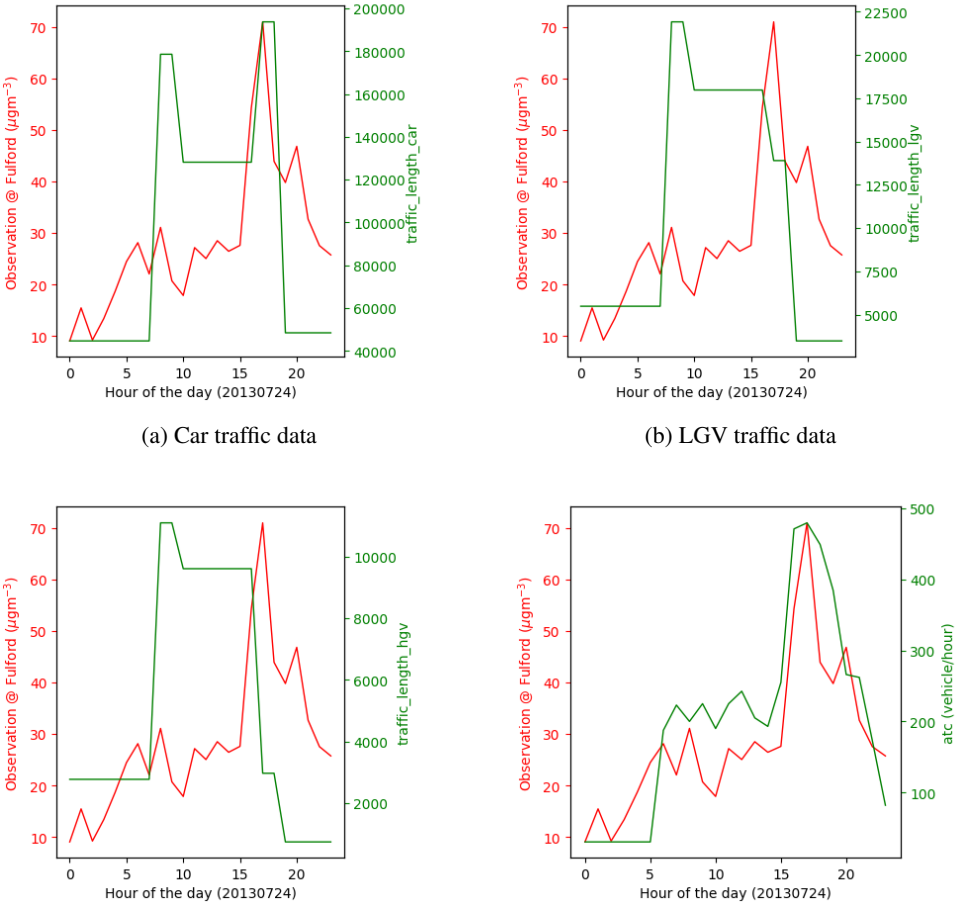
- Fulford station has 8228 observations using the ATC traffic data (8228 observations previously)
- Heworth station 7600 observations using the ATC traffic data (7600 observations previously)
- Fishergate station 8496 observations using the ATC traffic data (8496 observations previously)
- Gillygate station 6799 observations using the ATC traffic data (7490 observations previously)
- Lawrence station 7858 observations using the ATC traffic data (7948 observations previously)
- Nunnery station has no ATC station (7160 observations previously)
- Holgate station has no ATC station (8357 observations previously)

Figure 4.9 shows that the input data now can capture the real-world nature of the traffic which can explain some of the unusual $NO_2$ concentration levels observed by the monitoring stations. The opportunity to learn the connection between the real-world traffic volumes and the $NO_2$ concentration levels are given to the statistical regression approach, however, it is not clear what is the error levels on this new regression problem.

### 4.3.2 Evaluation of the usage of ATC data

The visualization of the new ATC traffic data shows that the new data can provide real-world observations of the traffic which can be an important information for the statistical regression model as the traffic is the primary pollution source for the $NO_2$ pollutant in the given modelling area (in York). It is not clear whether the Random Forest Regression would be able to utilize this data to make more accurate hourly $NO_2$ concentration level predictions, therefore, the evaluation of the algorithm using the new data source was executed:

- the new data source has its own data gaps, therefore, the usage of the ATC data is creating a slightly different regression problem
- the high-level accuracy of the Random Forest Regression method (using all the available data excluding the new ATC) is not known on this regression task, therefore, evaluation of the method is required by executing the developed evaluation framework
- the result of this execution will provide the baseline for further evaluation




(a) Car traffic data

(b) LGV traffic data

(c) HGV traffic data

(d) ATC data

Figure 4.9. Data visualization of the old traffic data and the ATC data including the concentration observation levels

- it is also not known that what high-level accuracy can be achieved by using the new ATC data compared to the old traffic data and how these accuracy levels compare to the baseline accuracy levels

To answer these questions, several evaluation runs were executed using the developed LOOCV evaluation framework. The new regression task allowed only 5 iterations for the cross-validation evaluation because the ATC data only have observations for 5 monitoring locations. All the runs were using the Random Forest Regression algorithm for the generating the hourly $NO_2$ concentration level predictions and the high-level RMSE error was calculated to describe the achievable error levels. The only difference between the runs was the input data given to the Random Forest Regression algorithm.

- Using all the previously available data (from all the data sources excluding the new ATC data) achieved 15.06 $\mu gm^{-3}$ RMSE error level. This level is slightly higher than the RMSE level observed in the previous regression task (14.45 $\mu gm^{-3}$) and this result is in line with the nature of the new regression task as this excludes two monitoring stations which have lower $NO_2$ concentration levels, therefore the current regression task is slightly harder as the observation levels are higher.

- Using only time and weather-related data (RFR+TW) generated a regression model with 12.68 $\mu gm^{-3}$ RMSE accuracy which is again in line with the previous findings

- Using only time and weather-related data plus the old traffic data introduced more error to the predictions as it generated a model with 14.38 $\mu gm^{-3}$ RMSE level (again, this is in line with the previous findings)

- Using the available time and weather-related data plus the new ATC data generated a statistical regression model generating the hourly $NO_2$ concentration level with 13.57 $\mu gm^{-3}$ error level. This result indicates that using the new ATC traffic data helps to make more accurate predictions than a model which is using the old traffic data, however, the error level is still greater than the error level of the RFR+TW model.

The result suggests that using the new ATC data (additionally to the time and weather-related data) does not provide a more accurate statistical regression model. From the experiment, it is not clear that using the ATC data with the existing features would give a more accurate model for the hourly $NO_2$ concentration level prediction task. The stepwise feature optimization task was executed again on all the available features (now including the new ATC data). Figure 4.10 shows the first 300 iterations of the feature optimization algorithm. The method found the global optima after 20 iterations and it did not find a better feature subset afterwards. The optimal subset of the features included again only the time and weather-related features (and it did contain neither the old traffic nor the new ATC traffic data).

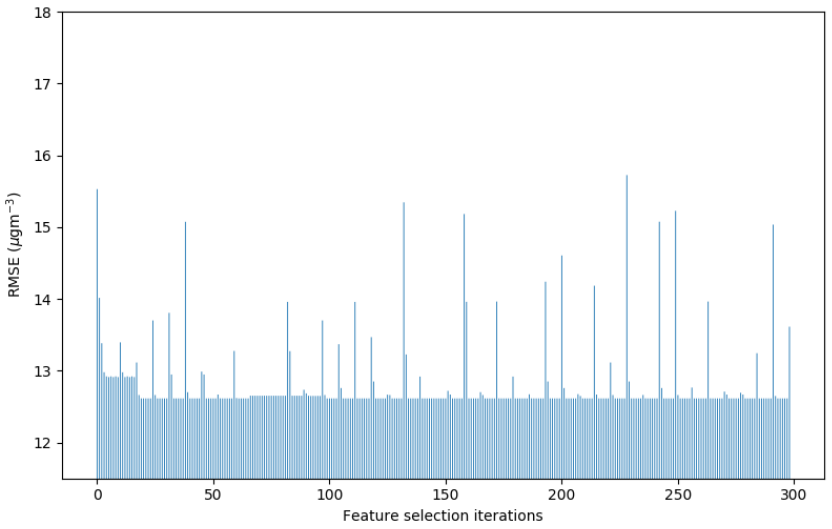


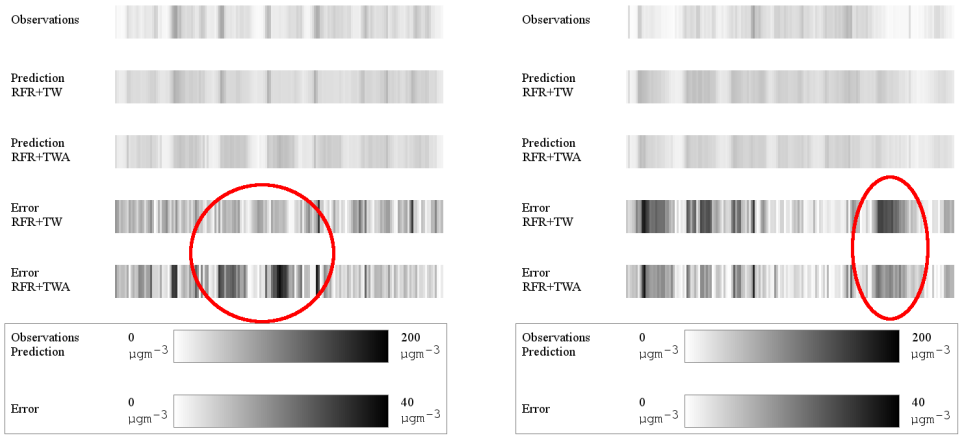

Figure 4.10. Visualization of the calculated RMSE accuracy level during the iterations of the stepwise feature optimization method

This result with the previous results (experimenting with the old and new traffic data) introduce a new problem for the model evaluation. The model has been evaluated by using observations for 7 stations (or 5 stations) geographically distributed in the modelling area. The weather and time-related data, however, only contains observations which are identical at each observation station:

- Weather-related data is identical at each station because the data source provided a high-level average weather condition for the whole city
- Time-related data is identical at each station because the features within this data group are identifying the observation/prediction time (e.g. hour of the day, month of the year)

The developed statistical regression model will give the same $NO_2$ concentration level prediction for the whole modelling area if it is only using the time and weather-related data which would give insufficient predictions as the main purpose of these models to understand the spatial and temporal changes of the $NO_2$ concentration levels in the modelling area.

The evaluation of the RFR+TW and RFR+TWA models introduced 12.68 $\mu gm^{-3}$ and 13.57 $\mu gm^{-3}$ high-level RMSE accuracy values, respectively. The difference between the RMSE values indicates that the RFR+TW model makes the hourly $NO_2$ concentration levels predictions more accurately, however, the high-level RMSE does not provide fine details of prediction errors (e.g. the hourly prediction errors in details). To investigate these final detailed errors, the visualization of the observation and prediction concentration levels of the RFR+TW and RFR+TWA




(a) Fishergate station between 25th March 2013 and 31st March 2013

(b) Fulford station between 22nd July 2013 and 28th July 2013

Figure 4.11. Visualization of the concentration level observations, predictions and prediction errors by the RFR+TW and the RFR+TWA models

models has been created. This visualization revealed that the RFR+TW and RFR+TWA models are mostly predicting the same hourly $NO_2$ concentration levels, however, there are certain episodes where one (RFR+TW) or the other (RFR+TWA) model generates predictions with higher error levels. Figure 4.11a and Figure 4.11b show two examples of these error episodes:

- Figure 4.11a shows that the RFR+TW model manages to predict the concentration levels accurately during the visualized period, but the RFR+TWA model generates a prediction error episode in the middle of the period as it fails to predict the concentration levels accurately (as it is using the additional ATC data and the model was trained on that data as well which now introduces this error episode). This result is in line with the high-level RMSE error level analysis as the RFR+TWA model is expected to produce more errors on average.

- Figure 4.11b shows the opposite process to the previous example as the RFR+TWA model generates more accurate predictions compared to the RFR+TW model. This was not expected from the high-level RMSE error analysis as this example shows that the RFR+TWA model produces more accurate predictions in some cases, however, on average the accuracy of this model is worse than the accuracy of the RFR+TW model.

Further investigation of the prediction error episodes of the RFR+TW and RFR+TWA models showed that these error episodes are non-overlapping. The second example (Figure 4.11b) indicates that the RFR+TWA model can produce more accurate predictions, however, on average the RFR+TWA is introducing more errors to the prediction. This finding is important as it means that there is a benefit to using the RFR+TWA model in certain circumstances, however, these

circumstances are not known. One possible explanation is that the traffic represented by the ATC data helps (RFR+TWA is better) when there is traffic jam combined with some specified weather condition, however, this case only represented a few times in the complete dataset, therefore, the Random Forest Regression algorithm ignores it, because it treats this case as an outlier.

Having established that RFR+TW and RFR+TWA models generate non-overlapping error episodes, the analysis of the input data was carried out to understand the prediction circumstances for these prediction error episodes. It is important to understand these circumstances as this knowledge can open the possibility of utilizing both models and set up the understanding of choosing the right model for the right predictions.

The analysis of specific rules (rules to decide what model to use for certain input data) was carried out which helps to create the systematic assessment of the prediction error of the two models. The rules were developed by using prior knowledge about the modelling area. In general, the RFR+TW model provides the most accurate predictions, however, it does not use information about the traffic. In cities, traffic peaks twice a day when commuters flood the roads (so they called morning and afternoon traffic peak period). We then separated two different time windows focusing on days where the weather does not affect the pollution (e.g. the wind speed is low):

- *morning*: before the morning traffic peak period, when the pollution has been cleaned out during the night (4AM-7AM)
- *afternoon*: during the afternoon traffic peak period, where traffic is high on the roads and traffic jams are highly likely (4PM-7PM)

Figure 4.12 shows the results of analysis of absolute error in prediction during these time windows using the model RFR+TW, RFR+TWA, and RFR+WA. The RFR+WA model was included in this analysis to investigate the accuracy of a model which does not have information about the time-related data. In the morning case, there is no benefit of using more data than the T+W. Using RFR+TWA model, however, shows less error in prediction when the traffic is peaking (afternoon case). Moreover, in this situation, using time-related data does not show relevance as the RFR+TWA and RFR+WA show similar prediction accuracy.

This result motivates the usage of complex modelling system where multiple random forest statistical regression models are being trained on different subsets of the input data and a model selector decides what model to use in which situation to exploit the non-overlapping error episodes of the different models.

## 4.4 Ensemble of the Random Forest statistical regression method

During analysing the detailed prediction errors of the RFR+TW and RFR+TWA models, the non-overlapping error episodes and the possibility of using a model selector to select the prediction output of the RFR+TW and RFR+TWA became apparent. Therefore in this section, an automated systematic model combination is developed and evaluated.

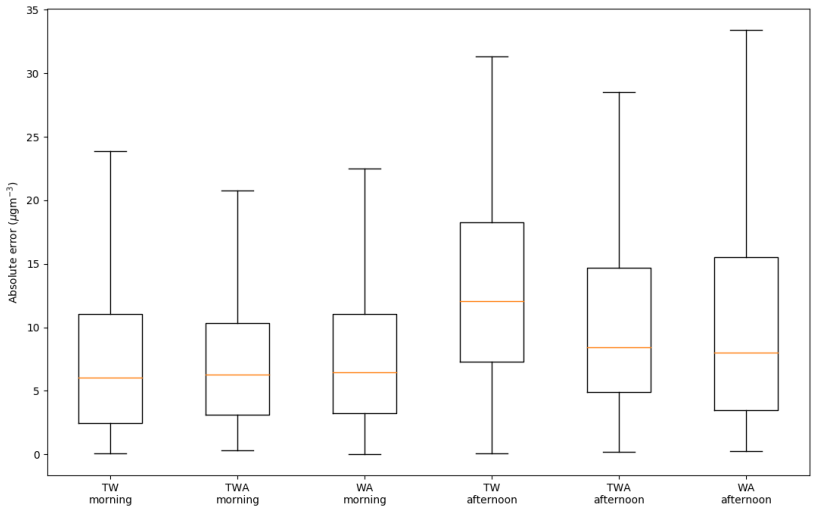



Figure 4.12. Absolute error plot of RFR+TW, RFR+TWA, and RFR+WA in the morning and afternoon time windows

### 4.4.1 Automated ensembling of the RFR+TW and RFR+TWA models

The process of using the predictions output of different machine learning models and combine them together is often referred as model ensembling [Dietterich (2000); Kotsiantis et al. (2007)]. One of the simplest model combination methods is to generate a classifier which decides (based on the input data plus the prediction output of the different models) when to use which model (in this case either to use the output of the RFR+TW or the output of the RFR+TWA). One of the advantages of using this model combination approach is to the possibility of calculating the best case scenario. Using the already developed cross-validation evaluation framework, it is possible to apply both RFR+TW and RFR+TWA models and select always the concentration level prediction which prediction is closer (has the smaler absolute error value) to the observed concentration level.

The perfect model combination of the RFR+TW and RFR+TWA model would give the following theoretical accuracy level:

- the current dataset using the ATC data contains data for 5 stations which contains 38981 data points
- using the RFR+TW model gives 12.68 $\mu gm^{-3}$ RMSE high-level accuracy
- using the RFR+TWA model gives 13.57 $\mu gm^{-3}$ RMSE high-level accuracy
- from the 38981 predictions the RFR+TW model gives more accurate predictions on 24558 occasions

- from the 38981 predictions the RFR+TWA model gives more accurate predictions on 14423 occasions
- using the perfect RFR+TW and RFR+TWA model combiner would give a statistical regression model with 5.83 $\mu gm^{-3}$ RMSE high-level accuracy

This result indicates that a statistical regression model using the perfect RFR+TW and RFR+TWA model selector can generate predictions with 5.83 $\mu gm^{-3}$ RMSE accuracy level on the current regression task, however, achieving this accuracy level needs a perfect classifier.

This result is promising as the achievable RMSE accuracy level is greater than the accuracy level of the single RFR+TW and RFR+TWA models, however, the result is only theoretical as it is challenging to develop a perfect model selector (classifier) for this regression task. It is evident that we can use an existing classification algorithm to do the model selection. Based on the success of the Random Forest Regression algorithm (on the regression problem), the Random Forest Classification algorithm was chosen to perform the classification task.

The model selection method performs the following steps to build the appropriate classification model:

- the current LOOCV evaluation framework utilizes data from only 4 stations to build the statistical regression model and the framework applies the model and evaluates the accuracy of the model on data of the fifth stations (and repeats this process 4 more times to apply the model on all the five stations)
- the model selection requires data to train a classification model which data includes the input data and the concentration level observations and prediction output of the RFR+TW and RFR+TWA models
- the model combination method first use only 3 stations data to train to RFR+TW and RFR+TWA models and applies it to the fourth station to generate the required concentration level predictions for the classification
- then based on the predictions given by the RFR+TW and RFR+TWA models, it assigns the value of 0 (prediction concentration level by the RFR+TW model is closer to the observed concentration level than the prediction concentration level by the RFR+TWA model) or 1 (prediction concentration level by the RFR+TWA model is closer to the observed concentration level than the prediction concentration level by the RFR+TW model) which provides the two classes for the classification
- this process is repeated four times to generate data for each station
- the Random Forest Classifier is trained based on the generated data
- RFR+TW and RFR+TWA models are trained using the data of the available 4 stations

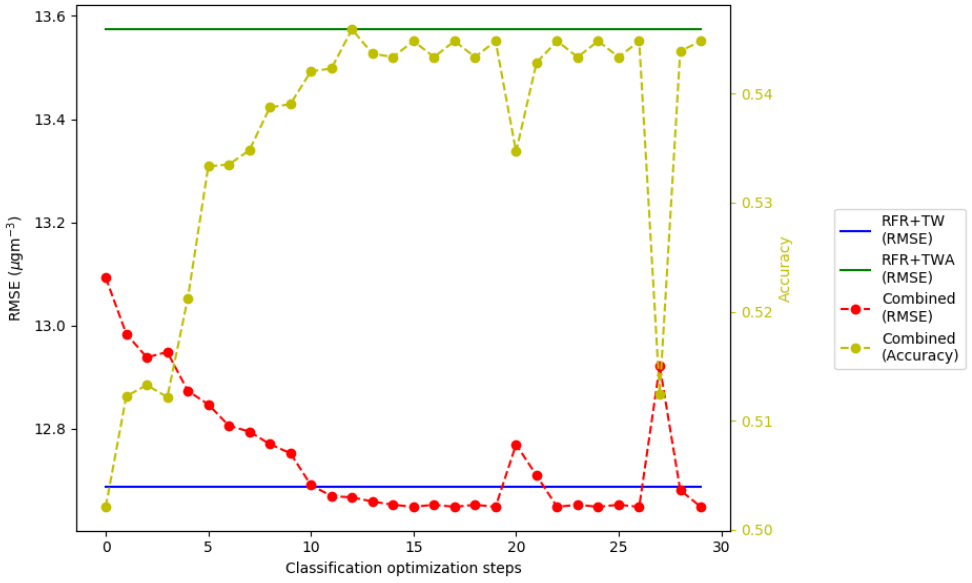



Figure 4.13. Visualization of the achieved accuracy levels (RMSE and classification accuracy) during the stepwise feature optimization run for the model selection classification

- RFR+TW and RFR+TWA models are applied to the fifth station as well as trained model selection classifier
- based on the output of the model selection classifier (it is either 0 or 1), the prediction output of the RFR+TW (if the classification output is 0) or the prediction output of the RFR+TWA (if the classification output is 1) will be selected for the final concentration level prediction
- the complete process is repeated 4 more times to cover all 5 stations

### 4.4.2 Optimization and evaluation of the ensemble method

The introduced approach provides a model selection classifier for the RFR+TW and RFR+TWA statistical regression models. This approach can be evaluated against the existing single RFR+TW and RFR+TWA statistical regression models, however, the underlying model selection classification needs to be first optimised for the given classification problem. A stepwise feature optimization can be executed for this optimization similarly to the previous feature optimization of the statistical regression approach.

Figure 4.13 shows the result of the stepwise feature optimization method for the classification method including also the accuracy of the classification of the model selection model. The stepwise feature optimization method started using all the available input data (the same input data which was developed for the regression task including all the data sources). Using all the data the model combination method generated 0.5021 classification accuracy which generated a statistical

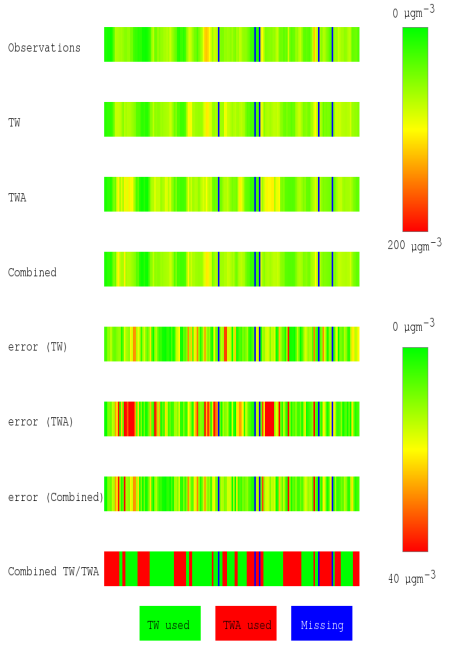
regression model with 13.09 $\mu gm^{-3}$ high-level RMSE accuracy. The stepwise feature optimization method then executed its iterations to find input feature subset to increase the accuracy level of the underlying model selection classification (which lead to improvement in the high-level RMSE regression accuracy). The global optima for the classification method were reached after 15 iterations which model was producing 0.5448 classification accuracy and 12.64 $\mu gm^{-3}$ RMSE high-level accuracy. The best subset of the input features includes the hour, day_of_week, month, bank_holiday, race_day, windspeed, temperature, rain, pressure, lane_length, length, leisure_area features which indicate that the classification model is using lane_length, length, leisure_area features (not only the time and weather-related data).

The result indicates that the presented method could provide more accurate hourly $NO_2$ concentration levels than the RFR+TW method utilizing the predictions of the RFR+TW and RFR+TWA models and selecting the appropriate prediction outputs. The achieved accuracy level of the combined method is far from the introduced model combination using the perfect classifier, but this result was expected as the achieved accuracy of the actual model selection classification is very low. Again, the high-level RMSE error does not provide fine details of the prediction errors, but the visualizations of the observations and predictions were generated to understand the introduced error by the new model (including the predictions of the existing RFR+TW and RFR+TWA methods).

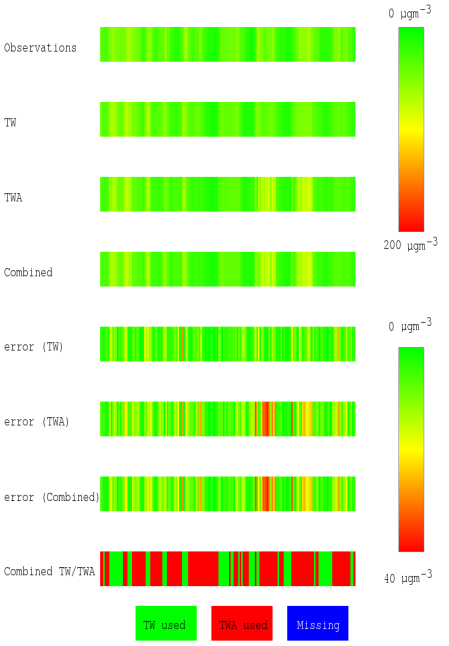
Figure 4.14 shows the visualization of the predictions of the developed model combination method as well as the predictions of the underlying RFR+TW and RFR+TWA models. The plot shows the four notable possible cases of the model selection classification outcome:

- the predictions of the RFR+TW model are showing an error episode, but the predictions of RFR+TWA model are close to the observations and the model selection selects the RFR+TWA model, therefore the final predictions are close the observations
- the predictions of the RFR+TW model are showing an error episode, but the predictions of RFR+TWA model are close to the observations and the model selection selects the RFR+TW model, therefore the final predictions are showing the error episode
- the predictions of the RFR+TWA model are showing an error episode, but the predictions of RFR+TW model are close to the observations and the model selection selects the RFR+TW model, therefore the final predictions are close the observations
- the predictions of the RFR+TWA model are showing an error episode, but the predictions of RFR+TW model are close to the observations and the model selection selects the RFR+TWA model, therefore the final predictions are showing the error episode

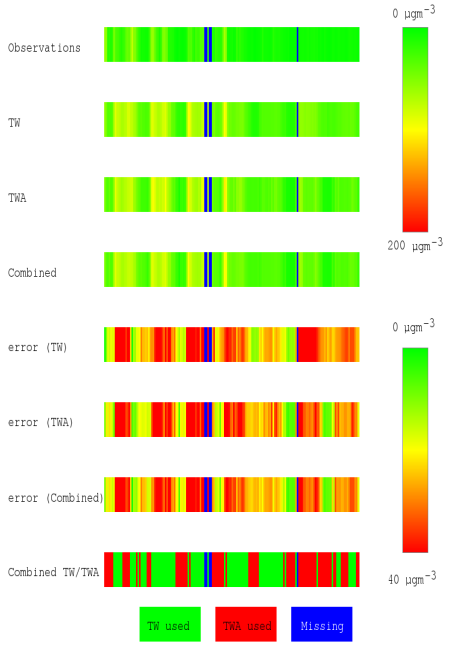
These examples are indicating that the model selection is capable of exploiting the differences in the predictions of the RFR+TW and RFR+TWA models and the model selection can



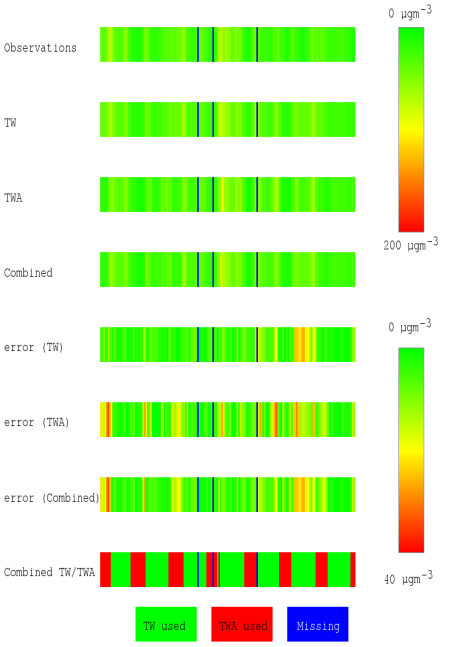

(a) Fishergate station between 18th February 2013 and 24th February 2013




(b) Fishergate station between 14th October 2013 and 20th October 2013




(c) Fulford station between 11th March 2013 and 17th March 2013




(d) Heworth station between 15th April 2013 and 21st April 2013

Figure 4.14. Visualization of the concentration level observations, predictions and prediction errors by the RFR+TW and the RFR+TWA and the combined models including the model selection classification prediction output

select the appropriate (more accurate) model in some circumstances. Understanding these circumstances is challenging because it requires the detailed analysis of the decision mechanism of the Random Forest classification method on this specific task (which include the analysis of hundreds of decision trees generated by the Random Forest classification algorithm). These circumstances need to be complex otherwise the underlying RFR+TW and RFR+TWA algorithms would have learned these and utilized the knowledge to generate more accurate predictions. The model selection, however, is not accurate enough to make the $NO_2$ concentration level predictions significantly more accurate than the underlying RFR+TW and RFR+TWA models, but the model selection classification method utilizes other data sources which make the developed model applicable to generate predictions for the whole urban area.

## 4.5 Summary

The aim of this chapter is to investigate the sensitivity of the applied input data to the prediction accuracy of the Random Forest Regression method. The analysis of the evaluation of the application of data from different data sources revealed that the time and weather-related data sources are crucial for the developed statistical regression approach. This result contributes to the Environmental Science field as it indicates what are the most important data for the Random Forest Regression statistical regression method. Moreover, the analysis highlighted that using the traffic data only introduces prediction error because the traffic data in question is only traffic volume estimates which do not represent the actual traffic in the observation hour.

The traffic data source has been changed to a different data source (ATC data source) which data provided actual traffic volume data for the regression model. The evaluation of using this new data source shows that using this data still increases the prediction error. The detailed analysis of the hourly $NO_2$ concentration level predictions revealed that the Random Forest Regression model trained on only time and weather-related data (RFR+TW) and the Random Forest Regression model trained on time, weather and traffic data (RFR+TWA) generates non-overlapping error episodes.

The existence of the non-overlapping error episodes in the concentration level predictions of the RFR+TW and RFR+TWA models suggests that selecting the prediction outputs of the two models at different input circumstances can utilize both models prediction power to further optimize the achievable prediction accuracy. A manual, simple rule-based case was investigated where the RFR+TWA could offer more accurate concentration level prediction compared to the RFR+TW model. The last section of the chapter describes the development of an automated model ensembling method which offers further improvement in the prediction accuracy by systematically selecting outputs of the RFR+TW and the RFR+TWA models.

The developed Random Forest ensemble method could produce more accurate hourly $NO_2$ concentration level predictions than the underlying RFR+TW and RFR+TWA models. This ensemble algorithm contributes to the Computer Science field as this novel ensemble algorithm can

be used to any other regression task where it is crucial to improve the overall regression accuracy.

The introduced model ensembling method provides further improvement for the hourly $NO_2$ concentration level predictions. It requires training two Random Forest Regression models (for concentration level predictions) and one Random Forest Classification model (for model selection). Building a Random Forest Regression (and Classification) model, however, is computationally expensive which raises the question on the scalability of the developed model. On the current dataset (which captures data of the current modelling area, York), the developed method is feasible and manages to make predictions within reasonable computational time, however, the approach on larger, more complex problems can struggle due to its high computation requirement. Also, applying the model to one dataset does not give enough information on the robustness of the developed approach. The next chapter will, therefore, investigate the scalability and robustness of the developed method by applying it to a larger, more complex environmental modelling problem.

# CHAPTER 5

## Robustness and scalability analysis of the Statistical Regression approaches

This chapter presents a detailed robustness and scalability analysis of the developed Random Forest Regression and Random Forest ensemble approaches.

The *scalability* is defined as the ability to carry out the model training and application on large environmental problems and the *robustness* is defined as the ability to produce accurate predictions in the case of a different modelling scenario. To understand the robustness and the scalability of the approaches, they will be applied to a large-scale environmental problem which covers the task of the hourly $NO_2$ concentration level prediction in the London area.

The chapter begins with the analysis of the developed Random Forest Regression and Random Forest ensemble methods application to the London dataset. The result of this analysis contributes to the Environmental Science field as the analysis indicates that the statistical regression approach can be applied to complex and large environmental modelling problems. The second part of this chapter explores the application of a different Random Forest ensemble method which algorithm contributes to the Computer Science field as it generates an algorithm which can be applied to any other regression task to further improve the regression accuracy.

In the first section (Section 5.1), the motivation of this work is explained which introduces the aim of the scalability and robustness analysis. Section 5.2 describes the large-scale environmental modelling problem including the input data collected for supporting the development of the statistical regression approach. The robustness and scalability analysis is described in Section 5.3. Based on the experience gathered during the analysis, a novel ensemble method is proposed in Section 5.4 which provides accurate hourly $NO_2$ concentration levels for the introduced large-

scale environmental modelling task. Finally, the Section 5.5 finalizes the chapter.

## 5.1 Motivation

The developed statistical regression approaches (the Randon Forest Regression method and the Random Forest ensemble method) were developed and evaluated using the York dataset. The complexity of the regression task given by this dataset is considered as average because

- there is only 7 stations (5 stations with the ATC data) in the modelling area producing 8760 hourly observations per year (approximately 61320 and 43800 data records if there is no gap in the data) which is easily processable for the existing scikit-learn implementation of the Random Forest algorithm
- the observation data contains concentration level observations in the same value range at the different monitoring locations, therefore, the complexity of the underlying environmental modelling problem is simple as the primary source of the pollution is the traffic in the modelling area (and there is no pollution heavy industry or any other major pollution source)

The statistical regression approach needs to use historical observations of the $NO_2$ concentration levels and other relevant information of the environment. The sensitivity analysis of the prediction accuracy to the dataset given to the Random Forest Regression algorithm (presented in the previous chapter) revealed that the accuracy of the predictions generated by this statistical regression approach highly depends on the quality of the available data. The algorithm requires intense computation to generate the underlying decision trees (compared to the standard Linear Regression statistical regression algorithm), however, the algorithm generated these internal data structures quickly enough to evaluate the accuracy on a single machine. The general scalability of the developed approach, however, is not known, therefore, this chapter aims to understand the feasibility of the application of the Random Forest Regression and the Random Forest ensemble methods on a large-scale environmental modelling problem. Modelling the $NO_2$ concentration levels on the urban area gives the most challenging predictions task in this $NO_2$ concentration level modelling field as the modelling area is complex and multiple independent processes are affecting the concentration levels (e.g. traffic is a primary source, but the urban geometry alters the concentration levels because it enables certain processes to release and keep the emitted pollution from the street level). From the environmental science aspect, one of the most analysed urban area is the London area:

- London has a very complex urban geometry including one of the most polluted street in the world (Oxford street)
- London has a very congested road network resulting high level of pollution emission by the vehicles

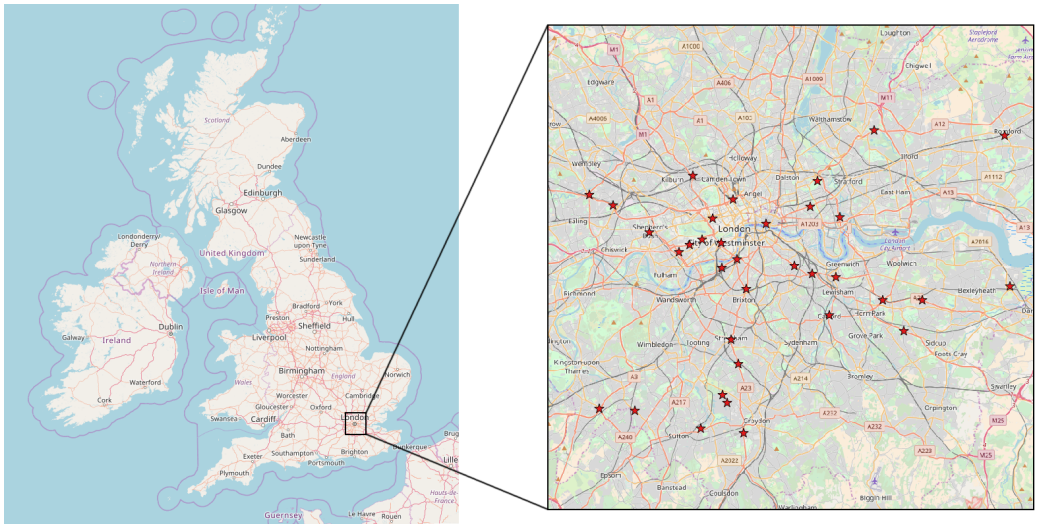



Figure 5.1. Geographical map of London with the monitoring station locations (red stars)

- London has large industry resulting in high level of pollution emission
- London has one of the densest pollution monitoring network in the world

These properties of London are suggesting that selecting London for the modelling area is desirable, therefore, the large-scale application of the developed models aims to predict the hourly $NO_2$ concentration levels for London. The chapter aims

- to develop the Random Forest Regression statistical regression approach for the London area to investigate the challenges of the development process itself
- to develop the Random Forest ensemble method for the London area to investigate the accuracy of the ensemble method
- to understand the scalability of the developed methods by investigating the computational time required to generate and evaluate the methods
- to understand the robustness of the developed methods by comparing the prediction accuracy results with accuracy results of the York model applications

## 5.2 Introduction of the large-scale environmental modelling problem

The first step of the application and evaluation of the statistical regression model is the data collection. Similar data was collected for the large-scale model application as in the previous chapters (Section 3.3) to feed the statistical regression model with historical observation data,

however, some of the data sources were not available for the London area. This section describes the collected data explaining the difference between the data collected here and the data collected for the previous model application. The description of the data transformation is also described to understand how the collected raw data was prepared to feed it into the statistical regression algorithm for hourly $NO_2$ concentration level predictions. The data collection method for this model application follows data sources in the existing studies of the literature [Hochadel et al. (2006); Stedman et al. (1997); Briggs et al. (1997)]. Figure 5.1 shows the London modelling area.

**Monitoring (observation) data** The most important data for the hourly $NO_2$ concentration level modelling is the hourly $NO_2$ concentration level observation data. In London, the London-air database contains data about the environment; this database is provided by the Environmental Research Group of King's College London. The database contains hourly concentration level measurements for $NO_2$ and other pollutants (e.g. $PM_10$, $PM_{2.5}$, etc.) from more than 100 monitoring locations. $NO_2$ concentration level data from 35 roadside stations (only these 35 stations have co-located traffic counter stations) have been acquired which covers the time period between 1st January 2016 and 31st December 2016. Figure 5.2 shows a boxplot of the observations produced by each station. These readings are considered to be high pollution levels as more than half of the stations have greater than 50 $\mu$gm$^{-3}$ median $NO_2$ concentration level observations. The collected observation data differs from the previously collected data as it has more stations (5 stations previously) and the observed concentration levels are different as there are stations with very high observed $NO_2$ concentration levels (compared to the previous data where the observations of the stations were close to each other).

**Land use data** Land use data has been collected using the Open Street Map database similarly to the previous model application. The available data describes the areas (in polygons format) usage scenarios (e.g. leisure, green areas, farm, etc.). The following data for each buffer area (around the monitoring stations) were extracted: "landuse_area" and "leisure_area" which are proportional area measurements of the specific subcategory of the polygons to the buffer area in the database.

**Building data** Building data has been collected using the Open Street Map database. The data source for this data is different from the previous model application as the Open Street Map database contains fine details of the existing buildings in the London area (which details have more detailed information about the buildings than the previously used Ordnance Survey's Mastermap database). The database gives spatial information about buildings within the area of interest. The raw data has been processed and the number of the buildings and area of the buildings covered by each buffer area generated the "buildings" and "buildings_area" features.

**Road data** Road data has been collected using the Open Street Map database. This data source is different from the previous model application however the Open Street Map database has very precise information about roads in the modelling area. Only static features were ex-

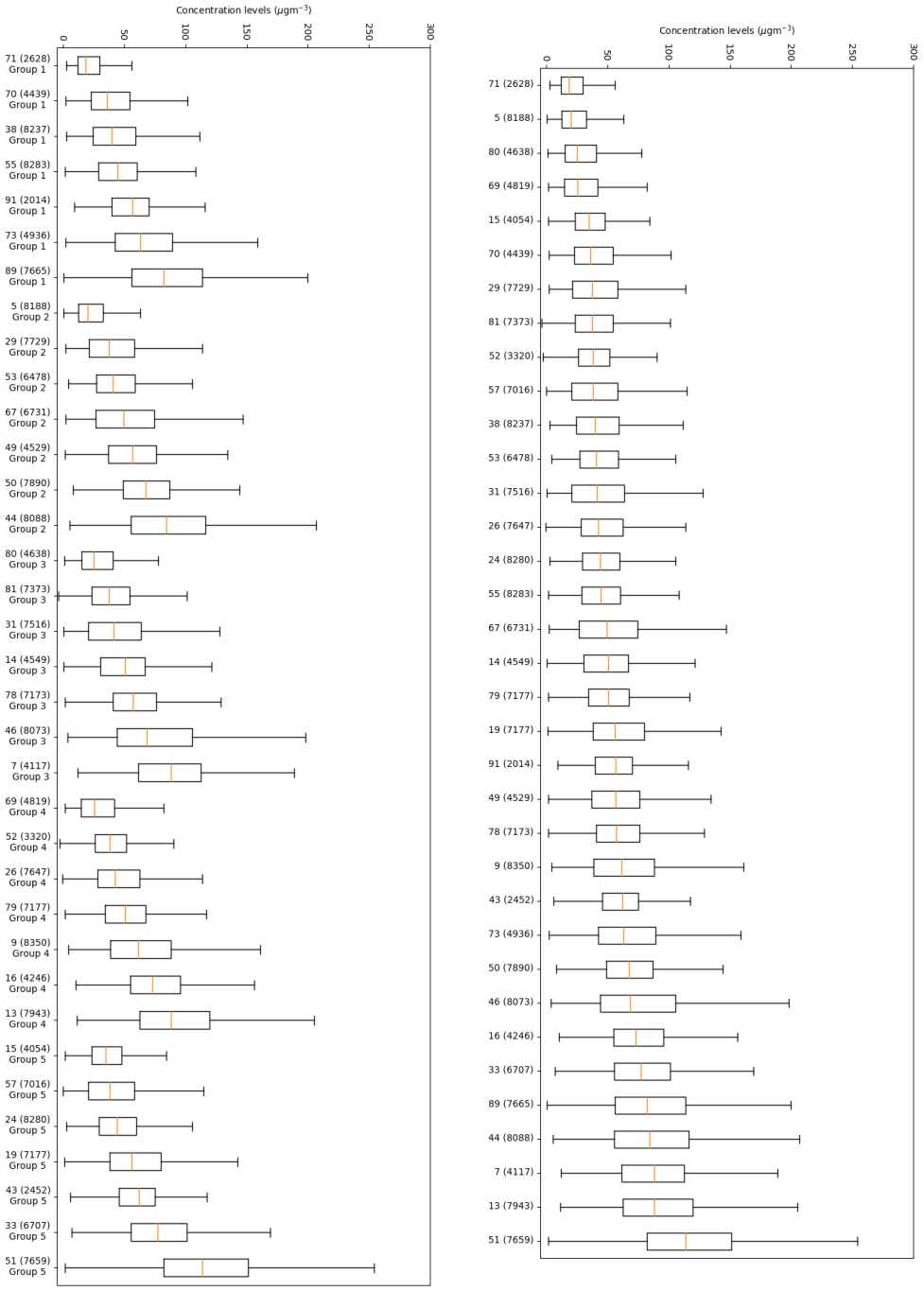



Figure 5.2. Monitoring data for the London modelling area (top) and the grouping of the monitoring data for the evaluation framework (bottom) including the station ID followed by the available observations for the station

tracted from this database such as the road_length (which covers the overall length of the roads within the buffer area) and the road_lane_length (which is using the lane number for a road as a multiplier for the given road's length).

**Automated traffic count data** Similarly to the previous model evaluation, Automated Traffic Count (ATC) data was collected from the Transport for London's Road Space Management group. This data covers the amount of the traffic around the monitoring stations (only 35 pollution monitoring stations have co-located traffic counter). This data is important as one of the main source of the $NO_2$ pollutant is the traffic and using this data helps the model to have information about this pollution source. The collected data captures the same time period as the monitoring observation data (time period between 1st January 2016 and 31st December 2016).

**Meteorological data** Meteorological data from the Weather Underground database (`https://www.wunderground.com/weather/api/`) has been acquired by using its API to download data. This database contains observations for cities and includes temperature, relative humidity, wind speed, wind direction, and pressure measurements. The data includes meteorological observations at all the stations because the modelling area is covering a larger area than the previous model application, therefore this data differs from the data previously used as the York meteorological data had only observations for the city on average. The time interval for this data matches the concentration level observation data time interval (hourly observations between 1st January 2016 and 31st December 2016).

**Time related data** Similar time-related indicators (e.g. hour of the day, day of the week, bank holiday, etc.) were generated as the previous model application, however, the York specific indicators (e.g. race_day) were excluded as these features are no longer valid for the given modelling area. It is practically hard to find similar features for this modelling area, because the area itself covers a much bigger area and events (such as football matches or concerts) only covers a small part of the complete modelling area.

Figure 5.2 shows the hourly $NO_2$ concentration levels at each station including the available observations per stations. The 35 stations produced 218121 number of observations which number is a magnitude higher than the 38981 observations in the previous model application.

The collected dataset contains very similar features to the previously used York dataset because the data collection process was focused to collect similar data. Unfortunately, collecting a complete London scale hourly road traffic was not possible, but all the other data source were available (e.g. Open Street Map database) or similar data could be collected (e.g. using the Open Street Map instead of the Ordnance Survey's Mastermap for building data). There is also an important property of the collected data as these data are all publicly available data:

- the Open Street Map database is an open-source database
- the Londonair database is publicly available to everyone
- the Weather Underground database is free until a certain number of daily queries

| Feature | Unit | Source | Data group |
|---|---|---|---|
| $no_2$ level | $\mu gm^{-3}$ | Londonair | - |
| road_length | meter | Open Street Map | R |
| road_lane_length | meter | Open Street Map | R |
| atc | traffic count/hour | Transport for London | A |
| buildings | - | Open Street Map | B |
| buildings_area | area | Open Street Map | B |
| landuse_area | area | Open Street Map | L |
| leisure_area | area | Open Street Map | L |
| wind_direction | degree (angle) | Weather Underground | W |
| wind_speed | m/s | Weather Underground | W |
| temperature | celsius degree | Weather Underground | W |
| rain | indicator | Weather Underground | W |
| pressure | hPa | Weather Underground | W |
| hour | - | Generated | T |
| day_of_week | - | Generated | T |
| month | - | Generated | T |
| bank_holiday | indicator | Generated | T |

Table 5.1. Summary of the collected data for the large-scale modelling scenario

- Transport for London provides publicly available data for everyone (including the traffic count data at their ATC sites)

The fact that the datasets are publicly available helps to generate reproducible research material as the data is available for everyone and researchers and scientist do not need to wait for special permissions to get the data. Also, this helps for the model application itself as the created data collection and transformation methods can be used to generate data for other modelling areas easily.

## 5.3 Evaluation of the developed statistcial regression methods

The data has been collected for the London modelling area which contains data for 35 stations from various data sources. The data was transformed into the right format (this transformation is essentially the same as the data transformation used for previous model application and evaluation).

The evaluation of the previous model application was carried out by implementing a leave-one-out cross-validation (LOOCV) method where data from one station was left out from the training phase of the statistical regression model and then the generated model was applied to this station data and the predictions were compared with the observations. The process was repeated for each station. The main purpose of this method is to determine the possible prediction error in

the case of applying the model to locations in the urban area unknown to the model and evaluate the average prediction error of the application of the model to the complete modelling area. The previous dataset only has 7 (or 5) stations due to the small size of the monitoring station network (which in fact means a very dense monitoring station network considering the size of the York area). Applying the LOOCV to the York dataset was an ideal choice as data was created for only a small number of monitoring stations (e.g. training the statistical regression model using even smaller number of stations and evaluating the predictions of data on more stations would result in higher error levels as the statistical regression approach would suffer from not having enough data).

LOOCV is a choice to evaluate the prediction accuracy of the statistical regression approaches on the London dataset, however, the dataset enables to create wider evaluation as it is possible to evaluate the prediction accuracy using data not only from one station but from multiple stations. It is still possible to follow the idea of the LOOCV evaluation method, however, not with single station but with a group of stations. The available data from 35 stations can be divided into 5 groups pseudo-randomly in the way that each group contains stations with low, medium and high hourly $NO_2$ concentration level observations, therefore, each iteration of the LOOCV method is going to evaluate data from all range of the stations. This helps to understand better the achievable prediction accuracy of a complete city-scale model application than using the standard LOOCV where only one station data is used for evaluation.

Figure 5.2 shows the groups of stations with their observed hourly $NO_2$ concentration levels. The figure shows that each group has observations from the station having all the range which will give bias to each validation iteration.

It is now possible to execute similar evaluation runs to the previous model evaluations, but the LOOCV is going to leave one group data out from training the statistical regression model and apply the generated model on the data left out from the training and compare the predictions with the observations. Investigating the result of a LOOCV gives us an understanding of

- the scalability of the statistical regression approach as it is possible to measure the time required to run each iteration and compare that time with the previous LOOCV execution time
- the robustness of the statistical regression approach as the validation will provide information about the quality (high level and low level) of the predictions generated by the approaches

To be able to assess the scalability and the robustness of the statistical regression methods, the Random Forest Regression method needs to be tuned for this new regression task. Using the result of the previous model application (Section 3.4), the *minleaf* train induction method was selected. Then similar hyperparameter search runs were executed on this dataset to properly tune the hyperparameters of this tree induction algorithm. The search was set to investigate the

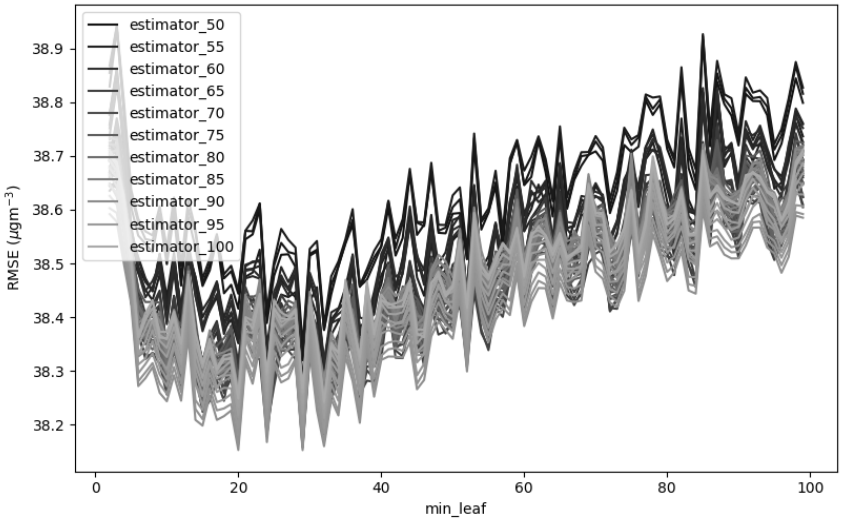



Figure 5.3. Hyperparameter investigation for the Random Forest Regression method using the *minleaf* tree induction technique on the London dataset

minleaf parameter between the value of 2 and 200 and the estimators parameter between the value of 50 and 100. Figure 5.3 shows the high-level RMSE results of this run. The figure shows similar trends to the hyperparameter searches of the previous model application:

- the result shows that the high-level accuracy is, in fact, sensitive to the applied hyperparameter as the high-level RMSE accuracy indicator depends on the given minleaf and estimators hyperparameters
- the RMSE high-level curve reaches its minima at minleaf=29 which suggest that the generated trees have more observations in their leaf nodes compared to the previous model application (where minleaf=2 gave the most accurate regression model) which is in line with the fact that the model now has more data to extract the necessary knowledge to make accurate hourly $NO_2$ concentration level predictions
- the result suggests that the Random Forest Regression approach is robust to the large-scale dataset as the curve reaches its minima and there are no unexpected spikes in the figure

Based on the result of the hyperparameter search, using the minleaf=29 and estimators=64 gives the most accurate model to the large-scale modelling task which model generates predictions with 32.16 $\mu gm^{-3}$ RMSE accuracy. The time required to run one LOOCV run takes 121 seconds on average on a computer with Intel Core i7-4770K processor and 32 GB memory hardware configuration. The same LOOCV run for the previous model application (for the York dataset) took 104 seconds using the same computer. This result indicates that the method, in fact,

requires more time to generate the underlying decision tree models, however, the large-scale model application evaluation can be executed within fairly short time using a desktop computer (therefore it does not require special hardware or a network of computers to carry out the model application). The main purpose of the scalability study of the statistical regression approach is to find out the method is feasible to carry out any large-scale environmental modelling task. The fact that the model application to one of the largest problem in the field takes a couple of minutes on an average desktop computer makes the model scalable and there is no need for further scalability investigation as all the regression tasks in the field (the field of hourly $NO_2$ concentration level modelling in the urban area) has the same (or smaller) problems regression problems.

Similarly to the previous model application and evaluation, tuning the Random Forest Regression approach was executed by feeding all the available data to the algorithm. The results of the previous chapter indicate that further accuracy improvement can be achieved by not using all the available features of the input dataset, however, it is not known that this behaviour still holds in the case of using the large-scale dataset. To understand the accuracy sensitivity of the Random Forest Regression approach to the applied data, the evaluation framework was executed using the all the possible subset of the input data by grouping the features by their data source. The evaluation framework was executed the determine the high-level RMSE accuracy for all the possible combination of the features collected from different data sources, similarly to the previous model evaluation (Section 4.2). Figure 5.4 shows the result of the high-level RMSE accuracy analysis of the input data analysis. The result is similar to the previous model evaluation:

- using all the available data gives a model (RFR+ALL) that generates the hourly $NO_2$ concentration level predictions with 38.16 $\mu gm^{-3}$ high-level RMSE accuracy
- the most accurate model is the model which using only the time and weather-related data (RFR+TW) generating a regression model which creates predictions with 31.88 $\mu gm^{-3}$ high-level RMSE accuracy
- adding the automated traffic count data to the time and weather-related data and using this dataset to feed the Random Forest Regression with training data generated a model with increased error rate (34.73 $\mu gm^{-3}$ high-level RMSE accuracy) compared to the previous (time and weather-related data only) case similarly to model evaluation for the York dataset

Similarly to the previous model evaluation, the relative RMSE high-level accuracy was visualized for using all the data subsets relative to the most accurate case (RFR+TW) (Figure 5.5). The figure indicates that the result is similar to the previous model application and evaluation as using the time and weather-related data generates the most accurate model to predict the hourly $NO_2$ concentration levels for the large-scale modelling task. This result, however, is expected as the same result for model evaluation of the York dataset as in this case the weather data is different at each station as the data source provided different meteorological data for each station.

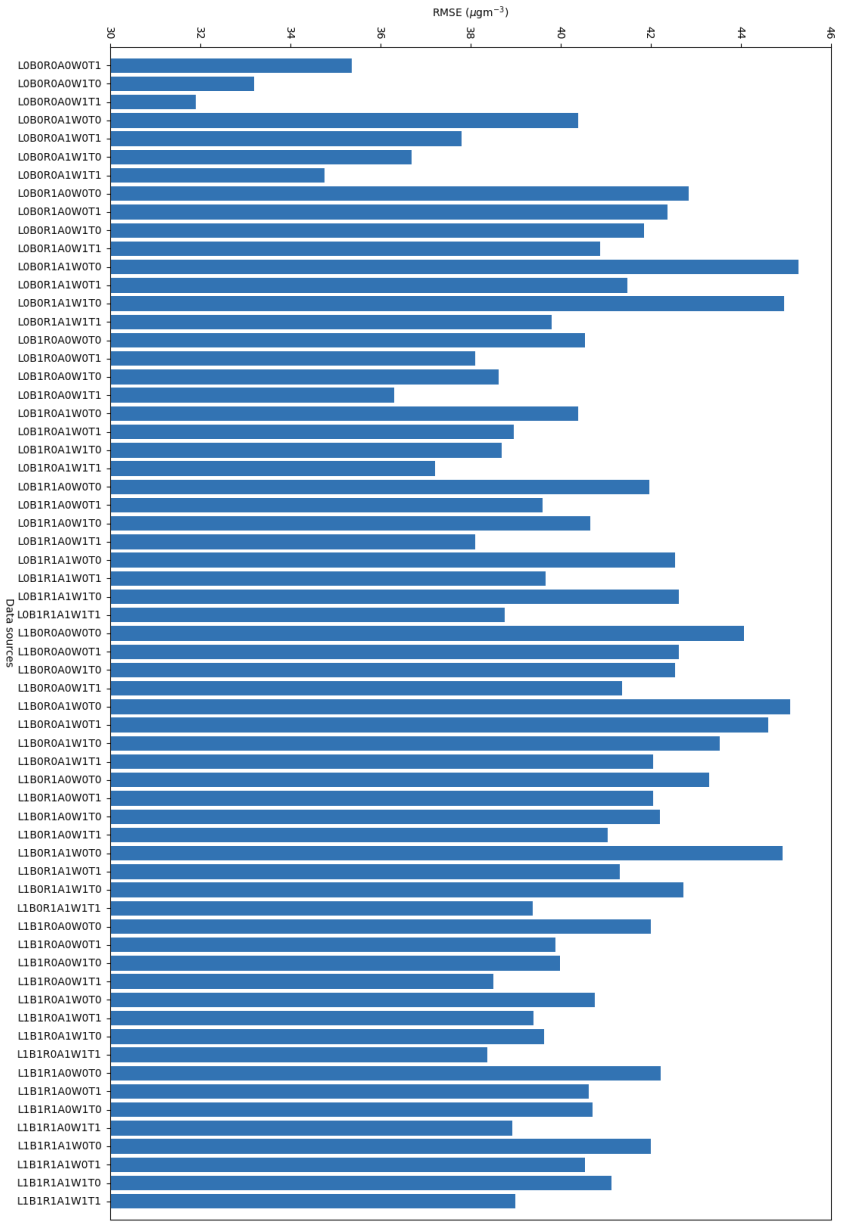



Figure 5.4. Accuracy investigation of the different input data subsets using the same labelling as the previous model evaluation

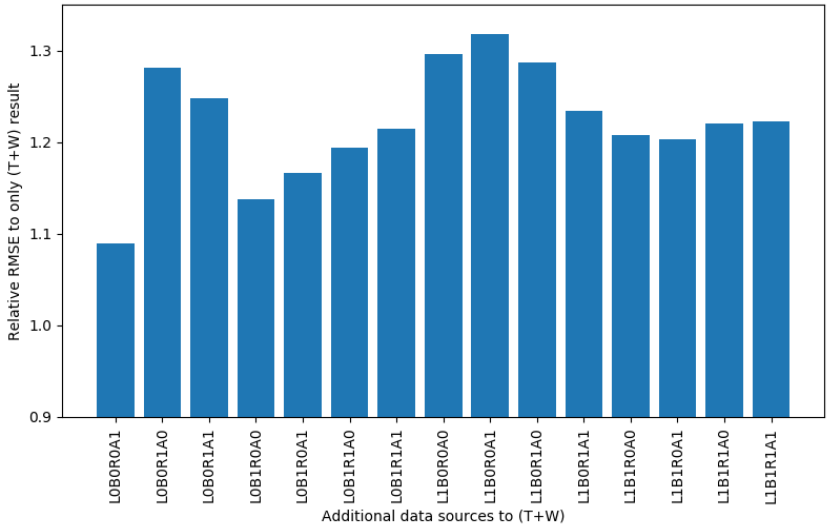



Figure 5.5. Relative RMSE accuracy using datasets compared to RFR method using only the Time and Weather data

Adding the automated traffic count data introduced more error to the predictions as indicated by the experiment which fact opens the way to the application of the developed Random Forest ensemble method. The previously developed ensemble method, however, was considering only individual stations, therefore, a simple modification had to be applied to be able to run on the large-scale dataset:

- the current LOOCV evaluation framework utilizes data from 4 groups of stations (compared to the previous model application where only data from stations was utilized) to build the statistical regression model and the framework applies the model and evaluates the accuracy of the model on data of the fifth group of stations (compared to the previous model application where only data from the remaining fifth station was used), and repeats this process 4 more times to apply the model on all the five groups of stations
- the model selection requires data to train a classification model; this data includes the input data and the concentration level observations and prediction output of the RFR+TW and RFR+TWA models
- the model combination method use only data from 3 groups of stations to train to RFR+TW and RFR+TWA models and applies it to data of the fourth group to generate the required concentration level predictions for the classification
- then based on the predictions given by the RFR+TW and RFR+TWA models, it assigns the value of 0 (prediction concentration level by the RFR+TW model is closer to the observed

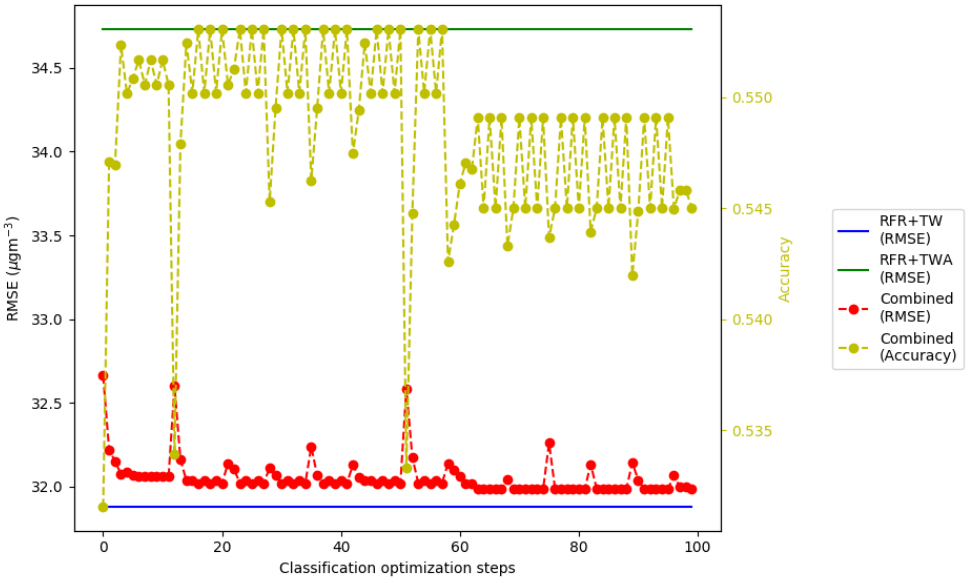



Figure 5.6. Classification feature optimization steps for the Random Forest ensemble method

concentration level than the prediction concentration level by the RFR+TWA model) or 1 (prediction concentration level by the RFR+TWA model is closer to the observed concentration level than the prediction concentration level by the RFR+TW model) which provides the two classes for the classification

- this process is repeated four times to generate data for each group of stations
- the Random Forest Classifier is trained based on the generated data
- RFR+TW and RFR+TWA models are trained using the data of the available 4 stations
- RFR+TW and RFR+TWA models are applied to the fifth group as well as trained model selection classifier
- based on the output of the model selection classifier (it is either 0 or 1), the prediction output of the RFR+TW (if the classification output is 0) or the prediction output of the RFR+TWA (if the classification output is 1) will be selected for the final concentration level prediction
- the complete process is repeated 4 more times to cover all 5 possible iterations

The classification method needs to be calibrated for this regression task (similarly to the previous ensemble method application). Figure 5.6 shows the result of the feature optimization technique which process helps to calibrate the classification method. The result indicates that the using the subset of the features to carry out the classification helps, however, the overall

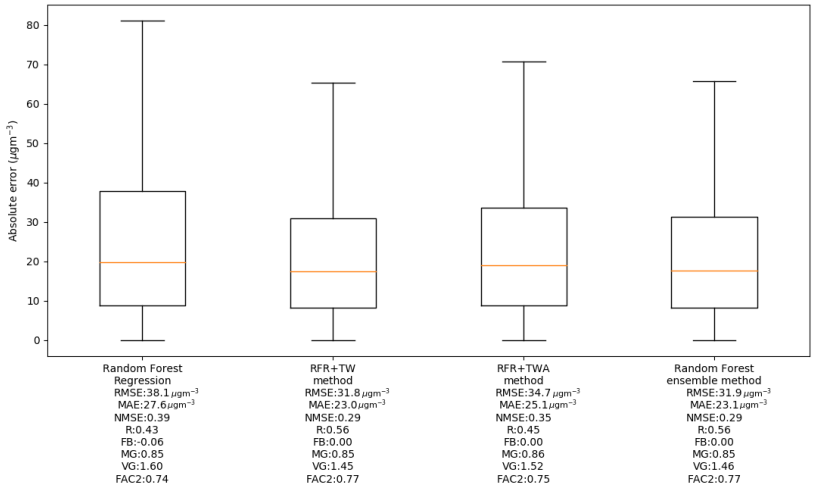


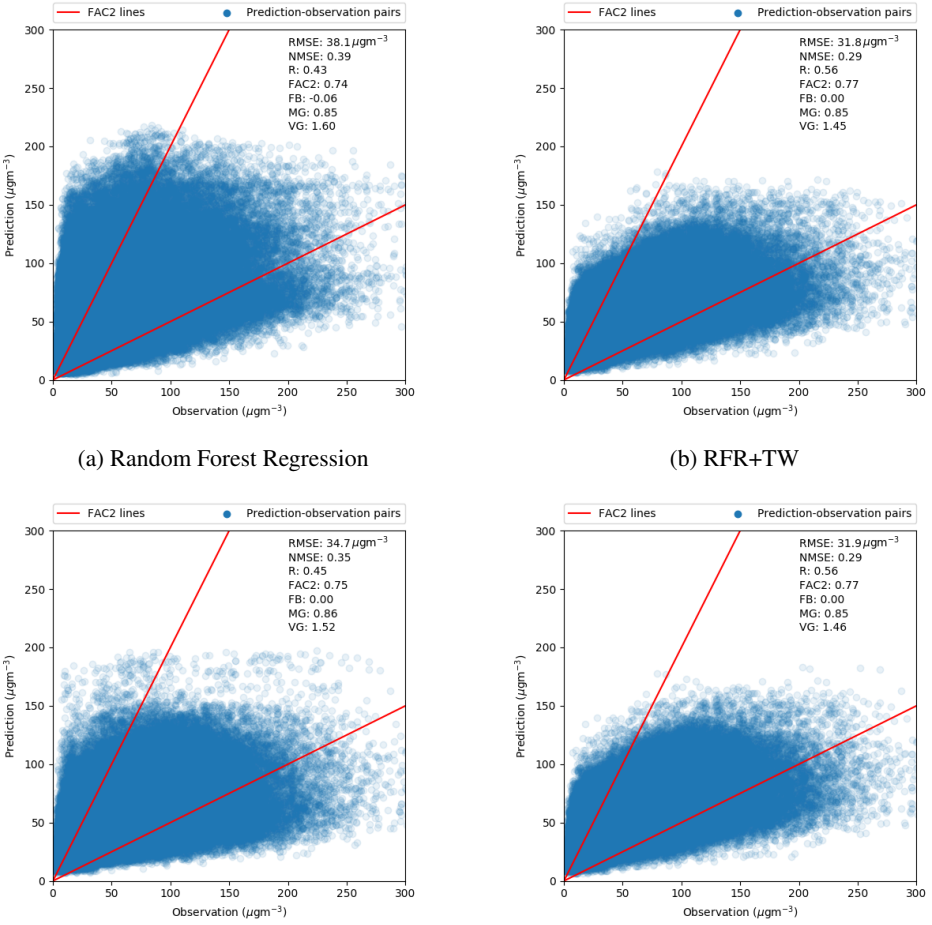

Figure 5.7. Boxplot of the absolute error for the RFR+ALL, RFR+TW, RFR+TWA and Random Forest ensemble methods

RMSE high-level error of 31.94 $\mu gm^{-3}$ for the Random Forest ensemble method on this dataset indicates that the ensemble struggles to utilize the RFR+TW and RFR+TWA methods to generate more accurate predictions than the RFR+TW model itself. There are multiple reasons for this result:

- traffic is not the only pollution source in the London modelling area, therefore, using the data gives information about one of the pollution source, but not all of them compared to the previous York modelling scenario
- the meteorological data in the York dataset contains observations from one single weather monitoring station and this data has been used at all the different pollution monitoring stations locations. On the other hand, the meteorological data in the London dataset contains weather observation data from multiple weather observation stations because the modelling area consists of multiple weather observation stations. This implies that the weather-related input data not necessarily the same at each pollution monitoring station locations for the same observation time.

To compare the three different approaches (RFR+ALL, RFR+TW, Random Forest ensemble), Figure 5.7 shows the boxplot of the absolute error of hourly $NO_2$ concentration level predictions and Figure 5.8 shows the observation-prediction pairs for all three models:

- the Random Forest Regression method using all the available data (RFR+ALL) produced the most inaccurate model as the model generated predictions with 38.16 $\mu gm^{-3}$ RMSE,




(a) Random Forest Regression

(b) RFR+TW

(c) RFR+TWA

(d) Random Forest ensemble

Figure 5.8. Observation-prediction plots for different methods on the London dataset

27.16 $\mu gm^{-3}$ MAE high-level errors and with 0.43 linear correlation value. This result is similar to the previous model application.

- the Random Forest Regression method using only the time and weather-related data (RFR+TW) produced a more accurate regression model as the model generated predictions with 31.88 $\mu gm^{-3}$ RMSE, 23.03 $\mu gm^{-3}$ MAE high-level errors and with 0.56 linear correlation value. Figure 5.7 shows smaller absolute prediction errors and Figure 5.8 shows tighter point cloud for the observation-prediction pairs which results are all in line with results of the previous model application.

- the Random Forest ensemble method which utilizes the RFR+TW and RFR+TWA models produced a regression model with 31.94 $\mu gm^{-3}$ RSME, 23.12 $\mu gm^{-3}$ MAE high-level accuracy and with 0.56 linear correlation value. The ensemble method struggled to effectively combine the predictions of the two underlying methods, therefore, the overall accuracy is lower than the accuracy of previous RFR+TW model. This result is different from the previous model application as using the Random Forest ensemble method gave accuracy improvement for the York dataset.

Comparing the high-level RMSE and MAE results to the previous model application suggests that the models struggle to make accurate predictions as the values of the RMSE and MAE levels are higher:

- the RFR+ALL models produced predictions with accuracy of 15.06 $\mu gm^{-3}$ RMSE and 38.16 $\mu gm^{-3}$ RMSE on the York and London dataset, respectively

- the RFR+TW models produced predictions with accuracy of 12.68 $\mu gm^{-3}$ RMSE and 31.88 $\mu gm^{-3}$ RMSE on the York and London dataset, respectively

- the Random Forest ensemble methods produced predictions with accuracy of 12.64 $\mu gm^{-3}$ RMSE and 31.94 $\mu gm^{-3}$ RMSE on the York and London dataset, respectively

MAE and RMSE are high-level prediction accuracy evaluation methods producing zero level for the perfect regression model, however, they both depend on the ranges of the regression target data. For example, a regression task where the regression target data range is between 0.0 and 1.0 has smaller MAE and RMSE values than a regression task where the target data range is between 0.0 and 100.0, because, the larger range gives more chance for larger individual prediction error.

To understand the difference between the calculated high-level RMSE values for the different model applications, the analysis of the differences between the hourly $NO_2$ concentration levels has to be carried out. The York and the London datasets are describing two very different regression tasks:

- York has the traffic as the main pollution source and has 5 (or 7) monitoring stations deployed to capture the $NO_2$ concentration levels which stations observed very similar processes affecting the concentration levels
- London has multiple pollution sources and the monitoring station network captures very different concentration level trends depending on the location of the monitoring station (e.g. the monitoring station deployed at Oxford street has very high-level of $NO_2$ concentration levels as this street has the worst pollution in the world while a monitoring station in one of the outer region observed low concentration levels similar to what the York dataset has)

Both RMSE and MAE measurements are relative to the actual observations, therefore, it is expected to have higher RMSE and MAE values on a regression task which has higher observation values simple because a misprediction can cause a higher absolute error on average, considering the complete regression task. This explains why the model applied to the London dataset produced higher RMSE and MAE high-level errors in general, however, the observation-prediction pairs and absolute error plot are required to be investigated and analysed for any outliers and anomalies in the predictions.

As the London modelling area contains an increased number of stations compared to the previous model application (35 stations compared to the 5 stations), the observation-prediction pair plot (Figure 5.8) and the prediction absolute error plot (Figure 5.7) have less meaningful information because the figures contain too many data points to visualize. These figures give an overall view of the quality of the predictions, but they do not provide information about anomalies and outliers in the predictions at the station level. To understand the prediction quality of the developed models, the visualization of prediction absolute errors were generated at each station:

- the data has been extracted from each iteration of the LOOCV process
- in each iteration, the prediction and observation data were captured and identified to match them to the corresponding station

Figure 5.9 shows the absolute prediction error at each station by the RFR+TW method. It indicates that there is no significant anomaly in the prediction data and the models generate prediction sensibly at each station. Investigating the same figures for the RFR+ALL and Random Forest ensemble method gives the same behaviour. The figure, however, reveals another unexpected property of the statistical regression approaches:

- the medians of the absolute errors of the predictions do not straightly follow the medians of the observations
- the medians of the absolute prediction errors reach the lowest values between the station 57 and station 50

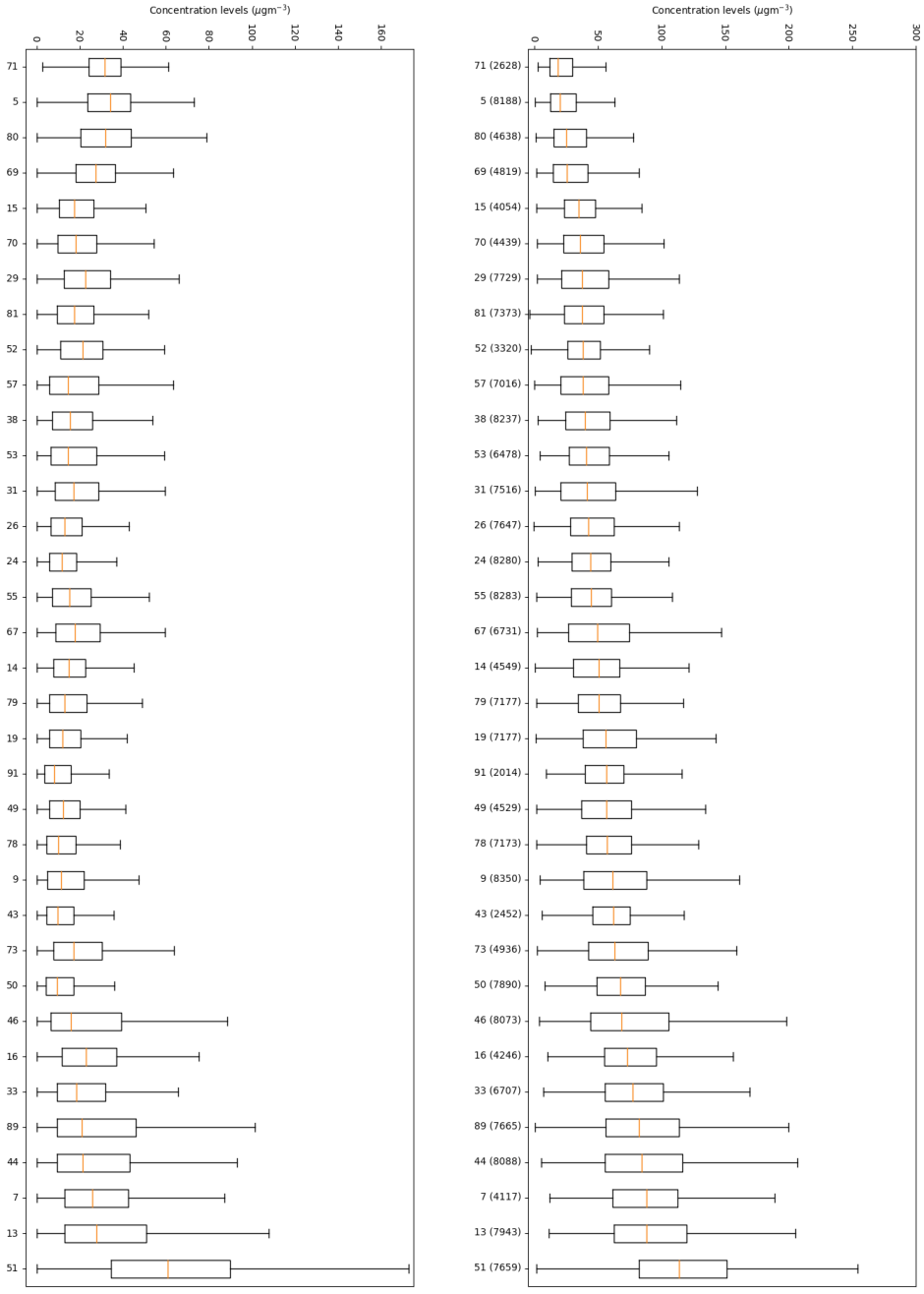



Figure 5.9. Absolute prediction errors by the RFR+TW model grouped by the stations and ordered by the median of the concentration level observation of the stations

- the medians are increasing leaving this middle section of the figure (left to the station 57 and right to the station 50)

- this suggests that the lowest absolute prediction errors are not presented at the station with the smallest observed hourly $NO_2$ concentration levels, but the stations which are close to the middle region which stations data are closer to an hypothetical average station data

Figure 5.9 indicates that the absolute prediction errors are smaller at the stations which has an average hourly $NO_2$ concentration level observations considering all the hourly $NO_2$ concentration level observations of all the available stations. This suggests that the Random Forest methods (RF+ALL, RF+TW, Random Forest ensemble) are generating accurate predictions considering an average monitoring station (a station which has an average hourly $NO_2$ concentration levels) and accuracy of the predictions degrades if the model needs to predict concentration levels at a place which has lower or higher average $NO_2$ concentration levels. This behaviour is the consequences of the internal tree induction mechanism of the Random Forest statistical regression algorithm as the tree induction algorithm is creating the internal tree to minimalize the mean squared error and the mean squared will be minimum if the model gives very accurate predictions at the stations where the concentration levels are close to the average.

This finding motivates the investigation of a different kind of ensemble method where multiple models are trained on different subsets of the available data based on the station's observations levels and the right model is selected to generate more accurate hourly $NO_2$ concentration levels.

## 5.4 Ensemble model for large-scale environmental modelling

Visualizing the absolute prediction errors at each station revealed that the Random Forest Regression methods produce accurate predictions at stations whose observations are close to the average and the methods produce uncertain predictions at stations either with low $NO_2$ concentration level profile or high $NO_2$ concentration level profile. To further investigate this property of the Random Forest Regression algorithm, the following visualization has been generated:

- a Random Forest Regression method trained using only one station's data (hourly $NO_2$ concentration levels and the time and weather-related part of the available input data), and the model applied to all the stations individually and evaluated using the RMSE high-level error metric

- The visualization of this experiment helps to understand the achievable average RMSE high-level prediction accuracy using models trained on observation data in one range and applying the generated model to a similar (and different) observation range

Figure 5.10 shows the visualization of the results which shows the following trend:

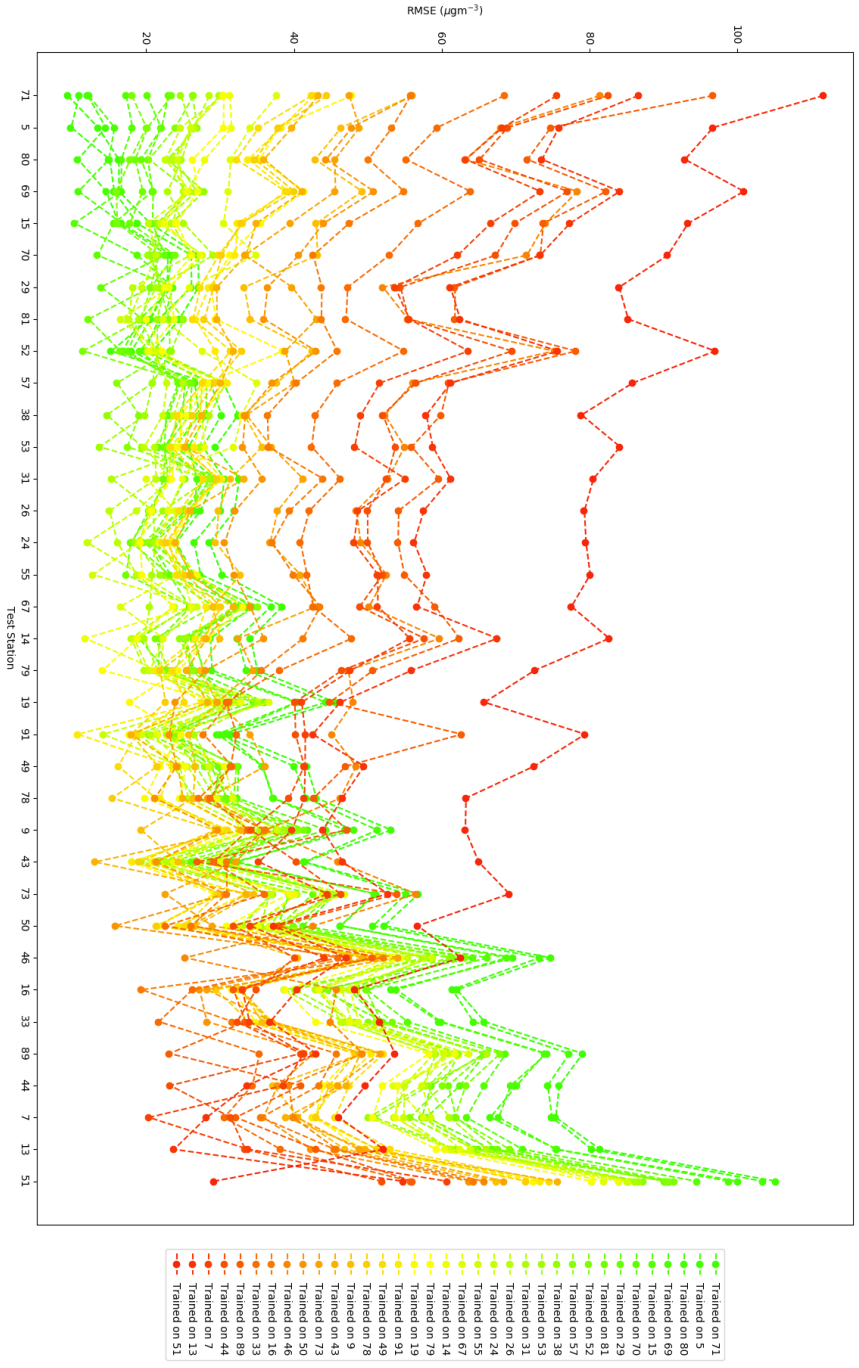



Figure 5.10. Error analysis of the Random Forest Regression models generated by using only one single station data and evaluated on all stations individually where the colour of the line indicates the concentration level profile observed by the single station (green has low concentration levels while red has high concentration levels)

- a Random Forest Regression model makes accurate predictions if it is applied to a station which has a similar $NO_2$ concentration level profile as the station which provided the data initially for the Random Forest Regression model

- this observation is true for all the spectrum of the stations as using the data of stations with low concentration levels observations is generating a model which is accurate prediction concentration levels of station with low concentration levels and the opposite is also true

- this information describes why the Random Forest Regression algorithm using observation data contains only low concentration levels would make predictions with low accuracy (as it has the opportunity to learn low concentration levels). The opposite case, however, is unexpected because using the Random Forest Regression model trained on observation data which contains high (but not always high as there are low concentration level observations at these monitoring stations) concentration level observations gives predictions on low concentration level observations with low accuracy.

This result reveals the nature of the prediction of the Random Forest Regression statistical regression algorithm as it can only predict events (concentration levels) that the model observed during the model training phase. This property implies that the input data needs to contain observation from all range of monitoring stations to give the right data to the Random Forest Regression algorithm otherwise it will generate predictions with high error levels.

This result also indicates that using two Random Forest Regression models (one trained on data from stations which observed low concentration levels and the other trained on data from stations observed high concentration levels) can potentially improve the accuracy of the statistical regression approach if the model selection can be implemented accurately. The ranges should not overlap in these models, otherwise, the process would give models where the models are less observation range specific, therefore they would give predictions similarly to the single model case. The model selection can be implemented as a binary classification task where the classification method needs to select the appropriate output of the available two models considering the current input. This classification can be carried out by using the Random Forest classification algorithm. Using this algorithm, the ensemble method needs to contain the following steps to be able to train and evaluate the model:

- the current LOOCV evaluation framework utilizes data from 4 groups of stations to build the statistical regression model and the framework applies the model and evaluates the accuracy of the model on data of the fifth group of stations (and repeats this process 4 more times to apply the model on all the five stations)

- the available data of 4 groups of stations used for training the regression model needs to be split into two parts based on the source stations concentration level profile: one group

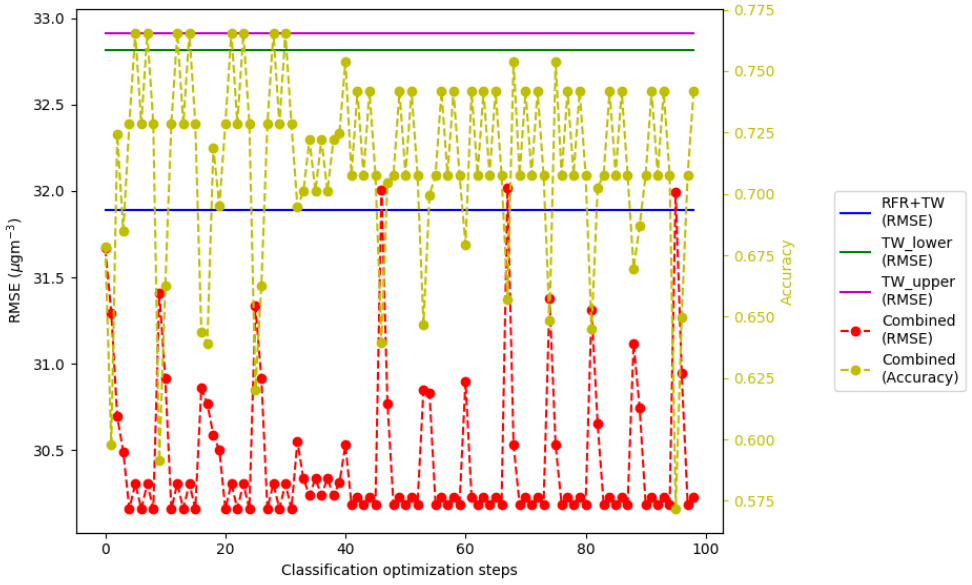



Figure 5.11. Stepwise feature optimization for the large-scale Random Forest ensemble method

contains the data from the lower part (exactly half of the available stations) and the other group contains the data from the upper part (the other half of the available stations)

- using the current London dataset and the 5-fold LOOCV method gives data of 14 stations for the lower model and data of 14 stations for the upper model
- the Random Forest Regression methods trained based on the lower and upper datasets
- another dataset is created for the classification which includes the original data from the available 28 stations (excluding the observation data) and adding one new feature which describes that the given observation belongs to the lower or to the upper datasets
- a Random Forest classification model trained on this new dataset to be able to decide which outputs to use for the final output generation
- all three models are applied to data from the fifth groups of stations (and the model selection is based on the output of the Random Forest classification method)
- the complete process is repeated 4 more times to cover all 5 iterations

This ensemble method contains a Random Forest classification method to automatically choose from the two outputs of two Random Forest Regression methods (RFR upper, RFR lower). This classification method, however, needs to be optimized as the overall prediction accuracy will depend on the classification accuracy of the model selection classification model. To optimize the classification method, a stepwise feature optimization technique was carried out

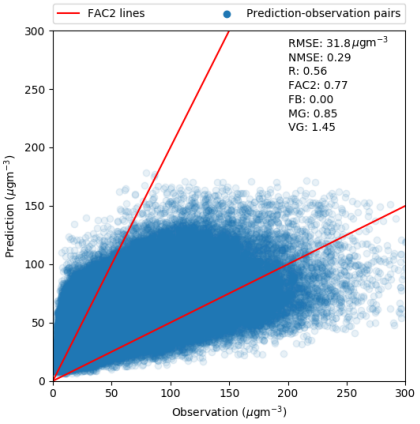
(similar to the previous stepwise feature optimization methods). Figure 5.11 shows the result of the stepwise feature optimization method for the proposed Random Forest ensemble method. The optimization follows the same process as the feature optimization of the classification of the previous Random Forest ensemble method as the high-level RMSE accuracy improves with the increase of the accuracy of the model selection classification. The feature optimization finds the global optima after 6 steps which solution includes the following features: building_area, natural_area, leisure_area, landuse_area, lane_length, wind_speed, wind_direction, rain, temperature, pressure, hour, month, bank_holiday. Again, the classification uses not just the weather and time-related data, but data from all the data sources to make the model selection process more accurate (therefore to increase the high-level RMSE accuracy of the Random Forest ensemble method) which indicates that the ensemble model generates prediction with good spatial variance as the concentration level prediction depends on all the features, not just time and weather-related features. The proposed Random Forest ensemble method generated a statistical regression model generates prediction more accurately than the single RFR+TW model:

- the Random Forest Regression method using only time and weather-related data generated a model which gives predictions with 31.88 $\mu gm^{-3}$ RMSE accuracy
- the Random Forest Regression method using only the data from stations with the low concentration level profiles generated a model with 32.81 $\mu gm^{-3}$ RMSE accuracy
- the Random Forest Regression method using only the data from stations with the high concentration level profiles generated a model with 32.91 $\mu gm^{-3}$ RMSE accuracy
- the feature optimized Random Forest ensemble method (which ensembles the RFR lower and RFR upper models) generated predictions with 30.09 $\mu gm^{-3}$ RMSE accuracy

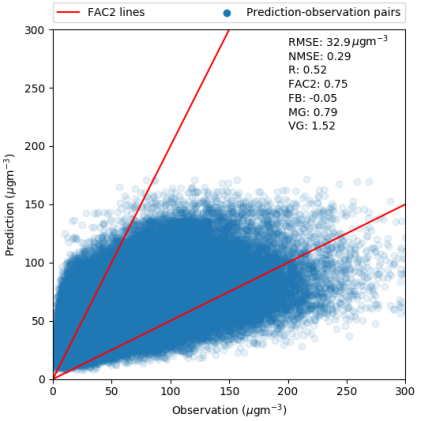
## 5.5 Summary

The aim of this chapter is to understand the scalability and the robustness of the developed statistical regression approach for the hourly $NO_2$ concentration level predictions. The large-scale modelling scenario was introduced in this chapter which task provided the opportunity to investigate the scalability and robustness of the method.

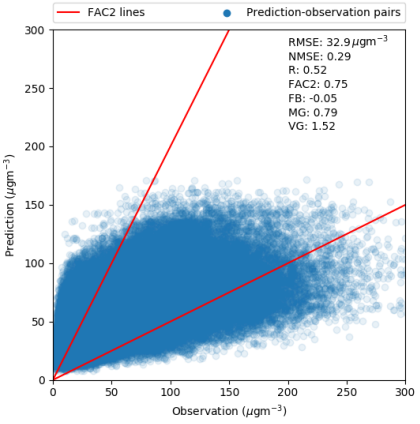
The analysis of the developed Random Forest Regression technique revealed that the statistical regression approach is robust to a large-scale environmental modelling task as it provided good high-level prediction accuracy levels. The Random Forest Regression method provided similar behaviour as the previous modelling scenario as it provided the most accurate model using only the time and weather-related data. The analysis of the developed Random Forest ensemble method indicated that the ensemble technique does not work on this large-scale modelling task as it provided predictions with less accuracy.



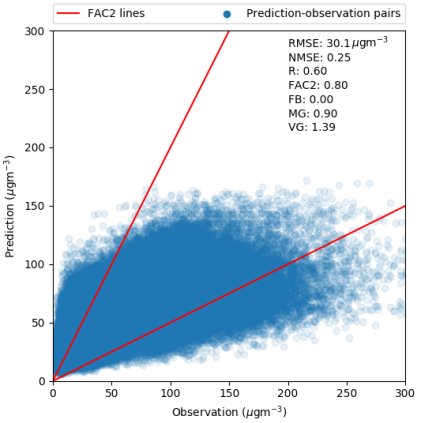

(a) RFR+TW




(b) RFR+TW trained on low concentration observations




(c) RFR+TW trained on high concentration observations




(d) Random Forest ensemble

Figure 5.12. Observation-prediction plots for different methods on the London dataset

Further investigation of the developed Random Forest Regression and Random Forest ensemble methods revealed a new property of the Random Forest Regression algorithm: it provides accurate hourly $NO_2$ concentration levels for a station which provided observations close to the average concentration levels considering all the observations by all the stations. This finding motivated the development of a different ensemble model where the ensemble method combines different Random Forest Regression models trained on different parts of the available data. The developed novel ensemble method generates more accurate (by using all the introduced accuracy evaluation metrics) hourly $NO_2$ concentration level predictions than the underlying Random Forest Regression models. Using the developed model helps to understand the pollution better in very complex modelling area such as London, because it can produce concentration level predictions with less error.

As this is the final chapter which provides technical work, the next chapter will summarize all the contribution of this research and propose possible future work related to the application of statistical regression methods to hourly $NO_2$ concentration level predictions for the urban area.

# CHAPTER 6

# Conclusion and future work

To conclude the work in this thesis, the hypothesis is stated as follows:

> *Through the appropriate ensembling of state of the art statistical regression methods, a more accurate, robust and scalable high-temporal environmental model can be created than the current state-of-the-art air pollution dispersion techniques*

The work presented in this thesis demonstrated that the prediction error of the current state-of-the-art air pollution dispersion modelling technique can be reduced by applying statistical regression ensemble technique for the urban-scale hourly $NO_2$ concentration level predictions. The developed novel statistical regression model in Chapter 3 generated more accurate predictions than the current state-of-the-art air dispersion model by evaluating all the introduced accuracy evaluation metrics. The introduced ensemble method generates even more accurate predictions in Chapter 4 by all the evaluation methods. The developed approach has been applied to a large-scale regression task in Chapter 5 and the results indicates that it makes good prediction accuracy on the hourly $NO_2$ concentration level prediction regression task.

## 6.1 Summary of the contribution

The contributions that have been presented in this thesis is summarised as follows:

**Evaluation framework for urban-scale hourly $NO_2$ concentration level predictions**

The work presented in Chapter 3 introduced the evaluation framework to measure the prediction accuracy of the different approaches including one of the current state-of-the-art air pollution dispersion models and the existing Land Use Regression approaches. The result of this work

indicates that the existing Land Use Regression approaches struggle to make accurate predictions on the hourly level. This result contributes to the Environmental Science field as it describes the difficulties of application of the existing models to the high-temporal $NO_2$ concentration level predictions.

**Advanced statistical regression method for the high-temporal environmental modelling problem**

The rest of the work presented in Chapter 3 focused on using advanced statistical regression algorithms to solve the given regression task more accurately. Using the developed evaluation framework, the sensitivity analysis of the hyperparameters of the algorithms were investigated and the most accurate algorithm was selected for this regression problem. Again, this work contributes to the Environmental Science field as it describes the efficient application of existing statistical regression algorithms.

**Prediction accuracy sensitivity study of the input data for the statistical regression approach**

Chapter 4 described the accuracy sensitivity analysis of the input data for the Random Forest Regression method. It highlighted that it is necessary to investigate the input data for a given statistical regression task for the Random Forest Regression method as using the appropriate data can increase the prediction accuracy of the model. This work contributes to the Environmental Science field (as it shows that which data is important for a Random Forest Regression based statistical regression approach) and to the Computer Science field (as it gives a systematic way of investigating the sensitivity of the Random Forest Regression method to the input data).

**Random Forest ensemble technique for more accurate hourly $NO_2$ concentration level prediction**

The second part of Chapter 4 investigated the prediction differences of the Random Forest Regression method trained on different subsets of the available input features. The investigation revealed the non-overlapping error episodes in the prediction which suggests that the effective combination of the models can provide accuracy improvement for the overall prediction task. A novel Random Forest ensemble method was proposed to utilize multiple Random Forest models and the method was evaluated and compared against the existing Random Forest Regression method. The development of this novel method contributes to the Computer Science field as the proposed method is a general regression algorithm which can be applied to any regression task to improve the overall prediction accuracy.

**Scalability and robustness analysis of the developed statistical regression method**

The work presented in Chapter 5 covered the scalability and robustness analysis of the developed Random Forest Regression and Random Forest ensemble methods by evaluating them on a large-scale environmental modelling scenario. The result of the evaluation suggests that the developed

methods are scalable and robust to large modelling scenarios despite the high computational requirement of the developed methods. The evaluation also revealed that the Random Forest ensemble fails to make more accurate predictions than the Random Forest Regression. These results contribute to the Environmental Science field as it indicates that the developed statistical regression approaches (including the Random Forest Regression and the Random Forest ensemble methods) can accurately predict $NO_2$ concentration levels even for large-scale and complex modelling scenarios without having issues with the high computational requirements of the underlying regression algorithms.

**Random Forest ensemble technique for more accurate large-scale $NO_2$ concentration level prediction**

The second part of Chapter 5 presented a different Random Forest ensemble method to effectively combine Random Forest Regression models trained on different subsets of the input data partitioned by data of the concentration level profile of the input monitoring stations. This method provided more accurate hourly $NO_2$ concentration level predictions than the previously evaluated statistical regression methods on the large-scale regression task. The development of this method contributes to the Computer Science field as it provides another ensemble method which method can be used to any regression task to further increase the accuracy of the predictions by utilizing multiple Random Forest models and efficiently combining them.

## 6.2 Limitations

Despite the contribution listed in the previous section, the work presented in this thesis does have limitations. The most significant limitations are discussed in this section.

**Data requirement of the statistical regression approach**

In Chapter 4, the sensitivity of the statistical regression approach to the input data was investigated. The result of the analysis indicates that the weather and time-related data can provide enough information to the Random Forest Regression method to generate a regression model which can accurately predict the hourly $NO_2$ concentration levels. However, historical observation is required for the algorithm to generate the internal regression model and this data needs to be collected for every model application scenario. On the other hand, air pollution dispersion methods provide established models which can be applied even on virtual data, therefore, the air pollution dispersion approach can provide some understanding without having the historical observations for the given modelling problem (e.g. rough estimates of pollution levels without actually having any pollution level observations).

**Limitation of the underlying statistical regression algorithm**

In Chapter 5, the large-scale evaluation of the developed Random Forest Regression method revealed that the underlying Random Forest statistical regression method can generate accurate

predictions for scenarios that were observed by the model during the training phase of the regression model. This indicates that the model will ever be accurate if the available input training data for statistical regression method covers all type of scenarios within the modelling area. An extreme example would be a case where the data points with heavy traffic are excluded from the input dataset, therefore, the model never observes such cases.

## 6.3 Future work

The last section of the thesis provides details about the future work related to the developed statistical regression approaches.

**Using future statistical regression algorithms to solve the hourly concentration level predictions more accurately**

The work presented in this thesis introduced efficient statistical regression methods for accurate hourly $NO_2$ concentration levels prediction. The most accurate existing statistical regression algorithm (the Random Forest Regression algorithm) was selected in Chapter 4 from many advanced statistical regression algorithms. The set of algorithms was selected based on a literature survey where studies were solving similar environmental problems with these algorithms. As the machine learning field is progressing forward, new algorithms will be developed to provide solution for the regression task, therefore, these algorithms can solve the given regression task more accurately than the most accurate model, the Random Forest Regression of this thesis (e.g. since the beginning of the work presented in this thesis, there are new approaches such as the boosted trees [Chen & Guestrin (2016)] and Gaussian process regression [Gal et al. (2014)]).

**Using deep-learning technique to efficiently ensemble models**

Chapter 4 and Chapter 5 introduced ensemble methods to further improve the prediction accuracy by exploiting the predictions of the different Random Forest Regression methods. These different Random Forest Regression methods were trained on different subsets of the input data and then combined using the Random Forest classification algorithm. This model combination flow suggests that the ensemble can be carried out using recently developed deep-learning techniques [LeCun et al. (2015); Gal & Ghahramani (2016); Qiu et al. (2014)] where a large number of machine learning models are connected together to solve the underlying problem more accurately and efficiently than a single model.

**Providing prediction data for exposure studies**

The work presented in this thesis provided novel statistical regression approaches for hourly $NO_2$ concentration level predictions. The accurate high-temporal large-scale predictions can give a new insight for Environmental Scientists to understand the pollution behaviour in the urban area by applying the methods to the complete modelling area. There are well-established methods to understand the health effect of low-temporal pollution [Cyrys et al. (2005); Cesaroni

et al. (2013)] exposure, however, interpreting the high-temporal pollution levels is challenging and new methods need to be developed to be able to understand the high-temporal dynamics of the pollution and the health effect of this high-temporal dynamics.

## 6.4 Final words

The work presented in this thesis indicates that machine learning algorithms can help to predict the air pollution concentration levels accurately on the high-temporal resolution by only providing historical observation data. In the data-driven future, data is likely to be the new gold standard and methods like the ones presented in this thesis will even further exploit the hidden knowledge. Using better more and better data will provide practical alternative methods to the current state-of-the-art techniques.

## 6.5 Availability of Source Code

The Python source code of generating all the research material including the figures of this thesis is available under the GNU General Public License version 3 and can be downloaded from https://github.com/gabormakrai/landuseregression.

# Bibliography


Alam, M. S. & McNabola, A. (2015). Exploring the modeling of spatiotemporal variations in ambient air pollution within the land use regression framework: Estimation of PM10 concentrations on a daily basis. *Journal of the Air & Waste Management Association*, *65*(5), 628–640.

Allwine, K. J. & Flaherty, J. E. (2006). Joint Urban 2003: Study overview and instrument locations. *Pacific Northwest National Laboratory Rep. PNNL-15967*.

Allwine, K. J. & Flaherty, J. E. (2007). Urban dispersion program overview and MID05 field study summary. *Pacific Northwest National Laboratory Rep. PNNL-16696*.

Allwine, K. J., Shinn, J. H., Streit, G. E., Clawson, K. L., & Brown, M. (2002). Overview of URBAN 2000: A multiscale field study of dispersion through an urban environment. *Bulletin of the American Meteorological Society*, *83*(4), 521–536.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, *46*(3), 175–185.

Berkowicz, R. (2000). OSPM - A parameterised street pollution model. In *Urban Air Quality: Measurement, Modelling and Management* (pp. 323–331). Springer.

Berkowicz, R., Ketzel, M., Jensen, S. S., Hvidberg, M., & Raaschou-Nielsen, O. (2008). Evaluation and application of OSPM for traffic pollution assessment for a large number of street locations. *Environmental Modelling & Software*, *23*(3), 296–303.

Brauer, M., Hoek, G., van Vliet, P., Meliefste, K., Fischer, P., Gehring, U., Heinrich, J., Cyrys, J., Bellander, T., Lewne, M., et al. (2003). Estimating long-term average particulate air pollution concentrations: application of traffic indicators and geographic information systems. *Epidemiology*, *14*(2), 228–239.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

Briggs, D. J., Collins, S., Elliott, P., Fischer, P., Kingham, S., Lebret, E., Pryl, K., Van Reeuwijk, H., Smallbone, K., & Van Der Veen, A. (1997). Mapping urban air pollution using GIS: a

regression-based approach. *International Journal of Geographical Information Science*, *11*(7), 699–718.

Briggs, D. J., de Hoogh, C., Gulliver, J., Wills, J., Elliott, P., Kingham, S., & Smallbone, K. (2000). A regression-based method for mapping traffic-related air pollution: application and testing in four contrasting urban environments. *Science of the Total Environment*, *253*(1), 151–167.

Britter, R. & Hanna, S. (2003). Flow and dispersion in urban areas. *Annual Review of Fluid Mechanics*, *35*(1), 469–496.

Carruthers, D., Blair, J., & Johnson, K. (2003). Validation and Sensitivity Study of ADMS-Urban for London. *Environmental Research Consultants, Cambridge*.

Carruthers, D., Holroyd, R., Hunt, J., Weng, W., Robins, A., Apsley, D., Thompson, D., & Smith, F. (1994). UK-ADMS: A new approach to modelling dispersion in the earth's atmospheric boundary layer. *Journal of wind engineering and industrial aerodynamics*, *52*, 139–153.

Cesaroni, G., Badaloni, C., Gariazzo, C., Stafoggia, M., Sozzi, R., Davoli, M., & Forastiere, F. (2013). Long-term exposure to urban air pollution and mortality in a cohort of more than a million adults in Rome. *Environmental Health Perspectives*, *121*(3), 324.

Champendal, A., Kanevski, M., & Huguenot, P.-E. (2014). Air pollution mapping using nonlinear land use regression models. In *International Conference on Computational Science and Its Applications*, (pp. 682–690). Springer.

Chang, J. & Hanna, S. (2004). Air quality model performance evaluation. *Meteorology and Atmospheric Physics*, *87*(1-3), 167–196.

Chen, T. & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 785–794). ACM.

Christensen, J. H. (1997). The Danish eulerian hemispheric model-A three-dimensional air pollution model used for the Arctic. *Atmospheric Environment*, *31*(24), 4169–4191.

Cimorelli, A. J., Perry, S. G., Venkatram, A., Weil, J. C., Paine, R. J., Wilson, R. B., Lee, R. F., Peters, W. D., & Brode, R. W. (2005). AERMOD: A dispersion model for industrial source applications. Part I: General model formulation and boundary layer characterization. *Journal of Applied Meteorology*, *44*(5), 682–693.

Corbitt, R. A. (1990). Standard handbook of environmental engineering.

Craig, K., De Kock, D., & Snyman, J. (1999). Using CFD and mathematical optimization to investigate air pollution due to stacks. *International Journal for Numerical Methods in Engineering*, *44*(4), 551–565.

Cyrys, J., Hochadel, M., Gehring, U., Hoek, G., Diegmann, V., Brunekreef, B., & Heinrich, J. (2005). GIS-based estimation of exposure to particulate matter and NO2 in an urban area: stochastic versus dispersion modeling. *Environmental Health Perspectives*, *113*(8), 987.

Dabberdt, W. F., Ludwig, F., et al. (1973). Validation and applications of an urban diffusion

model for vehicular pollutants. *Atmospheric Environment*, *7*(6), 603–618.

Daly, A. & Zannetti, P. (2007). Air pollution modeling – An overview. *Ambient air pollution*.

De Hoogh, K., Chen, J., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., Bauwelinck, M., van Donkelaar, A., Hvidtfeldt, U. A., Katsouyanni, K., et al. (2018). Spatial PM2.5, NO2, O3 and BC models for Western Europe - Evaluation of spatiotemporal stability. *Environment International*, *120*, 81–92.

Dezzutti, M., Berri, G., & Venegas, L. (2018). Intercomparison of Atmospheric Dispersion Models Applied to an Urban Street Canyon of Irregular Geometry. *Aerosol and Air Quality Research*, *18*, 820–828.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, (pp. 1–15). Springer.

Faulkner, M. & Russell, P. (2010). Review of Local Air Quality Management.

Gal, Y. & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, (pp. 1050–1059).

Gal, Y., Van Der Wilk, M., & Rasmussen, C. E. (2014). Distributed variational inference in sparse Gaussian process regression and latent variable models. In *Advances in Neural Information Processing Systems*, (pp. 3257–3265).

Galmarini, S., Vinuesa, J.-F., & Martilli, A. (2009). Relating small-scale emission and concentration variability in air quality models. In *Meteorological and Air Quality Models for Urban Areas* (pp. 11–19). Springer.

Gardner, M. & Dorling, S. (1999). Neural network modelling and prediction of hourly NOx and NO2 concentrations in urban air in London. *Atmospheric Environment*, *33*(5), 709–719.

Giavis, G. M., Kambezidis, H. D., & Lykoudis, S. P. (2008). Frequency distribution of particulate matter (PM10) in urban environments. *International Journal of Environment and Pollution*, *36*(1-3), 99–109.

Gidhagen, L., Johansson, C., Langner, J., & Olivares, G. (2004). Simulation of NOx and ultrafine particles in a street canyon in Stockholm, Sweden. *Atmospheric Environment*, *38*(14), 2029–2044.

Gilbert, N. L., Goldberg, M. S., Beckerman, B., Brook, J. R., & Jerrett, M. (2005). Assessing spatial variability of ambient nitrogen dioxide in Montreal, Canada, with a land-use regression model. *Journal of the Air & Waste Management Association*, *55*(8), 1059–1063.

Gokhale, S. & Khare, M. (2004). A review of deterministic, stochastic and hybrid vehicular exhaust emission models. *International Journal of Transport Management*, *2*(2), 59–74.

Guerreiro, C., de Leeuw, F., & Foltescu, V. (2013). Air quality in Europe - 2013 report.

Gulliver, J., de Hoogh, K., Fecht, D., Vienneau, D., & Briggs, D. (2011). Comparative assessment of GIS-based methods and metrics for estimating long-term exposures to air pollution. *Atmospheric Environment*, *45*(39), 7072–7080.

Gurjar, B., Butler, T., Lawrence, M., & Lelieveld, J. (2008). Evaluation of emissions and air quality in megacities. *Atmospheric Environment*, *42*(7), 1593–1606.

Hanna, S. & Chang, J. (2012). Acceptance criteria for urban dispersion model evaluation. *Meteorology and Atmospheric Physics*, *116*(3-4), 133–146.

Hanna, S. R., Britter, R., & Franzese, P. (2003). A baseline urban dispersion model evaluated with Salt Lake City and Los Angeles tracer data. *Atmospheric Environment*, *37*(36), 5069–5082.

Hanna, S. R., Egan, B. A., Purdum, J., & Wagler, J. (2001). Evaluation of the ADMS, AERMOD, and ISC3 dispersion models with the OPTEX, Duke Forest, Kincaid, Indianapolis and Lovett field datasets. *International Journal of Environment and Pollution*, *16*(1-6), 301–314.

Hochadel, M., Heinrich, J., Gehring, U., Morgenstern, V., Kuhlbusch, T., Link, E., Wichmann, H., Krämer, U., et al. (2006). Predicting long-term average concentrations of traffic-related air pollutants using GIS-based information. *Atmospheric Environment*, *40*(3), 542–553.

Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., & Briggs, D. (2008). A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment*, *42*(33), 7561–7578.

Hole, L. R., Christensen, J. H., Ruoho-Airola, T., Tørseth, K., Ginzburg, V., & Glowacki, P. (2009). Past and future trends in concentrations of sulphur and nitrogen compounds in the Arctic. *Atmospheric Environment*, *43*(4), 928–939.

Hosker Jr, R. (1975). Consequences of effluent release. *Environmental Research Laboratories*, 147.

Hunter, L., Johnson, G., & Watson, I. (1992). An investigation of three-dimensional characteristics of flow regimes within the urban canyon. *Atmospheric Environment*, *26*(4), 425–432.

Isakov, V., Johnson, M., Touma, J., & Özkaynak, H. (2012). Development and evaluation of land-use regression models using modeled air quality concentrations. In *Air Pollution Modeling and its Application XXI* (pp. 717–722). Springer.

Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahsuvaroglu, T., Morrison, J., & Giovis, C. (2004). A review and evaluation of intraurban air pollution exposure models. *Journal of Exposure Science and Environmental Epidemiology*, *15*(2), 185–204.

Kalhor, M. & Bajoghli, M. (2017). Comparison of AERMOD, ADMS and ISC3 for incomplete upper air meteorological data (case study: Steel plant). *Atmospheric Pollution Research*, *8*(6), 1203–1208.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging Artificial Intelligence Applications in Computer Engineering*, *160*, 3–24.

Kukkonen, J., Partanen, L., Karppinen, A., Walden, J., Kartastenpää, R., Aarnio, P., Koskentalo, T., & Berkowicz, R. (2003). Evaluation of the OSPM model combined with an urban background model against the data measured in 1997 in Runeberg Street, Helsinki. *Atmospheric Environment*, *37*(8), 1101–1112.

Kumar, P., Morawska, L., Martani, C., Biskos, G., Neophytou, M., Di Sabatino, S., Bell, M., Norford, L., & Britter, R. (2015). The rise of low-cost sensing for managing air pollution in cities. *Environment International*, *75*, 199–205.

Larkin, A., Geddes, J. A., Martin, R. V., Xiao, Q., Liu, Y., Marshall, J. D., Brauer, M., & Hystad, P. (2017). Global land use regression model for nitrogen dioxide air pollution. *Environmental Science & Technology*, *51*(12), 6957–6964.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436.

Levy, J. I., Spengler, J. D., Hlinka, D., Sullivan, D., & Moon, D. (2002). Using CALPUFF to evaluate the impacts of power plant emissions in Illinois: model sensitivity and implications. *Atmospheric Environment*, *36*(6), 1063–1075.

Liu, C., Henderson, B. H., Wang, D., Yang, X., & Peng, Z.-r. (2016). A land use regression application into assessing spatial variation of intra-urban fine particulate matter (PM2.5) and nitrogen dioxide (NO2) concentrations in City of Shanghai, China. *Science of The Total Environment*, *565*, 607–615.

Lu, H.-C. & Fang, G.-C. (2002). Estimating the frequency distributions of PM10 and PM2.5 by the statistics of wind speed at Sha-Lu, Taiwan. *Science of the Total Environment*, *298*(1-3), 119–130.

Marshall, J. D., Nethery, E., & Brauer, M. (2008). Within-urban variability in ambient air pollution: comparison of estimation methods. *Atmospheric Environment*, *42*(6), 1359–1369.

Martin, D. O. (1976). The change of concentration standard deviations with distance. *Journal of the air pollution control association*, *26*(2), 145–147.

McHugh, C., Carruthers, D., & Edmunds, H. (1997). ADMS-Urban: an air quality management system for traffic, domestic and industrial pollution. *International Journal of Environment and Pollution*, *8*(3-6), 666–674.

Morgenstern, V., Zutavern, A., Cyrys, J., Brockow, I., Gehring, U., Koletzko, S., Bauer, C.-P., Reinhardt, D., Wichmann, H.-E., & Heinrich, J. (2007). Respiratory health and individual estimated exposure to traffic-related air pollutants in a cohort of young children. *Occupational and environmental medicine*, *64*(1), 8–16.

Mueller, M., Wagner, M., Barmpadimos, I., & Hueglin, C. (2015). Two-week NO2 maps for the City of Zurich, Switzerland, derived by statistical modelling utilizing data from a routine passive diffusion sampler network. *Atmospheric Environment*, *106*, 1–10.

Namdeo, A., Mitchell, G., & Dixon, R. (2002). TEMMS: an integrated package for modelling and mapping urban traffic emissions and air quality. *Environmental Modelling & Software*, *17*(2), 177–188.

Nasrabadi, N. M. (2007). Pattern recognition and machine learning. *Journal of Electronic Imaging*, *16*(4), 049901.

Naughton, O., Donnelly, A., Nolan, P., Pilla, F., Misstear, B., & Broderick, B. (2018). A land use regression model for explaining spatial variation in air pollution levels using a wind sector

based approach. *Science of The Total Environment*, *630*, 1324–1334.

Nova, I., Ciardelli, C., Tronconi, E., Chatterjee, D., & Weibel, M. (2007). NH3-NO/NO2 SCR for diesel exhausts after treatment: mechanism and modelling of a catalytic converter. *Topics in catalysis*, *42*(1), 43–46.

Oettl, D., Kukkonen, J., Almbauer, R. A., Sturm, P. J., Pohjola, M., & Härkönen, J. (2001). Evaluation of a Gaussian and a Lagrangian model against a roadside data set, with emphasis on low wind speed conditions. *Atmospheric Environment*, *35*(12), 2123–2132.

Oke, T. R. (1988). Street design and urban canopy layer climate. *Energy and Buildings*, *11*(1), 103–113.

O'Neill, S. M. & Lamb, B. K. (2005). Intercomparison of the community multiscale air quality model and CALGRID using process analysis. *Environmental Science & Technology*, *39*(15), 5742–5753.

Owen, B., Edmunds, H., Carruthers, D., & Singles, R. (2000). Prediction of total oxides of nitrogen and nitrogen dioxide concentrations in a large urban area using a new generation urban scale dispersion model with integral chemistry model. *Atmospheric Environment*, *34*(3), 397–406.

Pasquill, F. (1961). The estimation of the dispersion of windborne material. *Meteorol. Mag*, *90*(1063), 33–49.

Pedone, M., Granieri, D., Moretti, R., Fedele, A., Troise, C., Somma, R., & De Natale, G. (2017). Improved quantification of CO2 emission at Campi Flegrei by combined Lagrangian Stochastic and Eulerian dispersion modelling. *Atmospheric Environment*, *170*, 1–11.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Pilling, M., ApSimon, H., Carruthers, D., & Carslaw, D. (2007). *Trends in Primary Nitrogen Dioxide in the UK*. Defra Publications.

Pohoata, A. & Lungu, E. (2017). A Complex Analysis Employing ARIMA Model and Statistical Methods on Air Pollutants Recorded in Ploiesti, Romania. *Revista De Chimie*, *68*(4), 818–823.

Qiu, X., Zhang, L., Ren, Y., Suganthan, P. N., & Amaratunga, G. (2014). Ensemble deep learning for regression and time series forecasting. In *Computational Intelligence in Ensemble Learning*, (pp. 1–6). IEEE.

Quinlan, J. R. (2014). *C4.5: programs for machine learning*. Elsevier.

Reynolds, S. D., Roth, P. M., & Seinfeld, J. H. (1973). Mathematical modeling of photochemical air pollution - I: Formulation of the model. *Atmospheric Environment*, *7*(11), 1033–1061.

Riddle, A., Carruthers, D., Sharpe, A., McHugh, C., & Stocker, J. (2004). Comparisons between FLUENT and ADMS for atmospheric dispersion modelling. *Atmospheric Environment*, *38*(7), 1029–1038.

Righi, S., Lucialli, P., & Pollini, E. (2009). Statistical and diagnostic evaluation of the ADMS-

Urban model compared with an urban air quality monitoring network. *Atmospheric Environment*, *43*(25), 3850–3857.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533.

Rzeszutek, M., Bogacki, M., Bździuch, P., & Szulecka, A. (2018). Improvement assessment of the OSPM model performance by considering the secondary road dust emissions. *Transportation Research Part D: Transport and Environment*.

Sahsuvaroglu, T., Arain, A., Kanaroglou, P., Finkelstein, N., Newbold, B., Jerrett, M., Beckerman, B., Brook, J., Finkelstein, M., & Gilbert, N. L. (2006). A land use regression model for predicting ambient concentrations of nitrogen dioxide in Hamilton, Ontario, Canada. *Journal of the Air & Waste Management Association*, *56*(8), 1059–1069.

Sánchez, A. S., Nieto, P. G., Fernández, P. R., del Coz Díaz, J., & Iglesias-Rodríguez, F. J. (2011). Application of an SVM-based regression model to the air quality study at local scale in the Aviles urban area in Spain. *Mathematical and Computer Modelling*, *54*(5), 1453–1466.

Scire, J. S., Strimaitis, D. G., & Yamartino, R. J. (2000). A user's guide for the CALPUFF dispersion model. *Earth Tech, Inc*, *521*, 1–521.

Seinfeld, J. H. & Pandis, S. N. (2016). *Atmospheric chemistry and physics: from air pollution to climate change*. John Wiley & Sons.

Smola, A. J. & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*(3), 199–222.

Srivastava, A. & Rao, B. P. S. (2011). *Urban Air Pollution Modeling*. InTech.

Stedman, J. R., Vincent, K. J., Campbell, G. W., Goodwin, J. W., & Downing, C. E. (1997). New high resolution maps of estimated background ambient NOx and NO2 concentrations in the UK. *Atmospheric Environment*, *31*(21), 3591–3602.

Stocker, J., Hood, C., Carruthers, D., & McHugh, C. (2012). ADMS-Urban: developments in modelling dispersion from the city scale to the local scale. *International Journal of Environment and Pollution*, *50*(1-4), 308–316.

Tso, G. K. & Yau, K. K. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy*, *32*(9), 1761–1768.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer Science & Business Media.

Vardoulakis, S., Fisher, B. E., Pericleous, K., & Gonzalez-Flesca, N. (2003). Modelling air quality in street canyons: a review. *Atmospheric Environment*, *37*(2), 155–182.

Vardoulakis, S., Valiantis, M., Milner, J., & ApSimon, H. (2007). Operational air pollution modelling in the UK-Street canyon applications and challenges. *Atmospheric Environment*, *41*(22), 4622–4637.

Verikas, A., Vaiciukynas, E., Gelzinis, A., Parker, J., & Olsson, M. C. (2016). Electromyographic patterns during golf swing: Activation sequence profiling and prediction of shot effectiveness. *Sensors*, *16*(4), 592.

Wang, J.-Z., Wang, J.-J., Zhang, Z.-G., & Guo, S.-P. (2011). Forecasting stock indices with back propagation neural network. *Expert Systems with Applications*, *38*(11), 14346–14355.

Watson, T. B., Heiser, J., Kalb, P., Dietz, R. N., Wilke, R., Wieser, R., & Vignato, G. (2005). The New York city urban dispersion program March 2005 field study: tracert methods and results. Technical report, Brookhaven National Lab.

Weisberg, S. (2005). *Applied linear regression*, volume 528. John Wiley & Sons.

Westmoreland, E. J., Carslaw, N., Carslaw, D. C., Gillah, A., & Bates, E. (2007). Analysis of air quality within a street canyon using statistical and dispersion modelling techniques. *Atmospheric Environment*, *41*(39), 9195–9205.

WHO (2000). Air quality guidelines for Europe.

WHO (2003). Health aspects of air pollution with particulate matter, ozone and nitrogen dioxide.

WHO (2009). *Global health risks: mortality and burden of disease attributable to selected major risks*. World Health Organization.

Yamartino, R., Scire, J., Carmichael, G., & Chang, Y. (1992). The CALGRID mesoscale photochemical grid model - I. Model formulation. *Atmospheric Environment*, *26*(8), 1493–1512.