MULTIVARIATE STUDY OF VEHICLE EXHAUST PARTICLES USING MACHINE LEARNING AND STATISTICAL TECHNIQUES

Aminu Suleiman

A thesis submitted to the University of Birmingham for the degree of DOCTOR OF PHILOSOPHY

School of Engineering

College of Engineering and Physical Sciences

The University of Birmingham

May 2016

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

Abstract

Road traffic is a major contributor to urban concentrations of particulate matter and has been shown to be detrimental to human health and the urban environment. The effects of these particles can be effectively controlled by providing adequate and efficient air quality monitoring control and mitigation measures. The efficiency of such measures is tested or predicted through effective and less expensive air quality modelling.

This research has examined the application of machine learning and statistical methods for developing roadside particle (number/mass concentrations) prediction models that can be used for air quality management. Data collected from continuous monitoring stations including pollutants, traffic and meteorological variables were used for training the models. A hybrid feature selection method involving Genetic Algorithms and Random Forests was successfully used in selecting the most relevant predictor variables for the models from the variables selected based on their correlation with the PM_{10}, $PM_{2.5}$ and PNC concentrations. The study found that the hybrid feature selection could be used with both statistical and machine learning methods to produce less expensive and more efficient air quality prediction models.

Among the machine learning models studied the Boosted Regression Trees (BRT), Random Forests (RF), Extreme Learning Machines (ELM) and Deep Learning Algorithms were found to be the most suitable for the predictions of roadside PM_{10}, $PM_{2.5,}$ and PNC concentrations. The machine learning models performed better than the ADMS-road model in spatiotemporal predictions involving monitoring sites locations. Moreover, they performed much better in predicting the concentrations in street canyons. The Artificial Neural Network (ANN) and BRT were found to be suitable for air quality management applications involving traffic management scenarios.

Dedication

This thesis is specifically dedicated to my family and the entire Humanity

Acknowledgement

This thesis would not have been possible without the support and inspiration from many people. Firstly, I would like to express my sincere gratitude and appreciation to Petroleum Technology Development Fund (PTDF) for sponsoring my PhD studies in the University of Birmingham United Kingdom.

My greatest thanks and appreciation goes to my supervisors Professor Miles Tight and Dr Andrew Quinn, for their invaluable contributions towards a successful completion of my thesis. Their guidance, patience and support throughout the process kept my hope alive, and I am short of words to show my gratitude to them.

I would like to thank Dr James Tate of the Institute of Transport, University of Leeds for providing part of the data used for this research. Also, special thanks go to London-air, DEFRA, BADC, DoT, and TfL for providing most of the air quality and traffic data used in the course of this study.

My deepest gratitude goes to my family for their love and unconditional support, particularly my wife Jamila and my sons Suleiman and Muhammad, who made many sacrifices and had always been on my side during this study.

I cannot thank my parents enough for their support and prayers throughout my life. Without their effort and love, I would not have been where I am today. Special thanks go to Malam Yahaya Ibrahim for his continuous and irreplaceable support and encouragement since my secondary school days to date. He is, in fact, a great source of inspiration and a mentor. Finally, I would like to thank my colleagues in Bayero University Kano and members of my research room F59B in the University of Birmingham for their friendly advice and encouragement throughout my study.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ADMS	Atmospheric Dispersion Modelling System
ANN	Artificial Neural Network
APS	Aerodynamic Particle Sizer
AURN	Automatic to Urban and Rural Network
BC	Black Carbon
BP	Backpropagation
BPP	Bivariate Polar Plots
BRNN	Bayesian Regularised Neural Network
BRT	Boosted Regression Trees
CART	Classification and Regression Trees
COE	Coefficient of Efficiency
COMEAP	Committee on the Medical Effects of Air Pollutants
CPC	Condensation Particle Counter
DEFRA	Department for Environment, Food and Rural Affairs
DfT	Department for Transport

DL	Deep Learning
DMA	Differential Mobility Analyser
DMS	Differential Mobility Spectrometer
DNN	Deep Neural Network
EC	Elemental Carbon
ELM	Extreme Learning Machine
ELPI	Electrical Low-Pressure Impactor
EU	European Union
FB	Fractional Bias
FCS	Fully Conditional Specification
FDMS	Filter Dynamics Measurement System
FMPS	Fast Mobility Particle Sizer
GA	Genetic Algorithm
GAM	Generalised Additive Model
GPS	Global Positioning System
HDV	Heavy-Duty Vehicles
IA	Index of Agreement

IOA	Index of Agreement
LAEI	London Atmospheric Emission Inventory
LAS	Laser and Aerosol Spectrometer
LDV	Light-Duty Vehicles
LEZ	Low Emission Zone
MATLAB	Matrix Laboratory
MB	Mean Bias
MBE	Mean Bias Error
MGE	Mean Gross Error
MICE	Multiple Imputations by Chained Equation
MIDAS	Met Office Integrated Data Archive System
ML	Machine Learning
MLPN	Multilayer Perceptron Network
MLR	Multiple Linear Regression
MRE	Mean Relative Error
MSE	Mean Squared Error
NMGE	Normalised Mean Gross Error

OSPM	Operational Street Pollution Model
PC	Principal Components
PCA	Principal Component Analysis
PCR	Principal Component Regression
PLS	Partial Least Squares
PLSR	Partial Least Square Regression
PM	Particulate Matter
PNC	Particle Number Count
RF	Random Forests
RMSE	Root Mean Square Error
RSS	Residual Sum of Squares
SA	Simulated Annealing
SLFN	Single Hidden Layer Feedforward Neural Networks
SMPS	Scanning Mobility Particle Sizer
SSE	Sum-of-Squared Errors
SVD	Singular Value Decomposition
SVM	Support Vector Machine

TEOM	Tapered Element Oscillating Microbalance
TfL	Transport for London
UAQM	Urban Air Quality Management
UNECE	United Nations Economic Commission for Europe
USEPA	United States Environmental Protection Agency
WHO	World Health Organisation
WRAC	Wide Range Aerosol Classifier

Publications

SULEIMAN, A., TIGHT, M. & QUINN, A. 2016. Hybrid Neural Networks and Boosted Regression Tree Models for Predicting Roadside Particulate Matter. *Environmental Modeling & Assessment*, 1-20.

SULEIMAN, A., TIGHT, M. & QUINN, A. Assessment and prediction of the impact of road transport on ambient concentrations of Particulate Matter PM_{10}. *Transportation Research Part D: Transport and Environment* (Under review)

SULEIMAN, A., TIGHT, M. & QUINN, A. Applying machine learning methods in managing urban concentrations of traffic-related particulate matter (PM_{10} and $PM_{2.5}$) (to be submitted to the *Environmental Monitoring and Assessment*)

Conferences

SULEIMAN, A., TIGHT, M. & QUINN, A. 2014. Extensive Evaluation of Neural Network Based PM_{10} Prediction Models Using Openair Package. 20th International Transport and Air Pollution Conference 2014, Graz, Austria.

SULEIMAN, A., TIGHT, M. & QUINN, A. 2015. Quantification and Prediction of the Impact of Road Transport on Ambient Concentrations of Particulate Matter PM_{10}. 2015 International Conference on Environment Pollution and Prevention-ICEPP 2015. Dubai, UAE.

Chapter 1

Introduction

1.1 Background

An outcome of the world's ever growing population is an unprecedented increase in vehicle population and use in urban areas. This growth in population and increasing transport demand is leading to rapid deterioration of air quality worldwide. Although transport and mobility choices are essential ingredients of urban lives and livelihoods, their effect on air quality is a primary source of concern today. The World Health Organization (2014a) reported that only a few of the world cities that are monitoring air quality had met the WHO guidelines for safe air quality levels. This continuous air quality deterioration puts people at a greater risk of premature deaths from diseases aggravated by poor air quality. In its report, World Health Organization (2014b) stated that outdoor air pollution is responsible for the premature deaths of about 3.7 million people under the age of 60 years worldwide in 2012.

Transport systems account for about 23% of the world's energy-related GHG emissions, and land transportation contributes nearly three-quarters of these emissions (Xia et al., 2015a). Several studies have shown that there is a correlation between asthma, bronchitis, pneumonia, and respiratory infections and settlements located near major roads (Brauer et al., 2002, Kim et al., 2004, McConnell et al., 2006, Lindgren et al., 2009, Heinrich et al., 2005). Moreover, long-term exposure to the gaseous and particulate matter pollutants released by traffic have a strong link to mortality rates (Namdeo and Bell, 2005). The fine particles ($PM_{2.5}$ or less, i.e. particles with aerodynamic diameters less than or equal to 2.5 mm) which are mainly contributed by road traffic, are often associated with the premature death which accounts for most of the costs of air pollution (USEPA, 2011, Yim and Barrett,

2012, COMEAP, 2010). In the UK for example, the annual mortality burden caused by air pollution is equivalent to about 40,000 deaths (Holgate et al., 2016). Gowers et al. (2014) estimated that the portion of mortality associated with the long-term exposure to the current levels of particulate pollution from anthropogenic sources in some Local Authorities in the UK range between 2.5 in rural areas of Scotland and Northern Ireland, 3 and 5% in Wales, to more than 8% in some polluted London boroughs. These mortality estimates might range between one-sixth and up to double due to uncertainties in mortality risk associated with ambient $PM_{2.5}$ (Gowers et al., 2014). Also for the UK, the annual cost of air pollution resulting from illness, premature deaths and the cost to society and businesses amount to about £20 billion (Holgate et al., 2016).

The ambient concentrations of particulate matter have been shown to be detrimental to not only human health but the urban environment itself (Anderson et al., 2012, Lawal et al., 2015, Brunekreef et al., 2009). Particulate matter also contributes to visibility impairment (Yang et al., 2012) and often contributes to road accidents (Abdel-Aty et al., 2011). The EU directive on ambient air quality and cleaner air for Europe (2008/50/EC), puts a limit on the level of PM_{10} that should not be exceeded for the purpose of reducing its impact on human health. The EU Directive states that the daily and annual mean of PM_{10} should not exceed $50\mu g/m^3$ and $40\mu g/m^3$ respectively. The daily limit should not be exceeded more than 35 days a year. The annual limit for $PM_{2.5}$ is set at $25\mu g/m^3$. These limits can be effectively controlled if the sources of particulate matter and the factors affecting its levels are adequately characterised and quantified.

The levels of particulate matter in some urban areas in the UK and other European countries are declining. However, emissions in areas close to major roads remain the challenge of the regulatory authorities due to frequent cases of exceedances of air quality limits and

objectives (Guerreiro et al., 2014). Particulate emissions from road vehicles can both emanate from the vehicle exhaust or non-exhaust sources such as wear and tear of vehicle parts (e.g. tyre and clutch). Another important source of road traffic particulate emissions is re-suspension of dust due to vehicle movements (Pant and Harrison, 2013). Hence the need for more studies on the accurate characterization, estimation and prediction of road traffic contribution to particulate concentrations in urban areas. The information obtained from such studies could be useful in identifying relevant and efficient control measures to the dominant sources of particles in a particular area e.g. proximity to major roads. If target levels of greenhouse gas emissions are met in the UK, 5700 deaths, 1600 hospital admissions for heart and lung problems and 2400 case of bronchitis would be prevented every year amounting to an economic cost of about €3.9 billion (Holgate et al., 2016).

1.2 Problem Statement

The effects of traffic-derived air pollution can be effectively controlled by providing adequate and efficient air quality monitoring control and mitigation measures (Holgate et al., 2016). The future effects of these measures can be designed and tested with the aid of air quality models. Air quality regulatory agencies have to complement measurements of air quality with models that can accurately predict pollutant concentrations and determine the causes of the air quality problems.

The models are calibrated using historical air pollution data and are used to forecast the likely air quality scenarios for the future. Air quality models currently used by regulatory agencies are mostly deterministic and are built on simple assumptions about the atmospheric processes and involve high computational cost which limits their application. They also require knowledge of the relationships between the variables involved and meteorological conditions.

The deterministic models are not only constrained by the accurate characterisation of the dynamics of the natural phenomena but also on the model configuration options. The use of default parameters and lack of real observations with the same spatial resolution with which to compare the model outputs are examples of the model configuration limitations (Chave and Levin, 2003, National Research Council Committee on Models in the Regulatory Decision Process, 2007). Steady-state Gaussian plume models are the most widely used air quality models and have been applied successfully in many air quality studies. However, despite their successful application, they are limited by assumptions regarding changes of wind and source emissions over time and do not include the detailed chemistry of particle pollutants (Lagzi et al., 2013).

In contrast, machine learning methods such as Artificial Neural Networks (ANN) (Haykin, 2005) and ensemble regression methods can be used to build air quality models with comparable prediction accuracy at a lesser computational cost and with no assumption of the atmospheric processes involved (Gardner and Dorling, 2000). The machine learning models are capable of handling complex and robustly nonlinear relationships that exist between air quality variables (Esplin, 1995) and produce prediction or forecasting models that can perform extremely well in practice. Machine learning methods such as ANN have been widely used in air quality studies (Taspinar, 2015, Ragosta et al., 2015, Elangasinghe et al., 2014a), involving prediction and forecasting of air pollutants ranging from the current hour to several days in advance (Russo et al., 2013, de Gennaro et al., 2013). ANN have also been applied in the prediction of pollution peaks (Catalano et al., 2016). Also ensemble learning methods like Boosted Regression Trees (BRT) and Random Forests (RF), and Support Vector Machines (SVM) have been applied in air quality studies (Yang et al., 2008, Sanchez et al., 2011, Sanchez et al., 2013, Wang et al., 2009, Xia et al., 2015b). Despite the quantum of research in the application of these methods in air quality prediction little is

known about their adoption by the regulatory agencies. Also, most of the studies focused on the ability of the methods to predict or forecast pollutant concentrations but not their application in evaluating air quality control measures. This might be attributed to the general notion that the machine learning models are hardly interpretable and are considered as black boxes.

This research is aimed at investigating the use of three machine learning and five statistical methods to develop air quality models for predicting roadside concentrations of PM_{10}, $PM_{2.5}$ and Particle Number Count (PNC). Also, to investigate the use of the machine learning methods in air quality management.

The machine learning methods include Artificial Neural Networks (ANN), ensemble regression trees (Boosted Regression Trees (BRT), Random Forest (RF)), and Support Vector Machines (SVM). The statistical methods are a stepwise regression, Lasso Regression, Elastic-net Regression, Principal Component Regression (PCR), Partial Least Square Regression (PLSR) and Multiple Linear Regression (MLR).

The statistical methods were selected based on their unique improvements in handling the limitations of ordinary least square regression. The ANN, ensemble regression trees, and SVM were chosen based on their popularity in artificial intelligence applications and their unique formulations for handling the machine learning problems (Hastie et al., 2008a). Also, different formulations of these main methods were considered to selecting the most appropriate for the prediction of the particles, and each has a particular feature for improving the original formulation. Each formulation was selected for its improvement in tackling the problem of overfitting, issues of generalisation, training speed, parameter tuning and prediction accuracy.

The response variables (i.e. PM_{10}, $PM_{2.5}$ and the PNC) were selected because of their higher impact on human health, and most of the operational models lack proper formulations for

their predictions. Also, PNC is of particular interest because of its potentials for being a better measure for regulating the ambient concentration of particles and its foreseeable future in the EU directive (2008/50/EC).

Among these methods, ANN, SVM and RF methods have been applied in modelling various air pollutant metrics, but little is known about their application to particle number count (PNC). BRT was rarely reported to have been used in air quality modelling (Carslaw and Taylor, 2009, Sayegh et al., 2016) despite its successful application in areas of ecology and medicine (Sharifi and Ghafourian, 2014, Elith et al., 2008). Although Lasso and Elastic-net regressions have the ability to model the relationships between air quality data and also perform feature extraction and variable shrinkage, they have rarely been applied to modelling air quality or used in conjunction with artificial neural networks as feature selection methods. However, they have been successfully implemented in Health related studies (Sun et al., 2013), Pattern Recognition (Tan et al., 2011) and forecasting (Korobilis, 2013, Aye and Gupta, 2013). They have also been used for model selection (Savin and Winker, 2013).

The most successful of either statistical or machine learning methods will be compared with the ADMS-Roads model (Carruthers et al., 1997). ADMS-Roads was selected because it is the most widely used air quality models in UK local authorities and it was developed and calibrated using UK data. Besides its popularity in the UK, ADMS models have been applied successfully in many air quality studies (Silva and Mendes, 2009, Righi et al., 2009, Hirtl and Baumann-Stanzer, 2007, Carruthers et al., 2001, Carruthers et al., 1994). Also, the ADMS Roads are widely used, and there is an active user community with experience of using ADMS models (Williams et al., 2011). Other models that are widely used in the UK include OSPM (Hertel and Berkowicz, 1989), Community Multiscale Air Quality (CMAQ) modelling system (Byun et al., 1998) and AERMOD (Cimorelli et al., 1998). A

comprehensive review of the commonly used air quality models in the UK can be found in Williams et al. (2011).

1.3 Aim and Objectives of the Research

1.3.1 Aim

To examine the application of Machine Learning and Statistical Methods for developing roadside particle (number/mass concentrations) prediction models that can be used for air quality management.

1.3.2 Objectives

1. 1. To identify sites with the required data availability that will be used for training and testing of the models to be developed. The data required include; roadside particle number concentration (PNC) and particle mass (PM_{10}, $PM_{2.5}$) concentrations, Traffic data, and meteorological parameters.

2. To determine appropriate feature selection procedures for selecting relevant predictor variables for the ANN and Statistical models to be developed. The techniques to be considered include Genetic Algorithm (GA) and Simulated Annealing (SA) combined with Random Forests (RF).

3. To use machine learning methods; ANN, BRT, RF and SVM and statistically based techniques in developing air quality models for predicting roadside particle concentrations. The statistical methods are Stepwise Regression, Lasso regression, Elastic-net Regression, Principal Component Regression (PCR), Partial Least Square Regression (PLSR) and Multiple Linear Regression (MLR).

4. To use ADMS-Roads (an operational model) to predict PM_{10} and $PM_{2.5}$ concentrations and compare its predictive performance with that of the machine learning models developed in (3).

5. To use some of the machine learning models developed in (3) above and the ADMS-Roads model to predict the implication of a hypothetical air quality management scenario and compare their results.

1.4 Expected Contribution of Research

This research seeks to add to the body of knowledge on traffic-related air pollution modelling. Thereby contributing to accurate monitoring and prediction of air quality levels and by extension help in reducing the impacts of vehicle exhaust emissions on health and overall urban air pollution. Machine learning based air quality models have been established to be superior to conventional statistical and deterministic models and have been extensively used to predict air pollutant concentrations. However, very few are focusing on Particle Number Count (PNC) as a metric for measuring the particles. Also, the use of Deep learning, Extreme learning machines, BRT, Elastic-net and Lasso regression in air quality modelling have rarely been reported despite their successful applications in other disciplines. Moreover, the use of Genetic Algorithm (GA) and Simulated Annealing (SA) combined with Random Forests (RF) as feature selection methods for the statistical and machine learning modelling involving particulate matter and PNC have not been reported to the best of the researcher's knowledge. Comparing the capabilities of the machine learning models and ADMS-Roads in managing air quality control scenarios is also first of its kind to best of the researcher's knowledge.

1.5 Structure of the Thesis

This thesis is organised according to following chapters

Chapter 2 reviews the effects of transport-related air pollutants on human health and the various modelling options available for the prediction of roadside pollutant concentrations. Also, the statistical methods, machine learning methods and the ADMS-Roads model (McHugh et al., 1997), as well as various air quality model evaluation methods, are reviewed.

Chapter 3 describes the processes followed in the execution of various modelling exercises involved in this research. The study is divided into data collection and analysis, model development and model comparisons. Therefore, this chapter is divided according to these categories where a detailed procedure involved in achieving the objectives of each category is given.

Chapter 4 presents the description of the air quality monitoring sites used in this study, a brief description and descriptive statistics of the traffic, meteorological and pollutant variables. The chapter continued with the analysis of the long-term trends in the particle pollutants and the correlation between various variables in the air quality data.

Chapter 5 examines the temporal and spatial relationships between road traffic and particle concentrations collected at the monitoring sites. These are analysed using bivariate polar plot techniques. The road traffic contribution to the overall roadside particle concentrations (PM_{10}) has been estimated.

Chapter 6 examines the effect of two feature selection methods (Genetic Algorithms (GA) and Simulated Annealing (SA) combined with Random Forests (RF)) on some selected

statistical models for predicting roadside particle concentrations. Also, the application of these methods in the predictions of roadside particles has been examined.

Chapter 7: this chapter discusses the use of three machine learning methods including Artificial Neural Networks (ANN), Ensemble regression trees (BRT and RF), and Support Vector Machines (SVM) in air quality modelling. The selected machine learning methods were trained to predict roadside particle concentration (i.e. PM_{10}, $PM_{2.5}$ and PNC). Also, the response of the machine learning methods to the feature selection procedure carried out in Chapter 6 was investigated.

Chapter 8 evaluates the application of the machine learning based air quality models discussed in Chapter 7 in both spatial and temporal prediction of PM_{10} and $PM_{2.5}$ and PNC concentrations, and their applicability in air quality management involving PM_{10} and $PM_{2.5}$ concentrations. The performance of the selected machine learning models is then compared with the performance of the ADMS-Roads model.

Chapter 9: in this chapter, the summary of the findings of the research has been highlighted. Also, the conclusions and recommendations drawn from the findings and suggestion for further research are given, and it is the final chapter of the thesis.

Chapter 2

Transport, Particulate Matter and Health

2.1 Introduction

Road transport is one of the major vehicles for socioeconomic development. However, it is also an important agent in polluting urban air which affects the health of people and other organisms and their respective environments. Several toxicological and epidemiological studies have shown that long-term exposure to traffic-related pollutants particularly particulate matter leads to premature death which takes the highest share of the costs of air pollution (Yim and Barrett, 2012, Balaguer and Carpin, 2012, Yim et al., 2013, COMEAP, 2010, USEPA, 2011, Holgate et al., 2016). These effects can be effectively reduced through provisions of adequate and efficient air quality control and mitigation measures which are designed and tested using air quality models. The environmental regulatory agencies have to supplement air quality measurements with models that can predict pollutants concentrations and determine the cause of the air quality problems. The use of such models provides an opportunity for using historical data to study the past scenarios of air pollution episodes and to forecast the likely pollution events for the future. This chapter reviewed the effects of transport-related air pollutants on human health and the various modelling options available for the prediction of the roadside pollutant concentrations. Section 2.2, the characteristics and the methods of measuring particulate matter are discussed. In Section 2.3 the health implications of the air pollutants, particularly particulate matter is presented. The air pollutants inventory in the UK and the relevant air quality standards are also highlighted. Section 2.4 focused on the air quality modelling methods where various statistical and machine learning methods are discussed as alternatives to the traditional dispersion models such as ADMS-Roads (McHugh et al., 1997). The evaluation methods for the air quality

model are discussed in Section 2.5. The chapter concludes with the summary of the main discussions in the chapter.

2.2 Particulate Matter

Airborne particulate matter consists of solid and liquid substances of various sizes ranging from a few nanometres in diameter to about 100 micrometres (100μm). Particulate matter exists in two different components i.e. primary components and secondary components. Primary components are released directly from the source into the atmosphere, and secondary components, are formed in the atmosphere by chemical reactions. Particulate matter comes from both human-made and natural sources. It contains a range of chemical compounds, and the identity of these compounds provides clues to its origin. Particulate matter is classified and measured in terms of its various sizes. For example, PM_{10} represents the mass concentration of particles that are less than or equal to 10μm in aerodynamic diameter; $PM_{2.5}$ often called fine particles describes the mass concentration of particles less than or equal to 2.5μm in aerodynamic diameter. Coarse particles fall between PM_{10} and $PM_{2.5}$. Ultrafine particles or Nanoparticles represents the particles with aerodynamic diameter less than 0.1μm which is usually expressed in nanometre (nm) (Colls, 2002b).

2.2.1 Particle Size Distribution

Although particles can be represented with an aerodynamic equivalent diameter, there will always be a small number of particles that will have a certain diameter. As such, a single value of a diameter is not being used to measure particle concentrations. Consequently, particle concentrations are expressed in $\mu g/m^3$ or equivalent. Number distribution is obtained simply by counting the number of particles within each diameter range. The definition of the particle size range of each mode as currently used in the literature varies, depending on the area of application. For example, in toxicology, ultrafine particles are

represented as particles with less than 100nm diameter, fine particles with less than the 1000nm diameter, and coarse particles with greater than the 1000nm diameter. While most of the regulatory agencies around the globe use expressions such as PM_{10}, $PM_{2.5}$ and PM_1 (PM is referred to particulate matter and the subscripts show cut-off sizes in μm)(Kumar et al., 2010). Several measurements of the particle size distributions of the airborne particulate matter in urban atmospheres have been idealised to give relationships between the number of particles, their surface area, volume and mass, and their particle diameters. Figure 2.1 shows the idealised particle size distribution of urban aerosols in the form of three curves showing the number, area, and volume size distributions. "*The vertical axis is scaled in dN/dlog D, dA/dlog D or dV/dlog D, and the horizontal axis in log D. The range of sizes from a given sample is expressed as the normalised number distribution – the number of particles per unit size interval"* (Colls, 2002b).

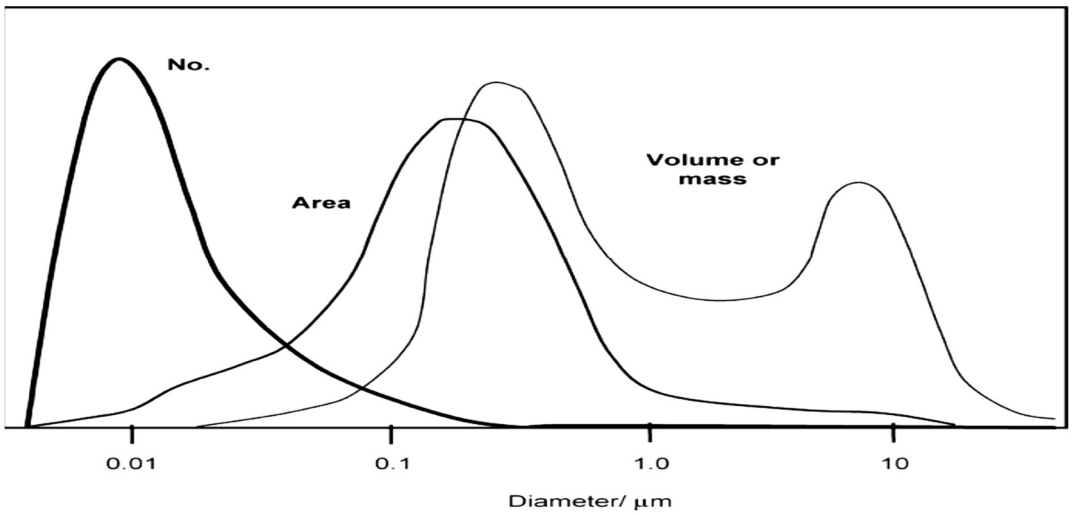


Figure 2.1 A simple distribution of particle number with diameter, and its transformation into surface and volume (or mass) distributions (Colls, 2002b).

2.2.2 Anthropogenic Sources of Airborne Particles

The principal anthropogenic sources of particles in the urban environment include particle emissions from diesel vehicles, shipping, trains and aircraft and emerging sources (i.e. biofuel derived road vehicles and manufactured particles). However, the focus here will be on road vehicles which are believed to be the dominant source in an urban environment.

Road transport contributes immensely to the particulate matter emission in an urban environment. NAEI (2012) revealed that road transport accounted for 24%, 15% and 19% of PM_{10} emission in 2010 in England, Wales and Northern Ireland respectively. A recent source apportionment study in a Western Mediterranean city by Pey et al. (2009) identified five primary particle emission sources and two secondary sources. The primary sources include vehicle exhausts, mineral dust, sea spray, industrial source and fuel-oil combustion. The two secondary sources include a photochemically induced nucleation and regional/urban background particles derived from coagulation and condensation processes. The study disclosed that vehicle exhausts contribution dominate the total number concentration in all the particle sizes (52–86%), especially in the size range of 30 – 200 nm. Vehicle exhaust emission can be further apportioned to vehicles with different fuel consumption e.g. diesel, petrol and biofuel engines. Diesel engines, when compared to petrol engines, emit a higher number of fine and ultra-fine particles (Morawska et al., 1998, Rose et al., 2006, Harris and Maricq, 2001).

Also, the road traffic emits gaseous pollutants, and may cause the re-suspension of the road dust deposited on the traffic lines (Schauer et al., 2006, Thorpe and Harrison, 2008). Colvile et al. (2002) established that PM_{10} emissions from diesel-engine vehicles were higher (66.7%) than those from petrol-fuelled vehicles (11.4%). Most of the mass concentration based studies show that vehicular sources contribute a major percentage of the total PM_{10} in urban Environments. Charron and Harrison (2005) in their study of fine and coarse

particulate matter concentration near roads, revealed that diesel vehicles are the main source of fine and coarse particulate matter with fine having the larger (median contribution of 42%) share mainly from heavy duty vehicles.

Another significant source of particle emission that is related to road transport is non-vehicle exhaust particle emission. This source was reported to have contributed up to 70% of coarse particle mode by mass (AQEG, 1999). Measurements during winter and summer carried out by Hussein et al. (2008) shows that non-vehicle exhaust emission contributes in large part to the concentration of PM_{10}. Factors such as the type of tyres, characteristics of the pavement and vehicle speed, were shown to have affected these concentrations especially if studded tyres are used. A comprehensive review of the sources and properties of non-vehicle exhaust emission can be found in (Thorpe and Harrison, 2008).

2.2.3 Particle Number Concentration (PNC)

Particle Number Concentration (PNC) is another metric used to quantify the particles in the atmosphere expressed in (number of particles/cm^3). Particulate matter especially ultrafine particles are best characterised by particle number count rather than by particle mass concentration as the particle number count dominates regarding population but contribute less to the mass concentration (Sánchez Jiménez et al., 2012). Beside their domination in the fine particles, PNC has a far-reaching effect on health compared to particle mass, because they can easily find their way into the Lungs (Seaton et al., 1995). Even though measurement of PNC and particle-bound metals is cumbersome and requires much time; only a few specific metals are currently regulated. Therefore, they are not routinely measured and monitored in air quality networks (Sánchez Jiménez et al., 2012). In their report, AQMRSG (2011) recommended that modelling for pollutant or metrics that are not currently regulated

should be addressed as early as practicable to be proactive in consideration of their likely future inclusion in the regulations.

Many studies have concluded that road traffic is major sources of particle number concentrations (PNC). Johansson et al. (2007) found that PNC at densely trafficked kerbside locations in Stockholm, Sweden are dominated by ultrafine particles (less than 0.1μm diameter) due to vehicle exhaust emissions. Keogh et al. (2009) disclosed that heavy duty vehicles (HDVs) in urban South- East Queensland were major emitters of particulate matter pollution. Although they contributed only around 6% of total regional vehicle kilometres travelled, their contribution to region's particle number (ultrafine particles) and PM_1 emissions were more than 50%. Measurements at the roadside (4m from the kerb) and downwind from the traffic (more than 25m from the kerb) by Shi et al. (2001) show that nanoparticles (10nm diameter) contributed more than 36 to 44% of the total particle number concentrations.

A study conducted in Barcelona (Pey et al., 2009) shows that vehicle exhausts emissions contribute about 52 to 86% to the number concentration in all the particle sizes considered. This contribution might be due to the fact that higher percentage of particles emitted from both diesel and petrol fuelled vehicles normally falls below 130nm and 60nm respectively (Kumar et al., 2010, Harris and Maricq, 2001) although there are elements of uncertainty in the estimation of petrol fuelled vehicles which depend highly on driving conditions (Graskow et al., 1998).

Biofuel engine vehicles are among the emerging sources of urban particulates particularly number concentrations. Although they are preferred alternative for road vehicles, they are found to have emitted high particle number concentration but less particle mass and gaseous concentrations (Agarwal, 2007, Kittelson, 1998). This behaviour raises some concern as to

whether they can meet the recent EU5 and EU6 particle number concentration limits (Kumar et al., 2010).

2.2.4 Measurements of Particulate Matter

Measurement of Particulate matter concentration suspended in ambient air is not a clear-cut process. Various methods exist for measurement of ambient particulate concentration but due its complexity, the accuracy of the measurement depends on the method used.

2.2.5 PM Mass Concentration Measurement

There are two methods in which PM mass concentration can be measured in ambient air.

1. Direct reading instruments which provide continuous measurement of particle concentration.
2. Filter- based gravimetric samplers that collect particulate material onto the filter which must then be weight in the laboratory to obtain the mass concentration.

2.2.6 Size-Selective Inlet Head

Both automatic samplers and gravimetric filter based methods are widely based on the size selective inlet to exclude large unwanted particles before measurement takes place. This selection is achieved by the use of aerodynamic principles. Convoluted route forces the larger particles outside the path of the small particulate matter, and they can then be stopped from travelling on to the filter in either of the two ways.

1. Use of impactor (particles collected on flat plate)
2. Cyclone (particles collected on inner surface of a ring)

These inlets are designed in such a way that 50% of the critical particle sizes are rejected (AQEG, 2005).

The commonly used techniques in the UK for the mass measurement of particulate matter include filter-based gravimetric samplers (including the European reference sampler) and Tapered Element Oscillating Microbalance (TEOM) analysers. The EU First Air Quality Daughter Directive (1999/30/EC) specifies that the reference method should be used in measuring PM_{10} concentrations as defined in European Standard EN12341. This standard specifies three sampling devices that may be used including super-high volume sampler – the WRAC (Wide Range Aerosol Classifier), High-Volume Sampler – the HVS PM_{10} sampler (68 m^3 h^{-1}) and Low-volume sampler – the LVS PM_{10} sampler (2.3 m^3 h^{-1}) "(AQEG, 2005). The samplers comprise a PM_{10} sampling inlet directly attached to a filter substrate and a regulated flow controller. The mass of the PM_{10} collected on the filter is estimated gravimetrically after completion of the sampling period. Also, before weighing the filter should be conditioned at a relative humidity of 20°C and 50%"(AQEG, 2005).

TEOM analysers are widely used in the UK for monitoring PM mass concentration. For TEOM to be used as USEPA equivalent method for $PM_{10,}$ a default adjustment factor (1.3 * TEOM reading + 3μgm^{-3}) must be applied. The factor is applied to account for possible losses of semi – volatile components which occur as a result of high temperature. The high temperature is maintained to account for the effects of changing humidity on the mass measurements. The measurement methods described above are usually employed for measuring PM_{10} mass concentrations. The same methods can be applied for $PM_{2.5}$ and PM_1 with the only difference when instruments that use optical methods are used which determine the size fractions with methods other than size selective inlet(AQEG, 2005). Table A.1 in Appendix A summarised the most commonly used methods for measuring PM mass concentration measurement.

2.2.7 Particle Number Concentration Measurements

It is widespread practice to measure particle in the ultrafine size range (< 0.1μm) in terms of the particle number concentration. Ultrafine particles contribute less to particle mass but dominate the total particle number. Many instruments exist for measuring particle number beside the conventional light scattering instruments which cannot detect particles within the ultrafine range. These instruments include; Condensation Particle Counter (CPC, TSI), Scanning Mobility Particle Sizer (SMPS, TSI). The CPC is based on the belief that supersaturated vapour will condense on small particles (AQEG, 2005). The sampled aerosol passes through a chamber saturated with n-butyl alcohol vapour and then proceed to a cooled condenser where the particles grow as a result of condensation of the alcohol on the particles. The optical detector will be used to count the particles. Subject to the CPC configuration, the analyser can measure particles in between 0.003 and 2.0 μm. In the UK, measurements between ~0.007 and 2.0 μm have been routinely carried out (AQEG, 2005).

SMPS uses electrical mobility technique to measure number and size distributions. It consists of a bio polar radioactive charger used for charging particles, differential mobility analyser (DMA) use for classifying particles by electrical mobility, and a condensation particle counter(CPC) for detecting particles. Electrical low-pressure impactor (ELPI), Aerodynamic particle sizer (APS), and Differential mobility spectrometer (DMS) are also used for measuring nanoparticles. Other instruments include Fast mobility particle sizer (FMPS), Ultrafine particle (UFP) monitor, Laser and aerosol spectrometer (LAS) (Kumar et al., 2010).

2.3 Traffic Air Pollution and Health

Transport is a fundamental ingredient of contemporary life. The modal choice available to travel short and long distances unlock the doors for personal development and other specialised activities and increases mobility options, promote economic activities and

interaction between people. The economic progress of the entire population depends on the ease of access to goods and services provided by modern transport technology. Unfortunately, these promising characteristics of modern day transportation are closely associated with air pollution which adversely affects the environment and human health (Moschandreas et al., 1996, Siegl et al., 1997, Dora and Phillips, 2000, Liu, 2002, Dabberdt et al., 2006, Pugalenthi et al., 2007, Krzyzanowski et al., 2011, Sharma et al., 2012, Vergun et al., 2013, Da Mota et al., 2014). The World Health Organisation observed that globally, household and ambient (outdoor) air pollution worldwide causes approximately 4.3 million and 3.7 million premature deaths per annum respectively (World Health Organization, 2014b, WHO, 2014).

Studies on human health showed that transport is an important contributor to the health consequences of air pollution (Michal Krzyzanowski and Schneider, 2005, Lelieveld et al., 2015). For example, respiratory diseases such as asthma, bronchitis, pneumonia, and respiratory infections were found to have correlated well with residential and schools proximity to major roads (McConnell et al., 2006, Kim et al., 2004, Brauer et al., 2002, Lindgren et al., 2009, Heinrich et al., 2005). Also, Brunekreef et al. (2009), revealed that long-term exposure to traffic-related gaseous and particulate matter pollutants have a strong link to respiratory mortality. Combustion emissions, especially from road transport make a large contribution to the overall atmospheric concentration of air pollutants. Road traffic, in particular, contributes a range of air pollutants both in gaseous form and particulate matter (PM) of different aerodynamic diameter and chemical composition through direct emission or as a result of the chemical transformation in the atmosphere. Particles are also released into the environment from unpaved roads, wear of tyres and brake linings and other industrial activities. Among the pollutants released by the traffic, particulate matter, particularly the finer particles, are said to have been more related to the health conditions of

the individuals exposed for a long time even at lower concentrations (WHO, 2005). In its publication, COMEAP (2010)revealed that in 2008 alone, the burden of anthropogenic air pollution caused by particulate matter had an effect on mortality equivalent to nearly 29,000 deaths in the UK and associated loss of total population life of 340,000 life-years. Although recently, UK meets the European air quality limit values for particulate matter (DEFRA, 2015a). The annual motility burden caused by particulate and nitrogen dioxide exposure in the UK is equivalent to 40,000 deaths with an estimated social costs of £22.6 billion (Holgate et al., 2016).

2.3.1 Effect of Particulate Matter on Mortality

Although the anthropogenic air pollutants are widely recognised to have an adverse effect on health, long-term exposure to particulate matter, particularly $PM_{2.5}$ or less (Barrett, 2012) is believed to be the air pollution metric that is mostly linked to premature death. It also accounts for the bulk of the health costs of air pollution, (COMEAP, 2010, USEPA, 2011). The burden of particulate matter resulted in nearly 29,000 deaths in the UK and rose to 40,000 when the effect of nitrogen dioxide is included (Holgate et al., 2016, COMEAP, 2010). COMEAP (2010) also estimated that if all particulate matter ($PM_{2.5}$) from anthropogenic sources is to be removed, about 36.5million life years of UK Population will be saved in the next hundred years. The COMEAP approach was based on a combination of modelling and measurements of PM concentrations with a scheme designed to achieve "mass closure" relative to measured concentrations, established that $PM_{2.5}$ has the strongest correlation with mortality and for every 10-$\mu g/m^3$ increase in long-term $PM_{2.5}$ concentration, there is a 6% increase in the risk of deaths from all causes. Yim and Barrett (2012) evaluated and applied a multi-scale air quality modelling system to assess the impact of combustion emissions on UK air quality through the use of epidemiological evidence. Also, Yim and

Barrett (2012) quantitatively relate $PM_{2.5}$ exposures to the risk of early death and revealed that combustion emissions in the UK cause about 13,000 premature deaths annually with additional 6000 deaths caused by non-UK European Union combustion emissions. The principal contributor to this menace is the transport, which contributes about 7500 premature deaths per year with the remainder shared between industrial, power generation and other sources emissions (Barrett, 2012). These are numbers that should not be accepted by any progressive society, hence, the need for more in-depth research to quantify the ambient concentrations of particulates and their resulting effect on human health.

2.3.2 Air Quality Pollutants Inventory in the UK

Although UK air quality has been improving recently, there are still cases of exceedances of EU targets in three key air pollutants namely PM, NO_x and O_3 (NAO, 2009). Among these pollutants, particulate matter is thought to have the most impact on health even at lower concentrations (WHO, 2005). According to UK-AIR (2013), in the year 2012, there was a total of forty days on which very high and high air pollution was recorded in the UK. Thirty-six of these days were due to particulate (PM_{10} and $PM_{2.5}$), three due to ozone, one due to SO_2 and none due to NO_2. Although road transport-related air pollutants concentrations have been declining since 1990, the concentration of particulate matter and nitrogen dioxides were the least decline pollutants apart from ammonia which is largely coming from agricultural sources(Salisbury et al., 2014). Despite the uncertainties involved in the estimation of particulate matters in the UK particularly the fine fraction ($PM_{2.5}$ or less), the trend of NO_x and PM concentrations from road transport sources over these years followed the same pattern. This correlation is confirmed by (Sánchez Jiménez et al., 2012, Götschi et al., 2005). They found that ambient concentrations of NO_x and fine particulate matter especially those from secondary sources have a strong relationship. The uncertainty

attached to $PM_{2.5}$ concentration estimation might be because it is being measured in very few locations across the UK. Therefore, modelling will play a significant role in assessing concentrations for the future assessments (Whyatt et al., 2007).

Although most of the particulate matter measurements and inventory are largely based on PM mass concentrations (DEFRA, 2010), studies have shown that adverse health effects cannot only rely on total particulate mass but other metrics such as size, number and surface area (Zissis Samarasa and Hallb, 2005). Nowack and Bucheli (2007), revealed that particle size plays a significant role in defining toxicity of nanoparticles, but not much is known about the effect of size on the particle behaviour and reactivity. On the other hand, some epidemiological studies support the number concentration as a preferable metric for determining health effects of atmospheric particles. Pekkanen et al. (1997), demonstrated that there is a link between deficits in peak expiratory flow among asthmatic children and exposure to fine and ultrafine particles in an area where the dominant source of ambient particulate matter is traffic. They concluded that daily variations in black smoke and particle number concentrations (size range 0.032 - 0.32μm and 1.0 - 10.0μm) were strongly correlated (correlation coefficients 0.9). However, correlations with PM_{10} were relatively lower (below 0.7). Limbach et al. (2007) suggested that chemical and catalytic properties of particles should receive more attention than physical properties such as size, shape, and degree of agglomeration.

Smaller particles have been claimed to have caused more adverse effects than large particles. Nemmar et al. (2002), demonstrated that inhaled ultrafine particles spread rapidly into the blood circulatory system, and accumulate in the lungs, liver, bladder and other parts of the body. The rapid spread might be an important property to consider for the cardiovascular morbidity and mortality related to ambient particle pollution. Particles especially those in the nanoparticles range have high surface reactivity and also can cross cell membranes, this

might increase their negative health impacts although they might be desirable properties in engineered nanoparticles technology (Tetley, 2007). This reason has led some countries like the USA to revise the air quality standard for particulate matter to include measurement of fine particles (i.e. PM_{10} - $PM_{2.5}$) and to support evaluation of the best metric for air quality standards worldwide (AQEG, 2005). Tighter controls on particulate emissions from vehicles are currently entrenched in the Euro-3 and Euro-4 for LDVs and Euro 3,4, and 5 for HDVs in Europe (AQEG, 2005).

2.4 Air Quality Standards

Different international organisations are concerned with air pollution and have different air quality standards. The organisations include but are not limited to World Health Organisation (WHO), European Union (EU), United Nations Economic Commission for Europe (UNECE) and the United States Environmental Protection Agency (USEPA). These organisations and much more are using different terms such as standards, guidelines and limit values. For example, European limit values are mandatory and must be met by member countries (Colls, 2002a).

Currently, in most of the air quality standards around the globe, particulate matters are controlled based on their mass concentrations in terms of $PM_{2.5}$ and PM_{10} (particles with aerodynamic diameters less than or equal to 2.5 mm and 10 mm respectively). However, these standards are somewhat conservative in that they cannot normally account for the effect of particles of smaller aerodynamic diameter since they have less mass but consequently large population. For example, in an urban area, most of the ambient particles are being emitted by traffic, and most of them are within the ultrafine range. Therefore, providing standards by mass tends to undermine the effect of these smaller particles, which are mostly related to the adverse effect on health. Providing alternative metric that will

account for the shortcomings of mass concentration is necessary. Although some other suggestions have been made on particle size distribution, surface area and chemical composition as alternative metrics to mass concentration, particle number concentration can be a good choice since it will be based on the particle size that dominates the particle population (Kumar et al., 2010, Keogh et al., 2009).

2.5 Air Quality Modelling

Air quality models are tools that can be used to describe the underlying relationship between emissions from various sources, meteorology, atmospheric concentrations, deposition, and other factors(Nguyen, 2014). The models can be used to predict pollutant concentrations and evaluate the effectiveness of various air quality control measures. Air quality monitoring through measurements provides information on a quality of air at the place of the measurement. However, it cannot give information on what is likely to happen in future or in another location where there is no measurement, hence the need for air pollution models is indispensable.

The effects of traffic air pollution can be effectively controlled through provisions of adequate and efficient air quality control and mitigation measures which can be designed and tested with the aid of air quality models. Air quality regulatory agencies have to complement measurements of air quality with models that can be able to predict accurately pollutants concentrations and determine the cause of the air quality problems. The use of such models provides room for using historical data to study the past scenarios of air pollution episodes and forecast the likely scenario in future. Air quality models, when properly developed, can predict future pollutant concentrations with greater accuracy and also give information about the air quality of a particular place using available information related to the air pollutants (e.g. meteorological and traffic variables).

Also, air quality models can be used to provide information about different sources of air pollution and their percentage contribution to the overall air pollution in an area. It can also give an insight about which source contributes a particular pollutant at a certain receptor where measurement might have been difficult. The models can also be used to determine the dispersion mechanisms, the transformation processes, distribution and deposition of the air pollutants (Holmes and Morawska, 2006). Dispersion models, photochemical models and receptor models are the most commonly used air quality models in research and practice (Nguyen, 2014).

Air pollution models can be classified based on the nature of the source of the pollutants which include point source, area-wide, and line source models. They can also be classified based on the modelling techniques e.g. deterministic models, stochastic models and hybrid models (Gokhale and Khare, 2004). Dispersion models are based on the thorough understanding of physical, chemical and fluid dynamical processes in the atmosphere (Colls, 2002b). In this type of models, mathematical relationships are used to describe the physical, dispersion; chemical processes involved within the plume to estimate pollutants concentrations at different locations (Gokhale and Khare, 2004). The theory of dispersion models uses meteorological data, geometry, and strength of the source to provide a means for calculating the concentration of pollutants in the atmosphere. Dispersion models include; Eulerian models, Gaussian models, Lagrangian models (Colls, 2002b).

Receptor models are based on the correlation between the concentration of pollutants at the receptor and the concentration at various sources of air pollution that might affect the concentration at the receptor (Vallero, 2014).

Stochastic models use data on pollutants concentration, meteorological parameters (wind speed, wind direction, solar radiation) and other factors that might affect those

concentrations. Especially such factors (traffic volume, vehicle speed) that are related to the source of the emission to establish empirical relationships between these factors and the pollutant. The variance of the concentration about the mean value is attributed to a particular factor using statistical correlations. Then an estimate of a likely concentration in future is made if a value for each factor is known. These models are often updated as the new data become available and are being used for a real time and short term forecasting (Gokhale and Khare, 2004, Colls, 2002b).

The models discussed above formed the background of almost all the models being used in practice to predict future concentrations, to estimate the contribution of a particular source, to derive the relationship between various factors associated with the air quality and to determine the extent of the compliance with air quality regulations. The operational air quality models currently used in the regulatory agencies are mostly deterministic which are built based on the simple assumption of atmospheric processes and involve high computational cost. They also require knowledge of the relationships between various variables involved and meteorological condition. However, air quality models developed using machine learning methods such as Artificial neural network (ANN, if properly developed can address some of these challenges. The ANN based air quality models have been developed, and their various capabilities have been tested on different air pollutants (Sharma, 2005, Amirsasha Bnanankhah, 2012). ANN model is capable of modelling a complex and nonlinear relationship between a large number of variables with reasonable accuracy and less computational efforts especially with the aid of modern software and computers

2.5 Statistical Modelling Techniques

In this section, the commonly used statistical methods used for the air quality modelling are described. The methods include Multiple Linear Regression (MLR), Stepwise Regression, Lasso regression, Elastic-net Regression, Principal Component Regression (PCR) and Partial Least Square Regression (PLSR). These methods were selected for this research because of their popularity in many air quality studies and other environmental studies (Benas et al., 2013, Chen et al., 2013b, de Paula et al., 2015, Deka et al., 2016, Guo et al., 2016a, He et al., 2015).

Besides their popularity, the methods were considered based on their individual improvement over ordinary multiple linear regression. MLR is easy to model and interpret the relationships between the predictor variables and the response variables. However, despite its simplicity, MLR is not robust in handling the trade-off between bias and the variance in the least square estimates, it minimises only the bias component. Also, it describes only linear relationships, and it cannot handle a case when a number of samples are greater than the number of predictors resulting in overfitting and consequently poor predictions on future observations not used in model training (James et al., 2014). Also, it has difficulty in dealing with highly correlated variables (Kuhn and Johnson, 2013). Although, the remaining statistical methods are also linear in nature they were developed with various improvements over ordinary MLR, the basis which formed their inclusion in this research.

PLSR and PCR methods use principal component analysis to transform the feature space into new sets of uncorrelated variables. The principal component analysis reduces the dimensionality of the input space, and the reduction of the dimension of the input space decrease the requirements for capacity and memory, and an increased efficiency given the processes taking place in smaller dimensions. The newly created variables are expected to

be less sensitive to noise in the data and are formed to cater for the problem of highly correlated variables in MLR (Karamizadeh et al., 2013). The main disadvantages of PCA are that the covariance matrix is difficult to be accurately evaluated, and even the modest invariance could not be captured by the PCA unless the information is explicitly provided in the training data (Karamizadeh et al., 2013). Table 2.1 summarises the application, strength and weaknesses of the selected statistical models

Table 2.1. Details of the selected statistical methods selected for this research

Method	Application	Strength	Weaknesses
MLR	Linear models are applied to estimate any linear relationship between variables and have been used in many applications (Vlachogianni et al., 2011, Kuhn and Johnson, 2013)	Linear models are highly interpretable; their mathematical nature enables computing standard errors of the coefficients.	When collinearity exists within the data, the estimated regression coefficients are not unique, thus losing the ability to interpret the coefficients meaningfully. The solution of MLR is linear in the parameters. MLR cannot identify curvature or nonlinear characteristics in the data, and it is susceptible to outliers.
PCR/ PLSR	Same as in MLR	PCR and PLSR use PCA to reduce the dimension of the predictor variables thereby de-correlating the predictors, or their combinations. PLSR is most suitable when there are correlated predictors because it takes into account the relationships between the predictors and the response variables(Kuhn and Johnson, 2013).	The new predictors are linear combinations of the original predictors, and thus, the practical understanding of the new predictors can become murky. PCA does not consider any aspects of the response when it selects its components (Kuhn and Johnson, 2013).
Stepwise Regression	The stepwise regression method has been applied in many studies involving air quality and other studies (Banerjee et al., 2011, Brown et al., 2015, Chen et al., 2013b, Diaz-de-Quijano et al., 2014, Krivtsov et al., 2009).	The stepwise regression has an advantage in avoiding the collinearity issues of the MLR (Chen et al., 2013a). It has the capability of selecting the most important predictor variables for the estimation of the regression coefficients.	The major limitations of stepwise regression consist of bias in parameter estimation, inconsistencies among model selection algorithms, and dependence on a single best model (Whittingham et al., 2006).
LASSO/ Elastic-net	They have been applied in genetics (Waldmann et al., 2013), air quality (Suleiman et al., 2016), epidemiology (Sampson et al., 2013) and in variable selection (May et al., 2011).	Ordinary linear regression finds parameter estimates that have a minimum bias, whereas the lasso, and elastic net finds estimates that have lower variance. Lasso is somewhat indifferent to very correlated predictors, and will tend to pick one and ignore the rest. The elastic net has the effect of averaging variables that are highly correlated and then entering the averaged variable into the model (Friedman et al., 2010).	Lasso have problems with highly correlated variables and it might lead to breaking down in extreme cases. Lasso usually selects only one predictor and ignore the others. The lasso method cannot select more predictor variables than the sample size (Waldmann et al., 2013)

Stepwise regression is a popular modification of MLR with variable selection property which combines the backward and forward procedure. The predictor variables are tested for addition or removal from multivariate regression models using forward and backward stages respectively. The variables are retained or dropped based on their statistical significance. Lower and upper boundaries of p-values of F-statistics are set such that for a variable to be kept in the model or removed must satisfy those boundaries (Singh et al., 2012). The stepwise regression has an advantage in avoiding the collinearity issues of the MLR (Chen et al., 2013a). However, in some cases, the *p*-value threshold for adding and removing predictors can be somewhat different (Derksen and Keselman, 1992). Although this makes the stepwise procedure less greedy, it worsens a problem of repeated hypothesis testing. The major limitations of stepwise regression consist of bias in parameter estimation, inconsistencies among model selection algorithms, and dependence on a single best model (Whittingham et al., 2006). The stepwise regression method has been applied in many studies involving air quality (Banerjee et al., 2011, Brown et al., 2015, Chen et al., 2013b, Diaz-de-Quijano et al., 2014, Krivtsov et al., 2009).

The Lasso/Elastic–net (Zou and Hastie, 2005), although not so popular in the air quality studies (Simons et al., 2016, Suleiman et al., 2016), are forms of penalised regressions aimed at reducing the variances in the least square estimates by using bias-variance trade-off. The penalty is added to the sum of the squared errors as the estimates become large. This trade-off between variance and the bias ensures a modest reduction in the mean squared errors (MSE) which translate to a better estimate. The estimates shrink to zero when the penalty becomes large. Therefore, feature selection becomes possible as the predictors with zero coefficients are discarded. The ability of the Lasso/Elastic–net methods to carry out feature selection could help improve their predictive ability, and could help in providing air quality

modellers with simple tools that can be used to predict and analyse air quality issues with greater certainty.

2.6 Feature Selection

The feature selection methods can be broadly categorised into filter and wrapper methods. The wrapper methods consider the relationships between the predictor variables and response variables during their selection process while the filter methods select their variables without regards to the response variables. The advantage of the wrapper methods over filter method is that it reduces the number of predictor variables such that more efficient and interpretable model can be obtained. They used subsets of predictor variables as inputs while considering the performance of the models as the output to be optimised (Kuhn and Johnson, 2013). The advantage of filter methods is that they are faster. However, they do not consider the efficiency of the models during the selection process. The wrapper methods are slow therefore require more computational effort than the filter methods. Moreover, there is also a risk of overfitting when using wrapper methods as they aggressively search the dataset. In this work, the two wrapper methods namely: Genetic algorithms (GA) and Simulated Annealing (SA) are considered.

2.6.1 Genetic Algorithms (GA)

A genetic algorithm is one of the methods that mimics the biological evolutionary processes (Fouskakis and Draper, 2002). The algorithms are based on the biological reproduction principles where the training data sets are considered to represent the population, and the data subsets are considered as individual candidates that undergo reproduction process to produce offsprings. The candidates in the biological context are chromosomes that consist of genes and are evaluated based on their fitness. The fittest chromosomes are selected to bear offspring. During this process, the heads of the chromosomes are exchanged

(i.e.crossover) and move to the mutation stage where the new chromosomes are randomly selected, and their genes are altered. In the framework of feature selection, the chromosome can be taken as a binary vector with a length equal to the number of predictor variables in the training data. The binary values in the vector indicate whether a particular predictor is present or absent in the data. The fitness of the set of the predictor variables indicated by the binary vector is determined by the model using them to fit the data. The GAs, in this case, are used to optimise the solutions from the 2^n possible combination of set of predictor variables (Kuhn and Johnson, 2013). The crossover phase restricts the search around the fittest variables which can cause overfitting. However, the mutation phase corrects the overfitting by randomly altering the binary values using probability of mutation. This probability is kept very low to ensure optimal solutions.

2.6.2 Simulated Annealing (SA)

The Simulated annealing method is a global search technique that mimics the metal cooling process (Lin et al., 2008). The algorithm randomly makes small changes to the initially selected subset of predictor variables. The perturbed subset is then used to create a model, and the initial error is estimated. The same procedure is repeated, and the error for the new model is compared with the previous error. If the performance of the new model is better than the previous model, then the current set of predictors is accepted. Otherwise, a probability of acceptance is determined based on the difference between the performance of the two models and the current iteration of the search. The probability is estimated such that it decreases as the number of iterations become large making it difficult for a suboptimal model to be accepted. The process is repeated until the specified number of iterations is reached, and the optimal combination of predictors is determined.

2.6.3 Random Forests

Random forests method is one of the variants of ensemble learning techniques designed to improve the prediction accuracy of regression trees (Breiman, 2001). Bagged Regression trees are built using bootstrapped subsets of the training data so that the final model is the average of all the individual trees. The out of bag errors of the individual trees are estimated using the remaining samples of the training data left during the resampling process. The averaging of the trees reduces the overall variance in their estimates. However, the trees are correlated in one way or the other which limits the reduction of the overall variance in the estimates. Random forests method seek to de-correlate the trees by introducing randomness in the tree building process. The algorithm first selects the predictor variables at random and then selects the best predictors out of the random samples to partition the data. This process reduces the variance in the estimates of the individual trees and thus, reduces the overall variance in the final estimate.

2.6.4 Hybrid Feature Selection Methods

The hybrid feature selection methods combine the powers of the search algorithms and the random forests i.e. GA and RF on one hand and SA and RF on the other hand. This combination is aimed at using the capabilities of the search algorithms in finding the possible subsets of the predictor variables that will optimise the out of bag errors estimated by the random forests. The combination reduces the chances of overfitting the estimate of the internal performance since the out of bag errors are estimated using unseen data samples. The external performance is estimated using 10 – folds cross-validation repeated five times.

2.7 Machine Learning Methods

Machine Learning (ML) systems are sets of algorithms seeking to find out automatically how to perform a certain task from the set of training examples presented to them in the training data. The main focus of ML discipline is to answer two highly interconnected questions (1) How to design a system that can improve its performance based on experience? (2) What are the underlying laws governing all the learning systems including human, computer, and organisations? (Jordan and Mitchell, 2015). In searching for the answers to these scientific, engineering and practical questions, many algorithms were developed to cater for the wide variety of problems and data types across many disciplines (Hastie et al., 2008b). The study of ML methods gained much attention in the last two decades with many focusing on the development of new algorithms that can handle much more complex tasks such as image processing, speech recognition, drug discovery and object detection to mention a few (LeCun et al., 2015).

The most widely used machine learning methods include but not limited to Artificial Neural Networks (ANN), Ensemble learning trees e.g. Boosted Regression Trees (BRT) and Random Forests (RF), and Support Vector Machines (SVM). The ML methods are broadly categorised into supervised and unsupervised learning methods. The supervised learning methods produce their prediction y for each x by mapping $f(x)$. During the training, the algorithms are presented with the samples of the outputs y given the values of x. However, in unsupervised learning, the samples of the outputs are not provided and the algorithms are tasked with finding the underlying structure of the data. Unsupervised learning methods are usually employed for dimension reduction and feature estimations, although their implementations in the actual prediction is actively being researched as it is closer to biological learning process than the supervised learning (Bengio et al., 2013).A typical ML process involved data representation, evaluation, and optimisation with the main goal of

achieving generalisation on the unseen data. The main challenge for the most of the ML methods are the issues of overfitting and inference (Domingos, 2012). The focus of this research is supervised machine learning and henceforth ML will mean supervised ML unless expressly stated otherwise.

2.7.1 Application of Machine Learning Methods in Air Quality Modelling

The application of Machine Learning techniques in air quality modelling dates back to the 90s (Yi and Prybutok, 1996, Gardner and Dorling, 1998). The search for the more accurate and easy to use models than the operational air quality models lead to the application of various ML methods in air quality modelling. The operational air quality models require the knowledge of the relationships between the variables involved and meteorological conditions. The models are mostly deterministic and are not only limited by the accurate characterisation of the dynamics of the natural phenomenon but also on the model configuration options. For example, the use of default parameters and the lack of real observations with the same spatial resolution with which to compare the model outputs (Chave and Levin, 2003, National Research Council, 2007).

Most of the operational air quality models are based on the steady-state Gaussian plume models and have been applied successfully in many air quality studies. However, despite their successful application, they are limited by the assumptions regarding the change of wind and source emission over time and do not include the detailed chemistry of particle pollutants (Lagzi et al., 2013, Pelliccioni and Tirabassi, 2006). Other sources of uncertainty in the operational models are the inherent uncertainty associated with data required to run those models. For example, the models rely on the emission rates estimated by the emission models that in most cases accommodate up to $\pm 50\%$ uncertainties (Debry and Mallet, 2014). Computational time and effort are also part of the constraints that lead to the simplification of the operational models.

In contrast, machine learning methods such as Artificial Neural Networks (ANN) and Boosted Regression Trees (BRT) can be used to build air quality models with comparable prediction accuracy at a lesser computational cost and with no assumption of the atmospheric processes involved (Gardner and Dorling, 2000). ANNs are capable of handling complex and robustly nonlinear relationships that exist between air quality variables (Esplin, 1995) and produce models that perform extremely well in predicting an unseen data (Elangasinghe et al., 2014b, Balsamà et al., 2014, He et al., 2015). One of the major challenges of using machine learning methods is the feature selection (variable elimination). Feature selection process helps in understanding the data, reduction of the computational requirements, reducing the effect of the curse of dimensionality and improving the prediction performance (Chandrashekar and Sahin, 2014). Table 2.2 summarises the strength and the weakness of the machine learning models selected for this study.

Table 2.2. Details of some selected machine learning models

Method	Application	Strength	Weaknesses
MLP	All problems involving classification and regression e.g. Pattern recognition (Dervilis et al., 2014), stock market index prediction (Moghaddam et al., 2016), image processing, air quality forecasting (Abderrahim et al., 2016) etc.	It can estimate any nonlinear relationships between variables with reasonable accuracy. It is considered as universal approximator. Shallow architectures have been Shown effective in solving many simple or well-constrained problems (Deng, 2014)	Solutions are trapped in local minima and it is susceptible to overfitting. It has slow training speed and lack proper generalisation. It is also difficult to choose model parameters. Neural networks are not a suitable tool for feature selection (Vidyasagar, 2015). They have limited modelling and representational power making them difficult to deal with natural signals such as human speech, natural sound and language, and natural image and visual scenes.
SVM	Same as in MLP	It can generalise easily and better than traditional MLP. It has the ability to minimise the effect of outliers. It uses small number of training samples to generalise.	It uses quadratic programming in the training which requires high computational capability. It is also scale sensitive. The kernels required has to satisfy mercer conditions (Torija and Ruiz, 2015).
ELM	Pattern recognition, image processing, sensors, signal processing and automatic control have significant results (Ding et al., 2011a, Quteishat and Lim, 2008, Zhang and Wang, 2009). Biomedical/Medical, Image/video understanding and processing. System modelling and prediction, Control and Robotics, Chemical process, Time series analysis, Fault detection and diagnosis and Remote Sensing (Huang et al., 2015).	It is a simple single layer feedforward network that can accept many training algorithms. It can achieve good generalisation at extremely fast learning speed. No iterative tuning of hidden layer parameters is required. It has higher scalability and require less computational complexity and can be applied to big data (Ding et al., 2015). Its proponents view it as the real universal approximation (Huang et al., 2012)	Requires a higher number of hidden neurons. It will also encounter some problems with small sample data set and also it is prone to overfitting phenomenon. ELM randomly select the input weights and hidden layer biases which may lead to non-optimal input weights and hidden layer biases (Huang et al., 2015).

Table 2.2 *continued*

Method	Application	Strength	Weaknesses
DNN	Performed better than shallow architectures in semantic parsing, transfer learning, natural language, processing, computer vision, visual recognition, object recognition (Google Goggles), image and music information retrieval (Google Image Search, Google Music), computational advertising (Dean et al., 2012, Bordes et al., 2012, Cireşan et al., 2012, Ren and Xu, 2015, Mikolov and Dean, 2013, Ciregan et al., 2012, Krizhevsky et al., 2012).	Their deep architecture allows them to discover more abstract features of the data making them be suitable for real life problems such as images, video and audio signals processing (Guo et al., 2016b) It properly initializes the network, which prevents poor local optima to some extent. Training is unsupervised, which removes the necessity of labelled data for training. Deals more robustly with ambiguous inputs by incorporating top–down feedback and produces better generative models by allowing the lower layers to adapt to the training of higher layers. The training process can be scaled and can handle much bigger data set than in the case of shallow MLP (Bengio, 2013) Discovery of abstraction, with the belief that more abstract representations of data such as images, video and audio signals tend to be more useful (Guo et al., 2016b).	Due to the initialization process, it is computationally expensive to create a deep network model. Also, Training deep supervised neural networks is difficult with less computing capacity. Therefore, it requires higher computational capacity (Bengio, 2013). The joint optimisation is time-consuming. Despite its success in computer vision, the underlying theory is not well understood, and it is not clear how to select the best-performing architecture (Guo et al., 2016b). It also need specific knowledge to choose sensible values such as the learning rate, the strength of the regularizer. Shortage of training data may limit the size and learning the ability of such models, especially when it is expensive to obtain fully labelled data (Guo et al., 2016b).
BRT /RF	Same as in MLP	The final prediction of RF and BRT is based on an ensemble trees. It is not necessary to pre-select or transform predictor variables. They are also resistant to outliers (Moisen et al., 2006). BRT and RF can handle high dimensional spaces as well as a large number of training examples very well. Though, RF is more computationally efficient, although it requires more trees. The RF algorithm is highly resistant to over-fitting and has consistent performance. It is robust to noisy and missing data, faster to train and robust towards parameter setting. It helps in identifying most important predictor variables (Malhotra, 2015).	BRT could be susceptible to over-fitting because the individual learners are susceptible to overfitting. Despite using weak learners, boosting still employs the greedy strategy of choosing the optimal weak learner at each stage. Computation time for boosting is often greater than for random forests since random forests can be easily parallel processed given that the trees are created independently.

2.7.2 Artificial Neural Network (ANN) modelling

ANN models are designed to mimic the behaviour of the human brain. The human brain is made up of interconnected synaptic neurones that are capable of learning and storing information about their environment (Bishop, 1995). A neuron model is made up of three elements, the connecting links characterised by their strength and a linear combiner that combines the weighted input signals. Moreover, it has an activation function for limiting the amplitude range of the neuron's output to some finite value. The commonly used ANN method is Multilayer Perceptron Network (MLPN) trained using a back propagation algorithm (BP). The MLPN method involves designing an appropriate neural network architecture consisting of serially interconnected layers, training the network on a training data and testing the network on a test data set. The network layers include input layer where the input variables are received and the hidden layer where the sum of the weighted inputs from the input layer are received through the connecting links of various weights. The weighted inputs are transformed into a higher dimension using the hidden layer activation function (e.g. sigmoid function). The last layer is the output layer where the outputs of the hidden layer are received through connecting links, and the final output of the network are estimated using output layer activation functions which are usually linear (see Figure 2.2).

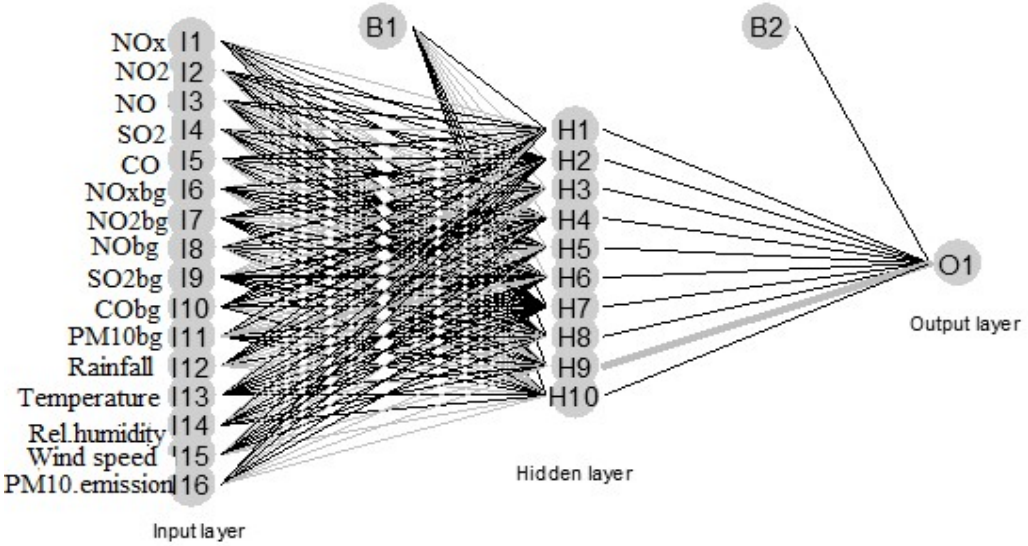


Figure 2.2. The typical structure of a multilayer neural network.

Note: The suffix "bg" in some of the variable names indicates background concentration.

The network outputs are then compared with the target samples, and the errors are estimated and propagated back to update the previously estimated network weights. This procedure is repeated up until the network with minimum error, and good generalisation is obtained.

The main task in the design of the neural network is the determination of an appropriate number of hidden neurons and the selection of input variables that will produce a model with the desired generalisation and prediction accuracy. The ANN methods have been successfully applied in many air quality studies (Taspinar, 2015, Ragosta et al., 2015, Elangasinghe et al., 2014b). They were used in studies involving predicting and forecasting of air pollutants ranging from the current hour to several days in advance (Russo et al., 2013, de Gennaro et al., 2013). Many studies involving neural network often used cross-validation or evolutionary algorithms such as Genetic algorithm and particle swarm optimisation methods to derive an optimum architecture for the ANN models (Ding et al., 2011c, Ding et al., 2011b, He et al., 2014). Also, the input selection methods often used in air quality studies

with ANN include PCA (Taspinar, 2015, Ragosta et al., 2015), stepwise regression (Russo et al., 2013, Lima et al., 2013), and cluster analysis (Elangasinghe et al., 2014b). The use of elastic-net and lasso regressions combined with ANN was also investigated by Suleiman et al. (2016). However, the conventional BP neural network has slow learning speed, and it is difficult to choose a proper size of the network. Also, its solutions can be easily trapped into local minima (Wang et al., 2015b). Several proposals have been made to improve upon the shortcomings of the traditional BP neural network or MLPN. These developments leads to the creation of several algorithms such as Support Vector Machines, (SVM) by Vapnik (2000), Extreme learning machine (ELM) (Huang et al., 2006, Huang et al., 2014) and Deep Learning (DNN) (Hinton et al., 2006, Schmidhuber, 2015) to mention a few. Other modifications include hybrid ANN algorithms where the traditional MLP is combined with some other algorithms to improve the performance of the traditional MLP algorithm especially in variable selection (May et al., 2011). In this research five, MLP modifications were used based on their individual ways of handling the shortcomings of MLP and their popularity. The selected methods include Multilayer perceptron with principal component analysis (PCA-MLP) (Santos et al., 2015, Balsamà et al., 2014, Taspinar and Bozkurt, 2014), Neural Networks Using Model Averaging (AVG-MLP) (Tumer and Ghosh, 1996, Perrone et al., 1993), Bayesian Neural Network (BRNN) (Dan Foresee and Hagan, 1997, Hernández-Lobato and Adams, 2015), ELM and DNN.

The PCA-MLP is the combination of MLPN and the Principal Component Analysis (PCA). The PCA was used to transform the input variables into new uncorrelated variables called Principal Components (PCs). Pre-processing the data in this way allows for the reduction of the dimensionality of the input space that is expected to enhance the performance of the MLP models. The PCs formed the inputs to the MLP models. Ul-Saufie et al. (2013) Combined the Feedforward Backpropagation (FFBP) with principle component analysis to

predict the next day, next two-day and next three-day PM_{10} concentration in Negeri Sembilan, Malaysia. The results obtained indicates that the performance of the FFBP models combined with PCA improved with an error reductions ranging between 12 – 46%. Santos et al. (2016) estimated the levels of lead (Pb), nickel (Ni), manganese (Mn), vanadium (V) and chromium (Cr) northern in France using principal component analysis (PCA) coupled with ANN. They suggested that both techniques can be considered acceptable air quality assessment tools for heavy metals in the studied area. However, the application of the PCA before ANNs training did not produce any improvements in the performance of the ANNs.

AVG-MLP combined several MLP networks which are trained with different random initiations, and their outputs are averaged to obtain a single output. The averaging is expected to provide trade-offs between the bias and variance. Therefore, the resulting output would have low bias and variance (Perrone et al., 1993, Siwek and Osowski, 2012).

The BRNN is a single hidden layer back propagation neural network implemented in the Bayesian framework. It regularised the sum of the squared error function to reduce the chances of overfitting and improve generalisation of the conventional back propagation neural network. The method imposes a constraint on the size of the neural network weights (i.e. regularisation) during the training. The principle behind the regularisation is to keep the size of the network's weights small so that the response of the network is smoothened (Dan Foresee and Hagan, 1997). This method has been successfully applied in air quality studies (Hoi et al., 2010, Tijani et al., 2016).

The Extreme learning machine (ELM) is a relatively new method for training single layer feed forward network that was designed to be more efficient than the traditional neural networks and support vector machines (SVM)(Huang et al., 2006, Huang et al., 2014). ELM seeks to offer a universal learning framework with least human input, computational efficiency and higher learning accuracy (Huang et al., 2015). The major drawback of the

traditional ANN trained using back propagation algorithms (Williams and Hinton, 1986), is that it is slow optimisation process, and the solutions are often trapped in local minima. Various improvements for the back propagation have been proposed (Hagan and Menhaj, 1994, Wilamowski and Hao, 2010, Kang et al., 2005, Branke, 1995). Though, much faster with increased generalisation, but their solutions could not be guaranteed as global optimal solutions (Huang et al., 2015). However, ELM, in its new approach adopted random initiation of the hidden layer nodes without parameter tuning and proposed that learning be only necessary for the weights connecting hidden layer and the output layer (Huang et al., 2014).

The performance of the basic ELM method and its variants have been compared with the SVMs (Huang et al., 2006) and traditional feedforward neural networks (Huang et al., 2012) and were found to have performed similarly or better despite their efficiency and fast training speed. The basic formulation of ELM consists of two major steps (1) random initiation of the hidden layer parameters (weight and biases) and (2) the linear estimation of the output weights by minimising the sum of squared losses of prediction errors (Huang et al., 2014). The ELM method has been applied in developing a warning system on the levels suspended particulate matter (Vong et al., 2014). The performance of the ELM warning system was compared with that of the SVMs and concluded that the ELM method was more accurate, efficient and faster than the SVMs. Lima et al. (2015) compared the speed and accuracy of extreme learning machines with ANN, SVM, MLR and RF on several environmental data sets including SO_2 data. They concluded that ELM was the fastest nonlinear model out of those tested using the smaller dataset, but RF was the most accurate and faster when dealing with large data sets.

The Deep learning framework allows for machine learning models with multiple processing layers consisting of several nodes to learn from the data with many levels of feature space

transformations (LeCun et al., 2015). These high levels abstractions allow the models to discover hidden features from the input space. The computational model selected for the application of deep learning framework is feedforward neural network and from now on it will be referred as Deep Neural Network (DNN). Deep learning is currently under active research and has recorded tremendous successes in the areas of image recognition (Krizhevsky et al., 2012), speech recognition (Szegedy et al., 2014), genomics (Leung et al., 2014) and language translation(Jean et al., 2014). The main advantage of deep learning is that it was designed to provide training stability, generalisation and scalability when dealing with big data. Ong et al. (2016) applied deep recurrent neural network in a time series prediction of $PM_{2.5}$ and concluded that it outperformed the $PM_{2.5}$ prediction system in Japan.

2.7.3 Support Vector Machine

A support vector machine (SVM) is one of the early statistical learning techniques (Vapnik, 2000) originally developed to solve binary classification problems and later extended to regression problems. The main advantage of the SVM over MLPN is its good generalisation ability, attained at a relatively small number of training data and large number of input nodes. SVM for regression is a robust technique that minimises the effect of outliers on the regression equations. The aim of Support Vector Regression (SVR) is to find the optimal hyperplane that minimises its distance to all the data points. The performance of the SVR depends on the user-defined parameters cost parameter C the insensitive zone values ε and the kernel function parameter. The choice of the kernel function parameter depends on the type of kernel function used which also depends on the software platform employed and may also reflect the distribution of the input variables in the training data (Cherkassky and Ma, 2004). The SVM method have been applied in many studies involving air quality ((Suárez Sánchez et al., 2011)

2.7.4 Ensemble Regression Trees

Regression trees are simple models that fit a response variable to predictor variables by partitioning the feature space using a series of partition rules (e.g. binary split). The partition rules are used to identify regions in the data having the most consistent responses to the predictor variables and then fit a constant (usually mean response for observations in a particular region for regression problem) to each region. Ensemble Regression Trees e.g. Boosted Regression Trees (BRT) and Random Forests (RF) are among the latest advancement in Classification and Regression Trees (CART) in which the strengths of several weak learners (regression trees) are ensemble together to model the relationship between the predictors and response variables with a view to achieving a better prediction performance. BRT derive its strength from two different algorithms; Regression trees and

Boosting while RF is an improvement over the Bagged regression trees which introduced randomness in the tree building process to reduce the correlation between the trees and hence improve the prediction accuracy. Details about these methods can be found in (Elith et al., 2008, Kuhn and Johnson, 2013). The BRT and RF techniques have been applied to identify air pollution sources and to predict the urban air quality of Lucknow (India) using the air quality and meteorological databases pertaining to a period of five years. The methods were found to have predicted the seasonal air quality indexes more accurately than the SVM. Both the BRT and RF have also been applied in the prediction of ultrafine particles using data collected from an aerial campaign, and they were found to have predicted the concentration levels of the UFP accurately (Pandey et al., 2013).

2.8 ADMS-Roads

The ADMS-Roads, a variant of the atmospheric dispersion modelling system (ADMS), is a computer-based software for modelling the dispersion of gaseous and particulate emissions from traffic and industrial sources to the atmosphere. The modelling in ADMS-Roads is achieved through the use of point, line, area, or volume source models. It is designed to allow the user to select the type of sources to be modelled and can accommodate simple cases involving say, single road source, as well as complex cases involving a large combination of road and industrial sources. For example, ADMS-Roads can be used to model up to 150 road sources each with 50 vertices and 35 industrial (3 points, 3 lines, 4 areas and 25 volumes) sources. In its formulation, the effect of vehicle wake, traffic induced turbulence and street canyons on the dispersion of road traffic emissions are incorporated.

Furthermore, the model allows the inclusion of the chemistry involving NO, NO_2 and Ozone, and the generation of sulphate particles from SO_2. The model used up-to-date parameterisations of the boundary layer structure based on a boundary layer height, Monin-

Obukhov length, a length scale dependent on the friction velocity and the surface heat flux to obtain a realistic representation of the changing characteristics of the dispersion with height. One important feature of the ADMS system is the integrated meteorological pre-processor that estimates boundary layer parameters from the various combination of input data: e.g. day, time of the day, wind speed and cloud cover or boundary layer height, surface heat flux and wind speed (CERC, 2013). The pre-processed meteorological variables can be extracted from the model for further analysis of the model outputs.

The ADMS-Roads model (Carruthers et al., 1997) is the most widely used air quality models by the UK's local authorities. Besides its popularity in the UK, ADMS models have been applied successfully in many air quality studies (Silva and Mendes, 2009, Righi et al., 2009, Hirtl and Baumann-Stanzer, 2007, Carruthers et al., 2001, Carruthers et al., 1994). ADMS Roads models have an active user community in the UK (Williams et al., 2011). Other similar operational air quality models used in the UK include OSPM (Hertel and Berkowicz, 1989), Community Multiscale Air Quality (CMAQ) modelling system (Byun et al., 1998) and AERMOD (Cimorelli et al., 1998). A comprehensive review of the commonly used air quality models in the UK can be found in Williams et al. (2011). Carslaw et al. (2014) compared the performance of some of these models in the predictions of NOx, NO_2, O_3, PM_{10} and $PM_{2.5}$ and concluded that the differences between the predictions of the models might be minimised if the inputs to the models are similar.

2.9 Models Evaluation

The last step in the model development is the evaluation of the model (Nagendra and Khare, 2006). The first task in model evaluation is defining the model objectives against which its performance will be measured (Jakeman et al., 2006, Bennett et al., 2013). It is imperative to use more than one performance criteria because the model to be evaluated might have

more than one objective, or the end user might want to use some different performance criteria to evaluate the model accuracy. Also, using many evaluation methods will help in augmenting the limitations of one method or the other (Chang and Hanna, 2004). To minimise subjective human judgement, quantitative tools are often used to estimate suitable numerical metrics that will give the overall performance of the model. These are often preceded by using graphical methods that represents the sensitivity of the model in the form of graphs, charts, or surfaces, which can be used to complement the results of mathematical and statistical methods for better representation (Fei et al., 2005).

Recently air quality modellers are increasingly exploring the use Artificial intelligence methods for air quality modelling due to their inherent advantages over most of the deterministic models recommended for regulatory purposes. If these emerging air quality models are to be accepted as operational models, they must be properly evaluated to command the confidence of the users. The performance of air quality models is often accessed statistically in terms of their fraction of predictions within the factor of two (FAC2) of the observations, coefficient of determination (R^2), Mean Bias (MB) and Normalised Mean Bias (NMB) as recommended by Derwent et al. (2010).

These performance statistics do not reveal the strengths and weaknesses of the models; rather they give a general idea of the performance of the models, hence the need to invoke techniques that will give more insight on the accuracy of a model. Other performance measures often used for environmental model evaluations include Fractional Bias (FB), Mean Absolute Error (MAE), Index of Agreement (IOA) and Root Mean Squared Error (RMSE) (Bennett et al., 2013). Model evaluation depends on the model objectives (Jakeman et al., 2006, Bennett et al., 2013). It is a common practice to have a model with more than one objective; therefore, different performance criteria must be evaluated. Graphical methods such as scatter plots, Taylor's diagram, conditional quantile plots, time variation

plots and polar plots have been used in many air quality studies (Thunis et al., 2012c, Thunis et al., 2012a, Dore et al., 2015, Carslaw D et al., 2013). These graphics give a representation of sensitivity in the form of graphs, charts, or surfaces, which can be used to complement the results of mathematical and statistical methods for better representation (Fei et al., 2005). Carslaw and Ropkins (2012) provided some straightforward and commonly used numeric model evaluation statistics in Openair package of R statistical software (R Development Core Team, 2015). The statistics include: Fraction of predictions within a factor or two (FAC2), Mean Bias (MB), Mean Gross Error (MGE), Normalised Mean Bias (NMB), Normalised Mean Gross Error (NMGE), Root Mean Squared Error (RMSE) and the correlation coefficient, R. others are Index of Agreement (IOA) and Coefficient of Efficiency (COE). Also, graphical functions including scatter plots, Taylor's diagram, conditional quantile plots, time variation plots and bivariate polar plots have been formulated and provided in the software.

2.9.3 Statistical Evaluation Metrics

2.9.3.1 Coefficient of correlation R and Coefficient of Determination (R^2)

The Coefficient of correlation R is a measure of collinearity between observed and modelled values often reported as the coefficient of determination (R^2) which indicates how much of the total variation in the observation is explained by the model prediction. The values of R range between -1 and 1 while that of the R^2 range between 0 and 1. A perfect model is expected to have an R or R^2 value of 1 while a value near 0 indicates little or no relationship between the modelled and observed variables. Although these metrics may appear attractive, their magnitudes are not consistently related to the predictive ability of the model especially when the prediction accuracy is defined as the degree to which the model predictions match the magnitude of the observations(Willmott, 1982). It is a common practice to report the statistical significance associated with these metrics to aid in interpreting the correlation

coefficients. However Willmott and Wicks (1980) demonstrated that statistical significance of R and R^2 might be misleading since they are not related to the sizes and differences between observed and modelled values. Willmott (1982) discourage the use of R or R^2 as part of the model performance measures due to ill-defined relationship and inconsistencies between the values of R or R^2 and the model performance. However, several model evaluation tools include these metrics as part of the array of model performance measures and are still being reported in recent researches involving model evaluation (Bennett et al., 2013, Thunis et al., 2012b, ASTM, 2010, Carslaw and Ropkins, 2012).

2.9.3.2 Root mean square error (RMSE)

The Root mean square error (RMSE) is a measure of an average error produced by a model and is among the best measures of overall model performance which can be easily interpreted since they carry the same unit as the modelled and observed values. Although it is sensitive to extreme values, it reveals the actual size of the error produced by the model unlike R^2 which is affected by the higher and low standard deviations of both observed and modelled values, however it does not reveal the types or sources of the error which will assist greatly in refining the models (Willmott, 1982, Willmott, 1981). The RMSE value varies between 0 and ∞, with 0 being RMSE for an ideal model. To enhance the understanding of the predictive ability of a model RMSE can be further broken into two major components systematic ($RMSE_s$) and unsystematic ($RMSE_u$), after fitting a line by a least square regression. $RMSE_S$ (model oriented error) describes the linear bias between observations and the model and is estimated by the difference between the expected predictions and the actual predictions. $RMSE_U$ (data oriented error) is a measure is a measure of the scatter about a regression line of estimated by the difference between the actual prediction and the expected predictions and the actual observations(Nagendra and Khare, 2005, P. Thunis, 2011).

2.9.3.3 Fraction of predictions within a factor of two (FAC2)

The Fraction of predictions within a factor of two (FAC2) of the observation is a measure of the fraction of the model prediction that falls in between the ½ times and 2 times the measured values (Dore et al., 2015). Chang and Hanna (2004) described FAC2 as a robust performance measure since it is not affected by outliers.

2.9.3.4 Mean Bias (MB)and Normalised Mean Bias(NMB)

The Mean Bias (MB) is the measure of the model under or over prediction estimated as the difference between the mean observed and the mean predicted values. It is estimated using the following relationship. MB values range from -∞ to + ∞ with 0 being MB value for an ideal model. Although MB is being used as model performance measure its major weakness is that it does not provide more diagnostic value than the mean values of observed and predicted. Willmott (1982) suggested that the mean values of observed and predicted should be reported instead of MB since they are more familiar to the researchers and contain little more information than MB. The is a normalised version of MB it is often used when comparing different pollutants concentration scales.

2.9.3.5 Mean Gross Error (MGE) and Normalised mean gross error (NMGE)

The Mean Gross Error (MGE) Is a measure of model error regardless of whether it is under or over prediction and it has the same unit as a model and observed values. Normalised mean gross (NMGE) error has the same interpretation as MGE with added advantage when comparing pollutants of different unit scales.

2.9.3.6 Coefficient of efficiency COE

The Coefficient of efficiency COE is the measure of model efficiency that is robust and easy to interpret (Legates and McCabe, 2012). This measure has an interpretation for zero and negative values. A perfect model has COE value of one. Zero values of COE indicate that the model's prediction accuracy is not more than the observed mean values of the data, and negative COE values indicate that the model's prediction accuracy is worse than the observed mean. It can be estimated using the following relationship.

Table 2.3 Statistical evaluation metrics

Name	Formula	Range	Ideal value	Strength	Weaknesses
Coefficient of correlation R	$\frac{\sum_{i=1}^{N}(M_i - \bar{M}) \times (O_i - \bar{O})}{\left[\sqrt{\sum_{i=1}^{N}(M_i - \bar{M})^2}\right]\left[\sqrt{\sum_{i=1}^{N}(O_i - \bar{O})^2}\right]}$	(-1, 1)	1	The coefficient of correlation measures the correlation of the measured and modelled values.	The magnitudes of R and R^2 are not consistently related to the predictive ability of the model especially when the prediction accuracy is defined as the degree to which the model predictions match the magnitude of the observations (Willmott, 1982).
Coefficient of Determination (R^2)	$\left(\frac{\sum_{i=1}^{N}(M_i - \bar{M}) \times (O_i - \bar{O})}{\left[\sqrt{\sum_{i=1}^{N}(M_i - \bar{M})^2}\right]\left[\sqrt{\sum_{i=1}^{N}(O_i - \bar{O})^2}\right]}\right)^2$	(0, 1)	1	Coefficient of determination it has the same interpretation of results, except its values range between 0 and 1.	Negatives to this model are linear model assumptions and the fact it can return an ideal result for a model with constant offset.
Root Mean square error (RMSE)	$RMSE_S = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(M_i - \hat{M}_i)^2}$	(0, ∞)	0	Root Mean Square Error (RMSE) is a measure of an average error produced by a model and is easy to interpret. It has the same unit as the modelled and observed values. Although it is sensitive to extreme values, it reveals the actual size of the error produced by the model, unlike R^2 which is affected by the higher and low standard deviations of both observed and modelled values.	Squaring the data may cause bias towards large events. Also it does not reveal the types or sources of the error which will assist greatly in refining the models (Willmott, 1982, Willmott, 1981).

Table 2.3 *continues*

Fraction of predictions within a factor or two (FAC2)	$FAC2 = 0.5 \leq \frac{M_i}{O_i} \leq 2.0$	(0, 1)	1	This is a measure of the fraction of the model prediction that falls between the ½ times and 2 times the observations. Chang and Hanna (2004) described FAC2 as a robust performance measure since it is not affected by outliers	
Mean Bias (MB)	$\frac{1}{N}\sum_{i=1}^{N}(M_i - O_i)$	-∞, ∞	0	Calculates the mean error.	The result of zero does not necessarily indicate low error due to cancellation. Although MB is being used as model performance measure, its major weakness is that it does not provide more diagnostic value than the mean values of observed and predicted.
Normalised mean bias (NMB)	$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{M_i - O_i}{\sum_{i=1}^{N} O_i}\right)$	-∞, ∞	0	NMB is a normalised version of MB, and it is often used when comparing different pollutants concentration scales.	Same as MB
Mean Gross Error (MGE)	$\frac{1}{N}\sum_{i=1}^{N}\lvert(M_i - O_i)\rvert$	-∞, ∞	0	Mean Gross Error is a measure of model error regardless of whether it is under or over prediction and it has the same unit as a model and observed values.	This reduces the bias towards large events; however, it also produces a non-smooth operator when used in optimisation.

Table 2.3 *continues*

Normalised Mean Gross Error ($NMGE$)	$\frac{1}{N}\sum_{i=1}^{N}(\frac{\lvert(M_i - O_i)\rvert}{\sum_{i=1}^{N} O_i})$			Normalised mean gross error (NMGE) has the same interpretation as MGE with added advantage when comparing pollutants of different unit scales.	
Coefficient of Efficiency (COE)	$1 - \frac{\sum_{i=1}^{N}\lvert(M_i - O_i)\rvert}{\sum_{i=1}^{N}\lvert(O_i - \bar{O}_{i_i})\rvert}$	(-∞, 1)	1	It compares the performance of the model to a model that only uses the mean of the observed data. It has an interpretation for zero and negative values. Zero values of COE indicate that the model's prediction accuracy is not more than the observed mean values of the data, and negative values indicate that the model's prediction accuracy is worse than the observed mean.	This metric suffers a potential bias. A model may have a significant offset and still yield ideal values of these metrics (Bennett et al., 2013) The efficiency coefficient is sensitive to extreme values and might yield suboptimal results when the dataset contains large outliers in it.
Index of Agreement (IOA)	$A\begin{cases} 1 - \frac{\sum_{i=1}^{n}\lvert M_i - O_i\rvert}{c\sum_{i=1}^{n}\lvert O_i - \bar{O}\rvert}, when \\ \sum_{i=1}^{n}\lvert M_i - O_i\rvert \leq c\sum_{i=1}^{n}\lvert O_i - \bar{O}\rvert \\ \frac{c\sum_{i=1}^{n}\lvert O_i - \bar{O}\rvert}{\sum_{i=1}^{n}\lvert M_i - O_i\rvert} - 1, when \\ \sum_{i=1}^{n}\lvert M_i - O_i\rvert > c\sum_{i=1}^{n}\lvert O_i - \bar{O}\rvert \end{cases}$	(0, 1)	1	This method compares the sum of squared error to the potential error. It is designed to be better at handling differences in modelled and observed means and variances.	Squared differences may add bias to large data value events (Willmott, 1981).

2.9.4 Visual Performance Evaluation

2.9.4.1 Scatter Plot

Scatter Plot is a simple visual technique plots the prediction and observation pairs to showing how they relate to one another. Visually the model's over or under prediction can be easily detected and quantified. In openair package, the scatter plot function estimates the linear relationship between the observed and predicted values, the coefficients of determination, slopes and intercepts of the linear equations, and 95% confidence intervals of the fit.

2.9.4.2 Conditional quantile plot

The Conditional Quantile is a simple way of looking at a model performance against real world observations for a continuous measurement (Wilks, 2011). The Conditional Quantile plot function in an Openair divides both prediction values and observations into bin pairs of equal length and estimates the median, 25/75th and 10/90th percentile of each bin; the estimates are then plotted to show how predicted, and observed values agree with one another. A perfect model would lie on the blue line and with very narrow spread (Carslaw and Ropkins, 2012). This plot differs from Quantile – Quantile plot in that for a particular interval, it does not use the distribution of the observations and prediction separately, but it uses the corresponding values of the observations in predictions. This plot is particularly important in revealing how well the distribution of predictions tally with that of the observations especially at lower and upper part of the distribution.

2.9.4.3 Time variation plot

The time variation plots are useful tools in describing how pollutant concentrations vary with time (an hour of the day, the day of the week, weekly and monthly, etc.). In air pollution studies these plots could be used to reveal information about the likely sources of the emission. In openair package, the time variation function produces four-time scale variation

plots: a combined hour of the day and day of the week, a mean hour of the day, the day of the week and monthly variation plots. It can also show the 95% confidence interval in the mean values. The uncertainty intervals are calculated using bootstrap resampling which would provide a better estimate than assuming normality, especially for small data. There is also an option for normalising the data which is useful when plotting data with different units. The normalisation is done by dividing the concentration or any other variable by their mean values. These plots can also be useful in comparing the model predictions and the observations to observe how the model predictions agree with the observations on a time scale (Carslaw and Ropkins, 2012).

2.9.4.4 Tylor's diagram

The Tylor's diagram is a useful tool for comparing the performance of various models graphically. It shows three model performance metrics; correlation coefficient, standard deviation and centred RMSE. Taylor (2001) showed that it is possible to relate these statistics through the use of the law of cosines on a 2D graph. Taylor's diagram should not be used alone because the values of the metrics used do not depend on the mean bias, they only measure unsystematic errors, therefore a model might systematically over or under predict but still has the same scatter as the observation yielding a perfect match for standard deviation (Chang and Hanna, 2004).

2.9.4.5 Bivariate polar plot

The Bivariate polar plots of pollutants concentrations have been used to discriminate against various types and characteristics of emissions sources (Carslaw et al., 2006). These plots describe the joint variation of pollutant concentrations, wind speeds and wind direction on a continuous surface using polar coordinates (Carslaw and Beevers, 2013).

2.10 Summary

This chapter reviewed the existing literature on the health issues surrounding transport related air pollutants. Particular preference was given to particulate matter where its various health implications, sources, its atmospheric processes and various methods of measurements are discussed. The chapter also discussed the air quality modelling methods currently used in the most of the regulatory agencies and host of machine learning and statistical modelling methods capable of producing air quality models. A brief review of the ADMS-Roads model is also given. Lastly, various methods of model evaluation have been discussed. The literature reviewed on the health implications of traffic-related air pollution revealed that, although most air pollutants are injurious to human health, particulate matter is often associated with greater health risk. Also, there is need to consider various measurement options besides mass such as size distribution and particle number for estimating the health implications of particulate matter. Road traffic was found to be a significant contributor of particulate matter in an urban area.

The particulate measurement and modelling methods often used by the regulatory agencies reviewed show that despite the successful applications of the operational models in many air quality studies, they have shortcomings regarding their formulations. In most instance, the models approximate the particulates to gases or do not include the complex chemical and mechanical processes involving the particles. The simplifications of the atmospheric processes involved, overly dependent on the validation data and emission data makes the uncertainties in the estimated outputs of the models very high and as a result, affects any policy decision that might be taken with respect to the model's outputs. Also, most of the models do not include the use of particle number metric because of the inherent difficulty in accounting for the chemical and mechanical properties of the particles. Moreover, these

models require the knowledge of the atmospheric processes involving the dispersion of air pollution in addition to high computing capacity required when dealing with “Big Data”.

These observations lead to the conclusion that there is a need for more robust, efficient and cheaper modelling methods that will cater for the shortcomings of the current operational models. For these reasons, statistical and machine learning were considered as the modelling methods that could be used to provide air quality models for the prediction of the roadside particulate matter. These methods do not require knowledge or simplifications of the atmospheric processes involving air pollutants, and they are relatively cheaper and easy to operate. Also, most of the methods can be found in freely available software packages.

Particulate matter including PM_{10}, $PM_{2.5}$ and PNC are considered as the pollutants of choice for this study because of the shortcomings in the existing models regarding their estimation and the need for their accurate estimation due to their high health implications.

Although the use of machine learning and statistical methods in air quality studies is under active research, their adoption by the policy makers is rarely observed. Therefore, this research seeks to develop models for point and Spatio – temporal prediction of the traffic-related particulate matter using statistical and machine learning methods. The performance of the models will be evaluated and compared with the performance of an operational model (ADMS-Roads) in the predictions and evaluating the effectiveness of a hypothetical air quality management scenario.

The result to be obtained if favourable it will give more insight as to whether these methods could be used in the regulatory agencies or not and it will boost the confidence of the regulatory in using these methods for policy making.

Chapter 3

Methodology

3.1 Introduction

This chapter describes the processes followed in the execution of various modelling exercises involved in this research. The whole study is divided into three main categories including data collection and analysis, model development and model comparisons as shown in Figure 3.1. Therefore, this chapter is divided according to these categories. Section 3.2 describes the procedure followed in data collection. Section 3.3 describes the procedure for the data processing and analysis. In Section 3.4, the procedure adopted in Machine learning and statistical modelling was described. The step by step procedure for feature selection exercise carried out is described in Section 3.5. ADMS-Roads modelling processes are described in Sections 3.6. The chapter concludes with its summary.

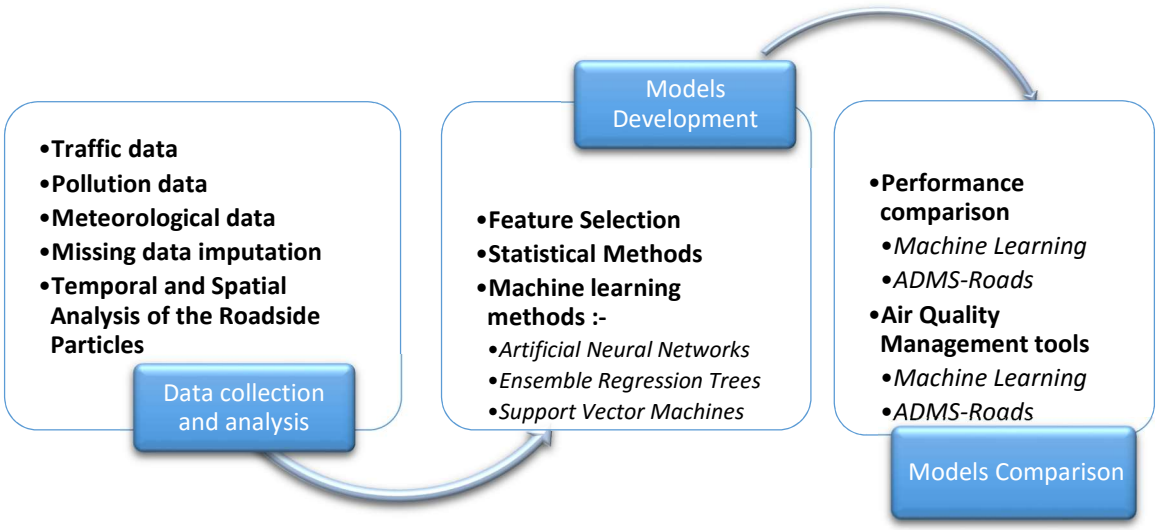


Figure 3.1 Methodology Flowchart

3.2 Data Collection

This research requires the use of historical data on traffic, pollutants and the meteorological data to achieve its aim. The various data variables are displayed in Figure 3.2 below. The data shown is required at each monitoring unit. However, meteorological data was collected from Heathrow airport weather station which is believed to be representative of the meteorological condition of London. Also, the traffic volume was not available for all the stations, therefore, estimated traffic data was used where necessary.

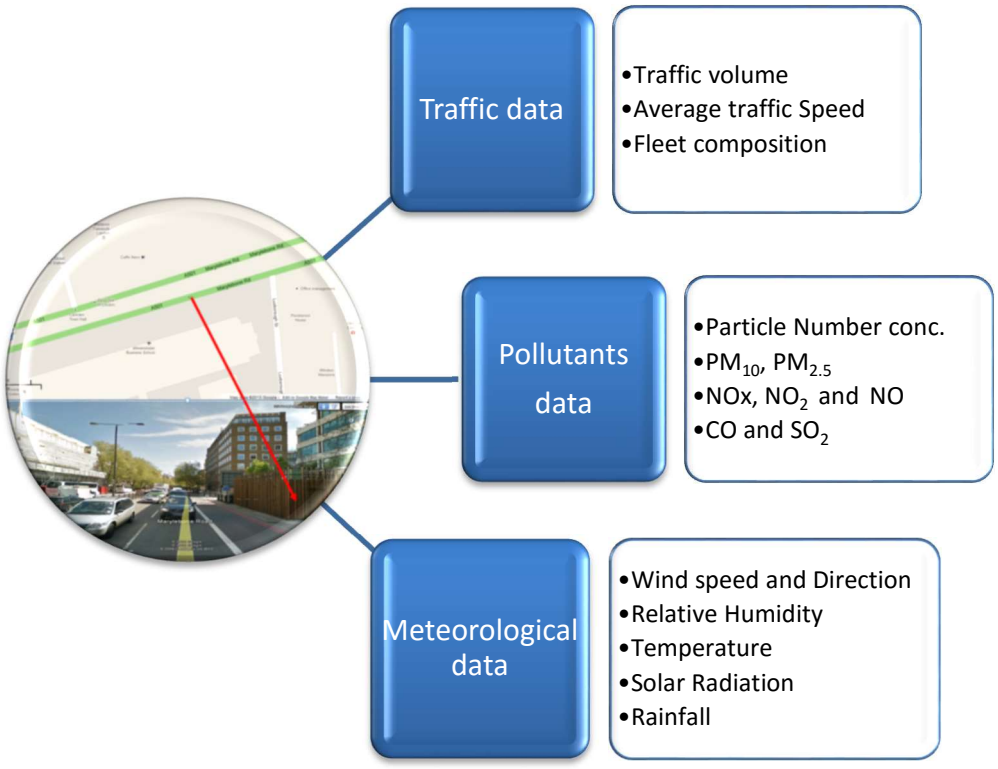


Figure 3.2 Summary of the data requirement at each monitoring unit.

3.2.1 Selection of Air Quality Monitoring Sites

There are many air quality monitoring sites in London maintained by different organisations notably Department for Environment, Food and Rural Affairs (DEFRA), London boroughs, British Airport Authority (BAA) and Transport for London (TfL). The sites are categorised into kerb sites, roadside, industrial, suburban background, urban background, and rural background sites. They are individually or collectively serving some purposes which include determination of the level of compliance with the air quality objectives and limits, providing information to academia for research and consultancies. It also includes providing information to developers for the preparation of air quality and environmental impact assessments. The sites also play a vital role in providing data for air quality models development and evaluation. Moreover, the sites also provide information on trends in pollutant concentrations for measuring the effectiveness of the air quality control strategies (Moorcroft and Marner, 2011).

The London sites used for this study were selected from the sites categorised as strategic by Moorcroft and Marner (2011). The strategic sites include the sites that are being used as Average Exposure Indicator Reference Sites for $PM_{2.5}$ and Low Emission Zone (LEZ) evaluation sites. The strategic sites are also part of the Automatic Urban and Rural Network (AURN). An additional criterion used in selecting the sites for this study is the data availability and the type of the site. First, the roadside and kerb sites with the available data were chosen and either an urban background or suburban sites located upwind of the roadside or kerb sites were used as their background sites. The sites at the Instrumented Junction in Leeds were selected based on their location and data availability.

3.2.2 Pollutants Data

The data itemised in Figure 3.2 were collected through the London air quality network (London Air, 2013), the UK Air Quality Archive (UK-AIR, 2013) of the Department for Environment, Food and Rural Affairs (DEFRA) and the Institute for Transport Studies at the University of Leeds for the sites in London and Leeds respectively. The Instrumented Junction is located in Headingley, Leeds. It is equipped with three air quality monitoring stations (ENV1, ENV2 and ENV3), and inductive loops where air quality and traffic variables are simultaneously being measured. Also, there are instruments for measuring wind and street inflows.

A detailed description of the various monitoring sites where the data was collected is given in Chapter 4. $PM_{2.5}$ and PM_{10} concentrations data were collected using Tapered Element Oscillating Microbalances (TEOM) Model 1400AB with different sampling heads design (Aurelie and Harrison, 2005). The TEOM consists of a filter, tapered hollow glass tube and PM_{10} impactor inlet for measuring PM_{10} mass while the sharp cut cyclone is attached to the TEOM for measuring $PM_{2.5}$. The filter dynamics measurement system (FDMS) have been installed at some of the monitoring stations to minimise the problems of loss associated with the TEOMs. At some of the monitoring stations ß-attenuation analysers were used for the PM measurements. The PNC data at MY1 and BL0 London monitoring sites was collected using the Scanning Mobility Particle Sizer system (SMPS). The SMPS system consists of Electrostatic Classifier (EC) model 3071A and a condensation particle counter (CPC) model 3022A for measuring the particle sizes and the particle concentration respectively (Aurelie and Harrison, 2005). Data from these sites are being measured according to EU protocols and are undergoing the quality assurance and quality controls according to Urban and Rural Network (AURN) and London Air Quality Network standards. For the monitoring sites at

the Instrumented Junction in Leeds, the Pollutant data were collected using the instruments listed in Table 3.1.

Table 3.1 Air quality monitoring instruments used at the Instrumented Junction Leeds

Instruments	**Manufacturer**	**Model**
WCPC particle counter	TSI	WCPC 3785
Butanol CPC particle counter	TSI	3775
NO/NO2/NOx analyser	Teledyne	200E
O3 analyser	Teledyne	400E
Data logger for NOx, O3 & WCPC	OPSIS	DL256
GSM Modem	Siemens	TS35i
Sonic Anemometer (on pole)	Gill	Solent R3
Rugged Versatility Data logger for Anemometer	Campbell Scientific	GPS & CR1000
Global positioning satellite (GPS)		

3.2.3 Traffic Data

The traffic volume data was collected from two sources. (1) Continuous traffic data monitored alongside the monitoring sites in London and the Instrumented Junction sites in Leeds. (2) Manual count data collected by the Department for Transport every year at some traffic count points on road links across the UK. The continuous traffic data was collected using induction loops buried on each lane with an estimated accuracy of 99% for counting and classification. The manual counts were conducted between 7:00 am and 7:00 pm for each link for a maximum of one day in a year. The traffic speed collected alongside the

continuous traffic count was also collected. The average speed of each road link under consideration was also obtained from the LAEI archives (LAEI, 2014). Traffic volume data was disaggregated into eight traffic categories (i.e. Petrol car, Diesel car, Taxi, LGV, Rigid, Artic, Bus and coach and Motorcycle) based on the UK traffic composition projections. Moreover, their corresponding hourly emission rates were estimated using LAQM emission factor toolkit version 6.0.1. The PNC emissions rates for Light duty vehicles (LDV) and Heavy-duty Vehicles (HDV) were estimated using emission factors derived by Jones and Harrison (2006) shown in Table 3.2. The particle emission factors used in this study were derived from the field studies conducted by Jones and Harrison (2006) using data collected at Marylebone Road and London Bloomsbury air quality monitoring sites. These sites are the main source of air quality data for this research which makes the emission factors more suitable to use. The estimated emission factors were compared with those estimated using laboratory test-bed measurements and other field studies and were found to be within only one to two standard deviations of the equivalent estimated emissions using the laboratory data (Beddows and Harrison, 2008). The emission factors estimated using field studies could be taken as more representative of the reality since the data was not taken under control conditions.

Table 3.2 PNC Emission Factors

Pollutants	**Units**	**Heavy vehicles**		**Light vehicles**	
		Emission factor	Standard error	Emission factor	Standard error
PNC_{11-30}	Number v^{-1} km^{-1}	2.14×10^{14}	4.14×10^{13}	2.30×10^{13}	7.38×10^{12}
PNC_{30-100}	Number v^{-1} km^{-1}	3.19×10^{14}	3.92×10^{13}	2.84×10^{13}	6.99×10^{12}
$PNC_{>100}$	Number v^{-1} km^{-1}	1.03×10^{14}	1.22×10^{13}	7.05×10^{12}	2.18×10^{12}
$PNC_{all\ sizes}$	Number v^{-1} km^{-1}	6.36×10^{14}		5.84×10^{13}	

(Jones and Harrison, 2006)

3.2.4 Meteorological Data

The meteorological data for London sites was collected from London Heathrow Airport Meteorological station through BADC data services (MIDAS Land Surface, 2013). However, those used in conjunction with the Leeds sites were collected from a 35m meteorological mast located 4km to the South of the Instrumented Junction being operated by the Leeds City Council. The meteorological data include wind speed, wind direction, solar radiation, relative humidity and ambient temperature. Others include rainfall and barometric pressure.

3.3 Data Analysis

The predictor variables consisting of traffic, pollutants and meteorological variables prepared for the models are shown in Table 3.3. These variables are expected to be collected for each roadside or kerbside monitoring station.

Table 3.3. Models predictor variables

Serial Number	Predictor Variables	London sites	Instrumented Junction Leeds	Units
1	Date	yes	yes	day/month/year hour: minutes
2	Carbon monoxide(CO)	yes	No	$\mu g/m^3$
3	Nitric oxide (NO)	yes	yes	$\mu g/m^3$
4	Sulphur dioxide(SO_2)	yes	No	$\mu g/m^3$
5	Nitrogen dioxide(NO_2)	yes	yes	$\mu g/m^3$
6	Wind direction	yes	yes	(0N)
7	Temperature	yes	yes	(0C)
8	Solar Radiation	yes	yes	W/m^m
9	Wind speed	yes	yes	m/s
10	8 Traffic emission rates	yes	yes	g/km
12	Barometric Pressure	yes	No	mBar
13	Relative Humidity	yes	yes	%
14	Average speed	yes	Yes	Km/hr
15	Rainfall	yes	No	mm

The response variables are PM_{10}, $PM_{2.5}$ and PNC concentrations for the models trained for London sites and only PNC concentration for the sites in Leeds (Instrumented Junction). The descriptive statistics of the variables which include mean, median, 1st and 3rd Quartiles, maximum and minimum values, and the number of missing data were obtained to observe the quality and distribution of the data. The wind rose plots, polar plots and time variation plots were also plotted to examine the prevailing wind direction and wind speed, an association of the target variable with the air pollution sources in the vicinity of the air quality monitoring sites. Correlation analysis was carried out to determine the correlation between the variables in the data. The long-term trend analysis was carried out to examine the trend in the particle concentrations over the study period.

3.4 Missing Data Imputation

In this study, the R software package, Multiple Imputations by Chained Equation (MICE) was used for the imputation of the missing data, and the effect of the imputation on the data and the accuracy of the models was investigated. The MICE software implements a semi-parametric approach to multiple imputations called Fully Conditional Specification (FCS). In FCS, an imputation model is specified for each variable by the use of conditional densities (Van Buuren, 2007, Buuren and Groothuis-Oudshoorn, 2011). The approach does not take into account nonlinear relations in the incomplete dataset and therefore produces bias estimate of the missing data. However, implementing recursive partitioning technique such as random forest and classification and regression trees (CART) in the framework of MICE bridges this gap and produces a better estimate of missing values (Doove et al., 2014). In this research, random forest method implemented in the mice framework was used for the missing value imputation because of its ability to handle nonlinear relationship which exists between the various variables prepared for the particulate matter prediction.

3.5 Model Development

The study was carried out using Matrix Laboratory (MATLAB) software, a freely available Open Source Statistical Software (i.e. R software) and an air quality management system for traffic domestic and industrial pollution (ADMS-Roads). Table 3.4 summarises the toolboxes and software packages used.

Table 3.4 Summary of software packages

Serial Number	Software	Toolbox/Package	Modelling method	Reference
1	R statistical software	The classification and regression training (caret)	Statistical, neural networks, BRT and SVM models training	(Kuhn, 2008)
2	R statistical software	Lasso and elastic-net regularised generalised linear models (glmnet)	Lasso and Elastic-net	(Friedman et al., 2009)
3	R statistical software	H2o package, R scripting functionality for H2O, the open source math engine for big data that computes parallel distributed machine learning algorithms	Neural networks, BRT and Random forests	(Fu et al., 2014)
4	R statistical software	Data Analysis Tools for the Air Quality Community (openair)	Air quality data analysis and model evaluation	(Carslaw, 2012)
5	R statistical software	Generalised boosted regression models (gbm)	Boosted regression trees (BRT)	(Ridgeway et al., 2013)
6	R statistical software	R software interface to Libsvm (E1071)	Support Vector Machines (SVM)	(Dimitriadou et al., 2008)
	R statistical software	The split-apply-combine paradigm for R (plyr)	Tools for Splitting, Applying and Combining Data Description	(Wickham and Wickham, 2013)
7	R statistical software	Breiman and Cutler's random forests for classification and regression (randomForest)	Random forests	(Breiman, 2006)
8	ADMS-Roads	Air quality management system for traffic, domestic and industrial pollution	Operational air quality modelling	(McHugh et al., 1997)

3.6 Feature Selection

Variable selection or feature selection was carried out to select among the available data the minimum combination of predictor variables that will be used to train the models for the prediction of roadside particle concentrations with an acceptable level of accuracy. In this work, the two wrapper methods, Genetic algorithms (GA) and Simulated Annealing (SA) each, combined with the Random Forests (RF) are used for the feature selection (see Section 2.8 for details of the methods).

3.7 Implementation of Hybrid Feature Selection Processes

The implementation of the Hybrid feature selection using GA, SA and RF by Kuhn (2012) in R software (R Development Core Team, 2015) is adopted in this work. The algorithm

carries out repeated searches in the feature space within the resampling iterations. Initially, the resampling method is specified in the control function, and then the entire genetic algorithm or simulated annealing process is implemented separately on each sample. Here 10 – fold cross-validation, repeated five times, was adopted as the resampling method for the external performance of the selection process. Therefore, for the first fold, the search is conducted on the initial nine – tenths of the training data while the external performance is estimated with the remaining tenth. The optimal number of generations in the GA or number of iterations for the SA is determined using the external performance since it does not take part in the search process. However, during the search, there is a need for the internal performance to guide the search, and this is determined using another resampling within the selected data. This procedure has the potential of overfitting the estimates; that is why the external performance is used for the selection of the final predictors.

The effects of the hybrid feature selection methods in improving the efficiency of the five popular linear regression methods are investigated. The methods include Multiple Linear Regression (MLR), Partial Least Square Regression (PLSR), Principal Component Regression (PCR), Stepwise Regression and Lasso/Elastic-net Regressions.

3.8 Statistical and Machine Learning Modelling Process

The process began with the selection of the monitoring sites that satisfy the aim of the research which is to model road traffic-related particulate matter and also the sites with the available data. The data collected were then processed to identify the nature of the data and imputed the missing data. The next step was the determination of the most relevant predictor variables among the various predictor variables collected for the modelling. This step was necessary because it reduces the complexity of the intended model and also reduced the number of input data required which will eventually reduce the cost of providing the model.

However, if the modelling method selected have an inbuilt feature selection method, then this step was carried out at the training stage. The data with the appropriate number of inputs was then divided into the training and testing dataset through the use of K-fold cross-validation or by using the date. The next step was the training and testing the model using their respective datasets. The model was then evaluated using several model evaluation metrics and visualisation tools after obtaining satisfactory training and testing results.

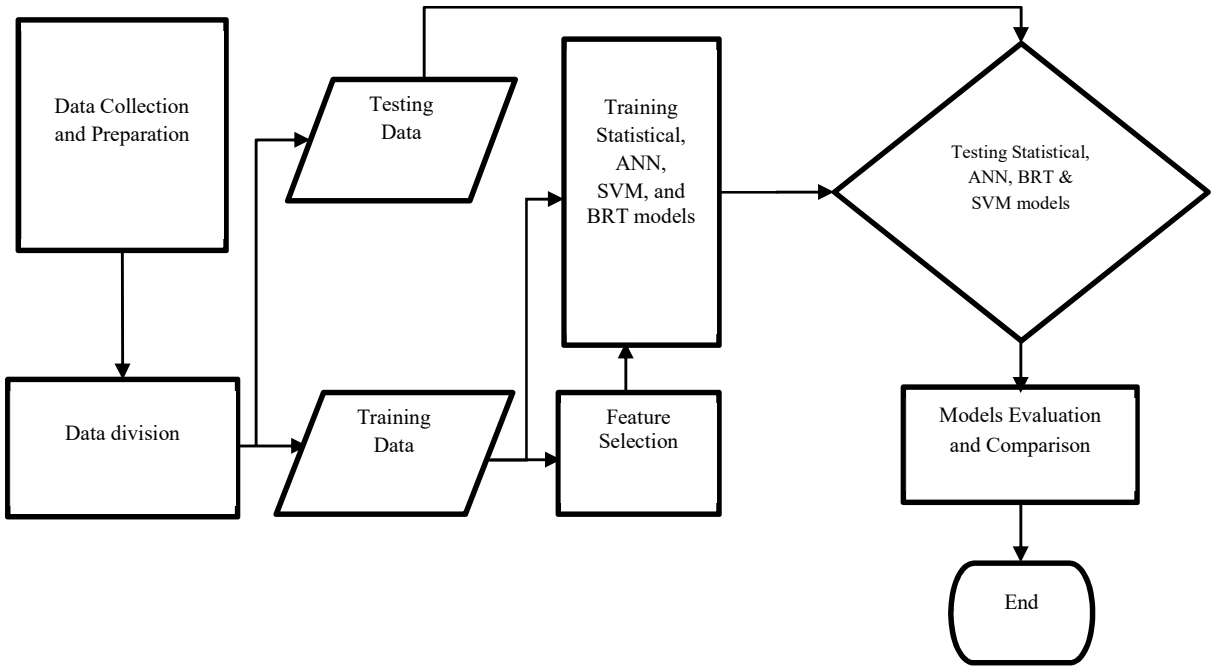


Figure 3.3 Flow chart for the Machine Learning modelling process

Figure 3.3 shows the flow chart of the modelling tasks carried out in the training. The data prepared for the modelling was divided into eighty percent training and twenty percent testing subset to train and test the prediction models respectively.

The training data subset was first used to select the most important predictor variables for models using the RF-GA method discussed in Chapter 6. After that, the data for the selected

variables was extracted from the training data set to form a separate training set with the selected variables giving rise to two training datasets.

The models were then trained with each data set using repeated k –fold cross – validation. The k – fold cross validation is a resampling technique designed to partitioned a training data into subsets (k – folds) such that a sample is held out while a model is trained with the remaining samples and then tested on the held out sample. The performance of the model on the held out sample is estimated using suitable performance metrics e.g. RMSE.

The procedure is repeated until all the samples have been used as held out samples. The performance of the model is then estimated as the average of the performance of the models on the held out samples. A repeated version of the k-fold cross-validation is adopted where the k = 10 was used and repeated five times.

The model parameters that require tuning were determined using grid search methods where the repeated cross-validation is applied to train the models with the specified range of the model parameters and the parameters that produced a model with the best performance are selected for the training of the final model. When each model is sufficiently trained, it is then tested using the test data set and subsequently its performance was evaluated using various model performance evaluation functions explained in Chapter 2. The final task is the comparison of the performances of the models for each target pollutant.

3.8.1 Neural Network Model Training Steps

1. Create a network: this involved the selection of the type of neural network, its architecture, activation functions and learning algorithms.

2. Configure network: the network created in (1) above was configured based on the input and target output data available

3. Initialise weights and biases: The network weights were set to small random values to prepare the network for training. This is similar to selecting a random point on error surface (Gardner and Dorling, 2000). Steps (2) and (3) are being carried out automatically when the configuration function in ANN package is initiated.

4. Train the network: This step involved the Normalizing the data, dividing it into training, validation and testing data subsets and using appropriate learning algorithm to estimate the network weights using the training data set and subsequently estimating the model performance

5. Validate the network

In this step, the performance of the network was determined using mean squared error (MSE) of the network output. Regression plots for network output, and the target output for all the data divisions will also be obtained. These plots were examined and in the case of any serious concern about the accuracy of the network prediction, the neural network was then be trained again and in some cases, steps (1) through (4) were repeated until the desired results were obtained.

6. Test the network: this step deals with using the test data to predict the desired output. The output was then compared with the actual response variable using various statistical performance statistics and graphical methods.

7. Use the network: once the network is validated, the network structure was saved and was used to predict any similar response variables with entirely different but related data.

3.5.1 ANN methods used in this study

In this research, we used five different ANN formulations including Multi-Layer Perceptron with Principal Component Analysis (PCA-MLP), Multi-Layer Perceptron with Model Averaging (AVG-MLP), Bayesian Regularised Neural Network (BRNN), Extreme

Learning Machine (ELM) and Deep Learning (DL). See Section 2.7.1 for the details of the methods.

3.8.2 Boosted Regression Trees Model Development Steps

1. Selection of Tree complexity, i.e. number of trees and learning rate
2. Computing the average response $\bar{y}$, and use it as initial predicted value for each sample
3. For k = 1, 2... K the following steps are executed
4. Estimate the residuals (i.e. the difference between observed and predicted values)
5. Fit a regression tree of a particular depth, D, using the residuals estimated in (4) as response variables
6. Use the regression tree fitted in (5) to predict each sample
7. Use appropriate shrinkage techniques, shrink the current prediction and add it to the predicted value in the previous iteration.
8. End

(Kuhn and Johnson, 2013)

The optimum BRT tuning parameters in this work were determined using the *train* function of the *caret* package. The function uses cross-validation to determine the optimum combination of the tuning parameters. For each pollutant, five different learning rates 0.001, 0.01, 0.05, 0.1, and 0.5, the number of trees from 1 to 10,000, tree complexities from 1 to 10 and a fixed bag fraction of 0.5 were tested. The model with the combination of tuning parameters that give the lowest RMSE value was taken to be the final model in each case. The final models were then tested and evaluated with the same testing data used for testing ANN models.

3.8.3 Model Tuning using *train* Function of the Caret Package.

The *caret* package of R software has a *train* function that is dedicated to streamlining the model building and evaluation process. Here the train function was used to tune all the statistical models and support vector machine models using the following steps.

1. Select the modelling method to be used. In this step the model parameter values would be specified according to the modelling technique e.g. for SVM, there were two model parameters to be optimised, the cost parameter (C) and the smoothing parameter (σ). The *train* function would then fit the models over a certain range of the parameters depending on the modelling technique.
2. The *train* function could then use the resampling method specified (e.g. bootstrap) to hold back a particular sample and fit the model on the remainder of the samples. The holdout samples would then be used to evaluate the performance of the trained models. This procedure is repeated until all the samples have been used for fitting the model and also as holdout samples. The optimal model parameters would then be determined based on the performance of the models built with them.

3. The final models were then fitted to all the training data using the optimal parameters (Kuhn, 2008).

3.9 ADMS-Roads Modelling Process

The modelling process in the ADMS-Roads began with the provision of different types of model inputs that were supplied to the model in six modules which included: Setup, Source, Meteorology, Background, Grid and Output modules as shown in Figure 3.4 below.

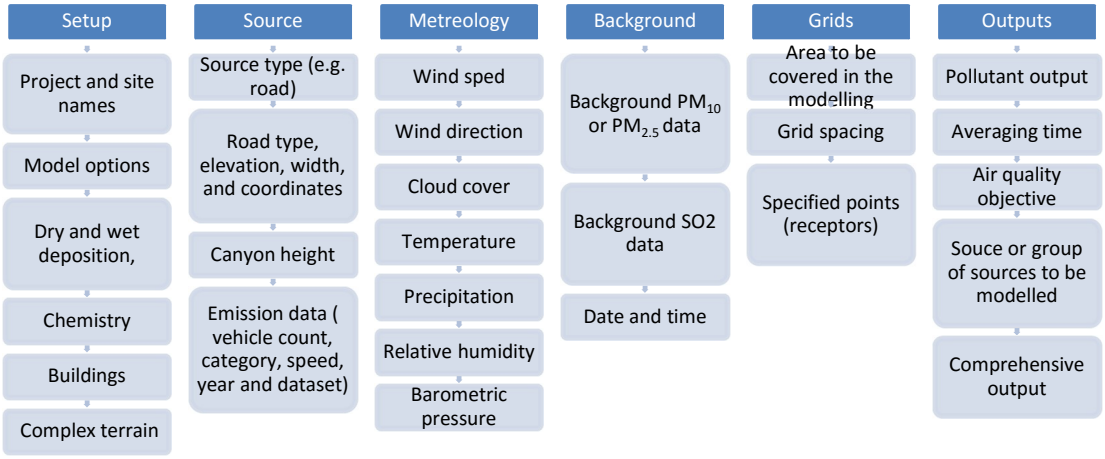


Figure 3.4. ADMS-Roads model modules

The setup module requires the user to provide the names of the project, the sites involved, and also to select the model options (i.e. deposition, odours, chemistry, buildings and complex terrain). The chemistry option if selected allows for the calculation of secondary particles and hence the SO_2 data needs to be provided in the background data file for this purpose. The second step was to provide information about the type of emission source (s) to be modelled. For example, if road source is selected, the road type, elevation, width and coordinates are provided. And if the road is located in a street canyon, the properties of the canyon i.e. height and width should be provided. Emission related data such as vehicle categories, counts, average speed, year and type of dataset to be used were also provided. The Third step was the meteorology input module where meteorology data, as shown in Figure 5, were provided in a specified format. The information on the meteorology and the dispersion sites was provided through the supply of the values of the surface roughness, Monin-Obukhov length, surface albedo and the height at which wind data was collected. The fourth step was to provide the background data of the pollutant to be modelled i.e. PM_{10} or $PM_{2.5}$ and the SO_2 was also provided to account for the information required for the chemistry option selected. The fifth step was to provide information about the grids of the dispersion area which include the grid spacing and the coordinates of the minimum and

maximum points bounding the dispersion area and the specified points of interest (e.g. monitoring sites and other receptors). The last step was the output module where the properties of the output(s) were specified. It requires information such as the name of pollutant(s), averaging time, the air quality objective, and the number of sources or group of sources to be modelled. The option for comprehensive output that allows for summarising the results after running the model was also selected. After providing the inputs for all the modules, the model was then run by pressing the run button. When the model run finished, it produced several files containing the results and various processed variables that were then extracted for model evaluation.

3.9.1 Emission inventory

The Emission Inventory is required in the second stage of the ADMS-Roads modelling where the information about the sources of emission is required. The inventory to be used with the ADMS-Roads is carried out in two ways: (a) entering the source data directly to the source module as described above or preparing the inventory separately using Microsoft Excel and Microsoft access database applications and then import the inventory into the ADMS-Roads model. The second method was adopted because it is more flexible when dealing with the large pollution sources. Therefore, the geometry of the sources was imported into the inventory file from either the map utilities of the ADMS-Roads or ArcGIS software.

The data required for the emission inventory include:

1. Road

a. Source name

b. Canyon height

c. Canyon width

d. Emission Dataset

e. Road type

b. Logical (Yes/No) to indicate using the traffic data or not to estimate emission rates

2. Source

a. Source name

b. Geometry type (e.g. Road)

3. Source geometry

a. Source name

b. Vertex number

c. X and Y coordinates

4. Traffic flow

a. Source name

b. Vehicle Category

c. Average speed

d. Vehicle count

5. Pollutants

a. Pollutant name

b. Other default options

6. Emission

a. Source name

b. Pollutant name

c. Emission rates

3.9.2 Euro4/VI Air Quality Management Scenario

An air quality management scenario is required to test the application of the machine learning models developed in this study in predicting the impact of traffic-related air quality management options. A hypothetical scenario which is called here Euro4/VI scenario was conceptualised to test the use of the machine learning models in real life application and compare their performance with the performance of an operational air quality model (i.e. ADMS-Roads). The scenario suggests that only petrol vehicles and diesel vehicles meeting EuroIV/4 and EuroVI/6 design specifications respectively would be allowed to enter the study area. After the scenario has been developed, the machine learning and the ADMS-Roads models were used to test the scenario in 2011 and 2015 for PM_{10}, and 2012 and 2015 for $PM_{2.5}$. The scenario was first implemented in London Westminster city where there were only two sites with the sufficient PM_{10} data and one site with sufficient $PM_{2.5}$ data. The Westminster city was chosen because is the location of Marylebone road monitoring site, where all the particle metrics and the traffic variables are being measured. The site is also one of the AURN super sites and it is located in an area of major air quality concern in London.

The emission inventory used in the ADMS modelling include all the major roads within the city of Westminster. However, for the machine learning models, traffic data obtained from Marylebone road was used as the representative of the traffic in the area. The remaining sites were later included after obtaining satisfactory results from the three sites, but only the roads closed to the monitoring units were considered in both the ADMS-Roads and the machine learning models.

3.10 Summary

This chapter described the methods and procedures adopted while carrying out the research. The data collection methods and the various instruments used for the collection are briefly described. The step by step processes of statistical, machine learning and ADMS-Roads modelling processes are described.

Chapter 4

Air Quality Monitoring Sites and the Description of Data

4.1 Introduction

The main objective of this research is to carry out an in-depth study on the use of machine learning and statistical methods in modelling roadside particle concentrations consisting of PM_{10}, $PM_{2.5}$ and PNC as discussed in Chapter 1. This task requires an observation data from continuous monitoring stations where the air pollutants, traffic and meteorological data are collected simultaneously. Therefore, based on this requirement Twenty-one air quality monitoring sites in London and four in Leeds were selected for the study. See Section 3.2.1 for details of the sites selection criteria.

The data collected from the sites in London was for the years 2000 to 2012 and those collected in Leeds covered the period between March 2009 and March 2010. However, there are some sites where the data is incomplete and only less than or equal to 10% missing data was tolerated. The Multiple Imputations by Chained Equation (MICE) imputation algorithm developed by Buuren and Groothuis-Oudshoorn (2011) was used to impute the missing data at those sites to ensure completeness of the data.

The Meteorological data was collected from Heathrow meteorological station and Leeds City Council Meteorological Station for the sites in London and Leeds respectively. Traffic data were gathered from a range of sources including UK-AIR (2013), Transport for London (TfL), Department for Transport (DfT) and the London Atmospheric Emission Inventory (LAEI) database.

This chapter presents the description of the air quality monitoring sites and the measurement instruments used in this study in section 4.2. Brief description and descriptive statistics of

the traffic, meteorological and pollutant variables are given in section 4.3 – 4.5. Also the analysis of the long-term trends in the particle pollutants is provided in section 4.5. The correlations between variables in the air quality data are described in section 4.6. Section 4.7 presents the validation of the missing data imputation. The summary of the chapter is given in section 4.8.

4.2 Air Quality Monitoring Sites

4.2.1 London Air Quality Monitoring Sites

The spatial distribution of the air quality monitoring sites in London is shown in Figure 4.1, and their properties are summarised in Table 4.1 below.

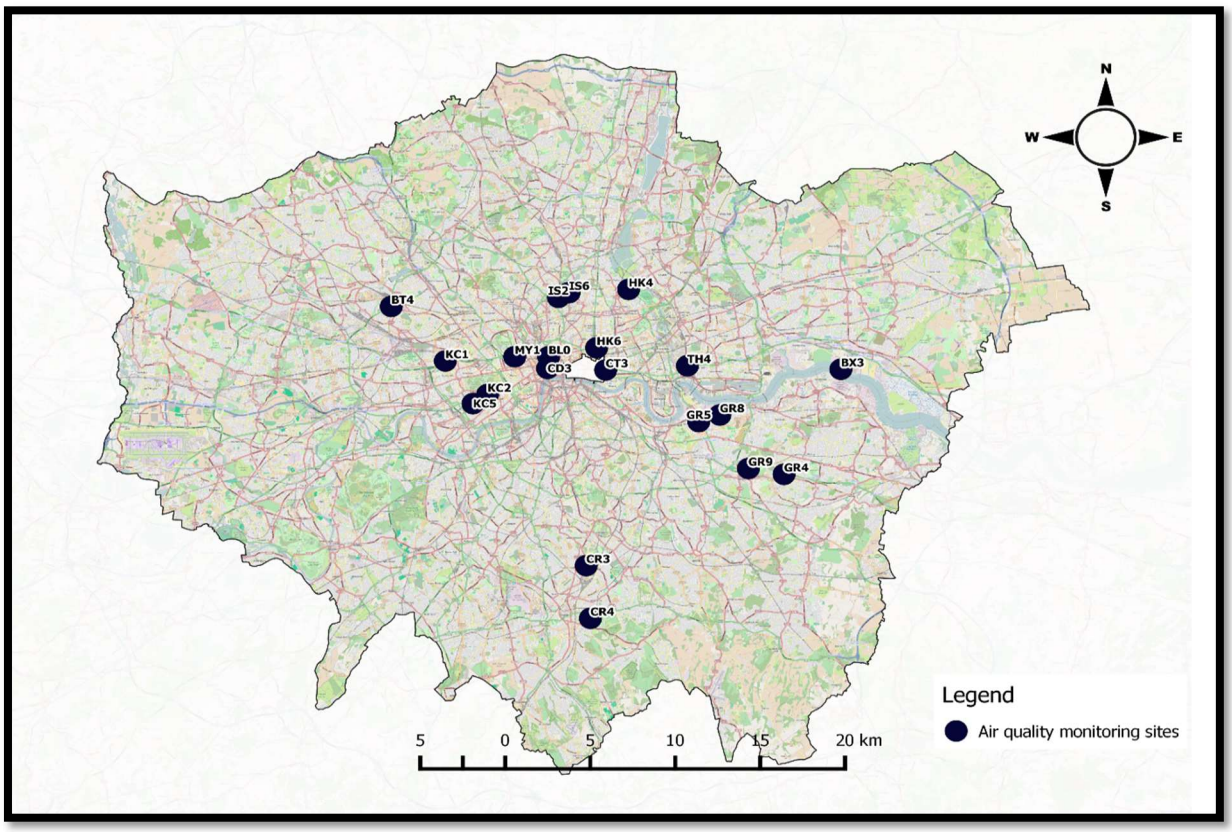


Figure 4.1 Map showing the location of the London monitoring sites (OpenStreetMap, 2015)

Table 4.1 Properties of the London monitoring sites

Site code	Easting	Northing	Site name	Site type	Distance to the road (m)	Traffic volume (veh/hr)	Average PM_{10} ($\mu g/m^3$)	Average $PM_{2.5}$ ($\mu g/m^3$)	PM_{10} (%) available	$PM_{2.5}$(%) available	Sampling Height (m)
BL0	530123	182014	Camden - Bloomsbury	Urban background			21.76	16.59	92.1	93.4	
BT4	520866	185169	Brent - Ikea	Roadside	Not available	4389	43.25	9.20	90.4	91.4	Not available
BX3	547323	181231	Bexley - Thamesmead	Suburban				9.450	-	99.2	
CD3	530057	181285	Camden - Shaftesbury Avenue	Roadside	3	1700	34.00		91.4	-	
CR3	532336	168934	Croydon - Thornton Heath	Suburban			21.13		91.0	-	
CR4	532583	165636	Croydon - George Street	Roadside	8	2500	25.00		95.0	-	4
CT3	533480	181186	City of London - Sir John Cass School	Background			27.51		91.5	-	
GR4	543978	174655	Greenwich - Eltham	Suburban			21.91	15.88	99.6	-	
GR5	538960	177954	Greenwich - Trafalgar Road	Roadside	5	1500	23.37		99.6	98.2	3
GR8	540200	178367	Greenwich - Woolwich Flyover	Roadside	3	7000	40.00	16.90	-	91.1	3

Table 4.2 *continued*

Site code	Easting	Northing	Site name	Site type	Distance to the road	Traffic volume	Average PM10	Average PM2.5	PM10 (%)	$PM_{2.5}$ (%)	Sampling Height (m)
GR9	541879	175016	Greenwich - Westhorne Avenue	Roadside	12	2700	22.00	16.74	94.1	95.5	3
HK6	532947	182575	Hackney - Old Street	Roadside	6	2500	31.83	16.62	98.6	-	3
IS2	530698	185735	Islington - Holloway Road	Roadside	3	2000	30.73		97.5	-	3
IS6	531325	186032	Islington - Arsenal	Urban background			22.40		96.7	92.4	
KC1	524046	181750	Kensington and Chelsea - North Ken	Urban background			21.11	14.68	81.4	-	
KC2	526527	179646	Kensington and Chelsea - Cromwell Road	Roadside	4	2800	33.71	15.39	98.9	-	2
KC5	525671	179080	Kensington and Chelsea-Earls Court Rd	Kerbside	Not available	1600	35.83		97.5	82.4	Not available
MY1	528125	182016	Westminster - Marylebone Road	Kerbside	1.5	3327	43.25	21.68	-	89.3	
TH4	538290	181452	Tower Hamlets - Blackwall	Roadside	4	6000	31.68	18.00	-	96.0	4

The London monitoring sites consist of twelve roadsides and seven background sites where PM_{10} and $PM_{2.5}$ data are collected. Among the sites considered in London, only MY1 and BL0 sites had sufficient PNC data available. Therefore, PNC data was only collected from those sites. Among the roadside sites, HK6, IS2, KC5 and MY1 are located in street canyons. GR5, GR8, KC2, CR4 and CD3 are located at junctions, while BT4, GR9 and TH4 are situated in an open area. A brief description of the roadside and kerb sites is given below. The BX3, BL0, CR3, CT3, GR4, HA1, IS6 and KC1 are the background monitoring sites selected for the roadside and kerb sites. They are mostly located in areas where there is less influence of local pollution sources. Other pollutants being measured at these sites include NOx, NO_2, NO, SO_2, CO, and O_3. Weather sensors are also available at some sites where local meteorology is being measured. The data is openly accessible through the London Air Archives (London Air, 2013) and UK Air Quality Archive (UK-AIR, 2013).

4.2.2 Air Quality Monitoring Sites at Instrumented Junction Leeds

The Instrumented Junction is a signalised junction connecting Otley Road (A660) and North Lane (B6157) located in the busy Headingley area of Leeds city 3km northeast of the city centre. The Junction often exceeds its planned capacity with average daily traffic of about 17,000 vehicles. The southern side of the junction is demarcated by a broken street canyon of aspect ratio (H/W ≈0.8). Moreover, to the north, a continuous 20m height shopping complex demarcates the eastern side of the Otley Road. Irregularly spaced buildings demarcate the western side of the Otley Road.

The junction is equipped with three Kerbside air quality monitoring stations (ENV1, ENV2 and ENV3) approximately spaced between 40 to 45m as shown in Figure 4.2. There is also a background site ENV4, located at approximately 180m from the junction. The ENV1 is situated on the eastern side of Otley Road and approximately 50m north of the junction and

ENV2 is located at the heart of the junction on the western side of Otley Road. ENV3 is located on the southern side of the North Lane approximately 25m from the junction housed by a deep street canyon with an aspect ratio of ≈1.3. Adjacent to each air quality monitoring stations is the traffic monitoring unit for measuring traffic flow, fleet composition and spot speeds.

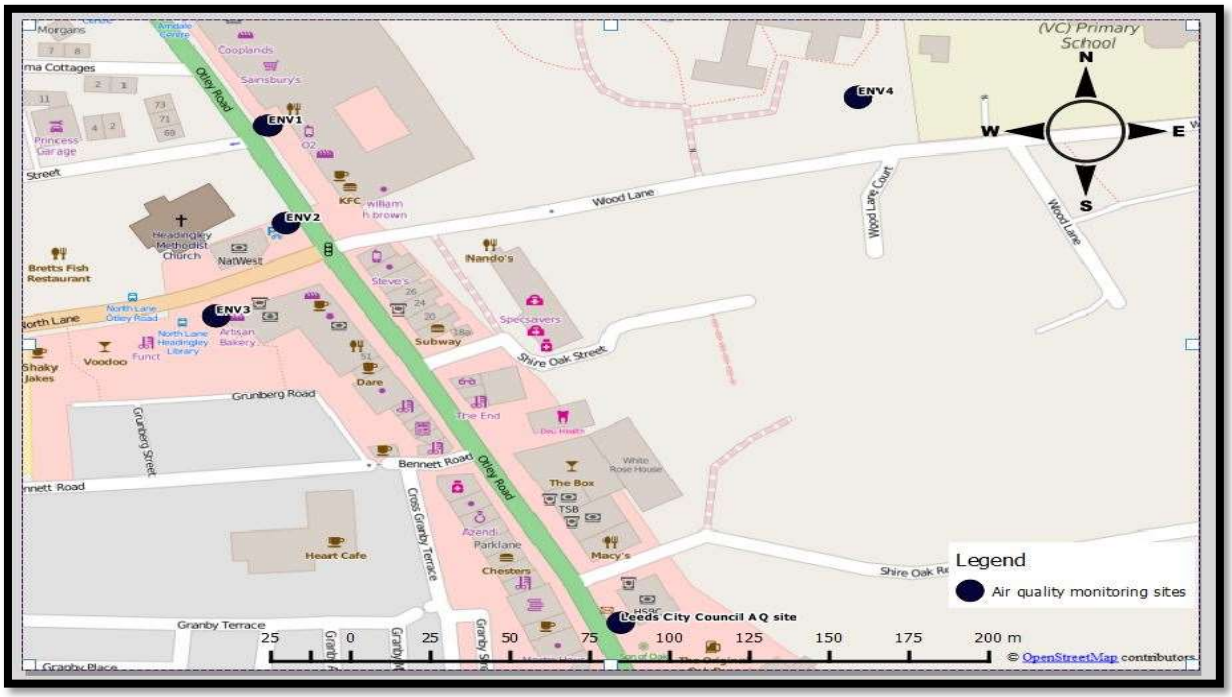


Figure 4.2 Map showing the location of the monitoring sites in Leeds (OpenStreetMap, 2015)

The prevailing wind data for the sites were obtained from a 35m meteorological mast located 4km to the South of the junction operated by Leeds City Council. Air pollution data monitored at the sites include NOx, NO_2, NO and PNC. The Meteorological data collected at the sites include wind speeds, wind direction, in-street flows, solar radiation, relative humidity and temperature. The data from these sites was collected for a period between March 2009 and March 2010 and was made available by the Institute for Transport Studies at the University of Leeds.

Data from the sites above (Sections 4.21 and 4.22) were only considered when training the machine learning models for spatiotemporal prediction and testing air quality management scenario (see Chapter 8). However, for comparison between the prediction accuracy of the models, only two sites (MY1 and EV1) were used, one each from London and Leeds.

4.3 Description of the Traffic Data

The Continuous traffic data in London available to us was only for MY1, HK6, BT4 and KC2 sites, therefore, at the remaining sites where there is no data an estimate of the traffic was provided based on the manual count data. The average traffic volume at Marylebone Road (MY1) site between 2000 and 2007 was about 3327veh/hr, and the 95th percentile was 5853veh/hr. The HDV vehicles constituted about 10% of the total traffic. The traffic volume on London North Circular Road near BT4 site was higher than the traffic volume on Marylebone Road. It carries an average hourly traffic of 4389veh/hr and can reach up to a maximum of 7955veh/hr during peak periods. Cromwell Road (KC2) and Old-Street (HK6) have lower traffic volume than Marylebone Road with an average traffic volume of 2124veh/hr and 1995veh/hr respectively. The average traffic volume on Otley Road at the Instrumented Junction in Leeds was 1198veh/hr much higher than the 450veh/hr on the north lane. The maximum traffic volume on Otley Road was 2776veh/hr while on the North Lane; it was 1068veh/hr. The percentage of the HDV vehicles at the Instrumented Junction roads was about 11%. Figures 4.3 and 4.4 shows the time variation plots of the traffic volumes on some London roads and at the Instrumented Junction Leeds respectively. The traffic flow reaches a peak at around 8:00 am in the morning and maintain that volume until around 7:00 pm. when it begins to fall until next morning.

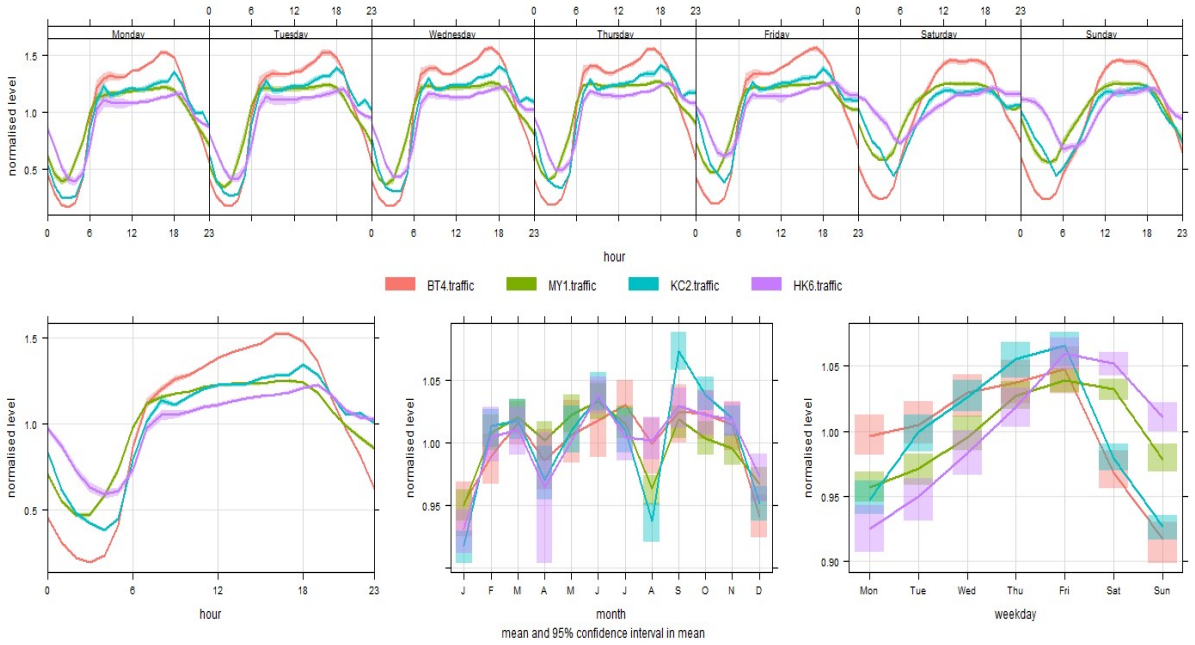


Figure 4.3 Temporal variation plot for traffic volumes (veh/hr) on some roads in London

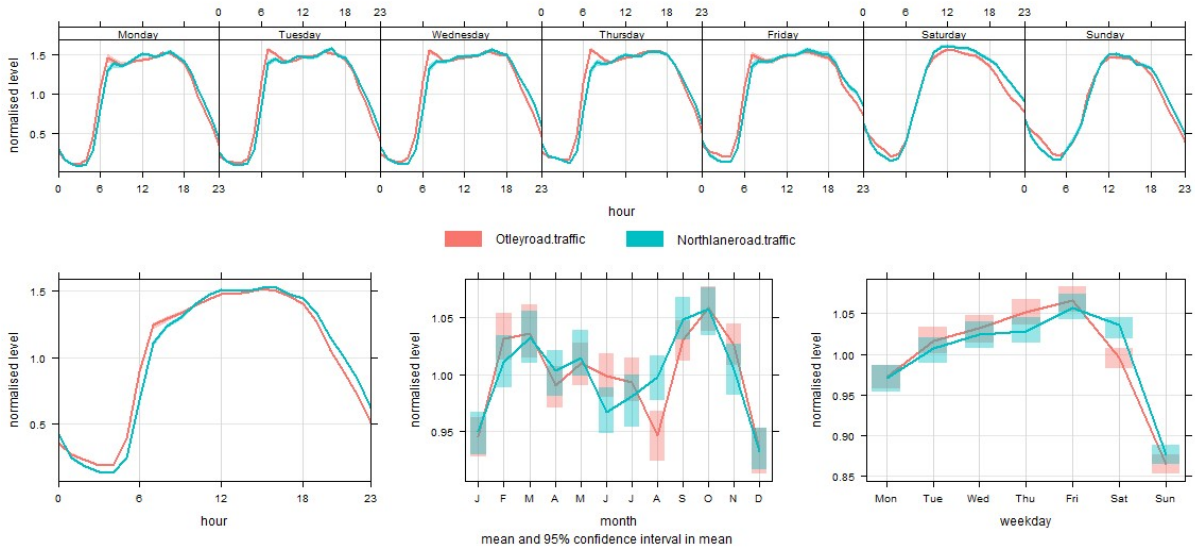


Figure 4.4 Temporal variation plot for traffic volumes (veh/hr) on some roads at Instrumented Junction in Leeds

4.4 Description of Meteorological Data

The meteorological data include wind speeds, wind direction in-street flows, solar radiation, relative humidity and ambient temperature.

The average wind speed measured at Heathrow airport between 2000 and 2012 was 2.1m/s and the 95th percentiles, and the maximum wind speeds were 9.9m/s and 4.3m/s respectively. However, in Leeds, the average speed measured between March 2009 and March 2010 was 2.8m/s and reaches up to a maximum of 12.1m/s with the 95th percentile of 5.8m/s. The winds at the two sites mostly blow from south-west and west directions as indicated by the wind roses shown in Figure 4.5. The wind rose plot represents an average London (left) and Leeds (right) hourly averages of wind speed and wind direction distributions. In the plots, the wind direction intervals were rounded up to 30^0 and the wind speed interval was set to 1.2m/s. The widths of the "paddles" represent the wind speed while their lengths describe the frequency of counts by wind directions.

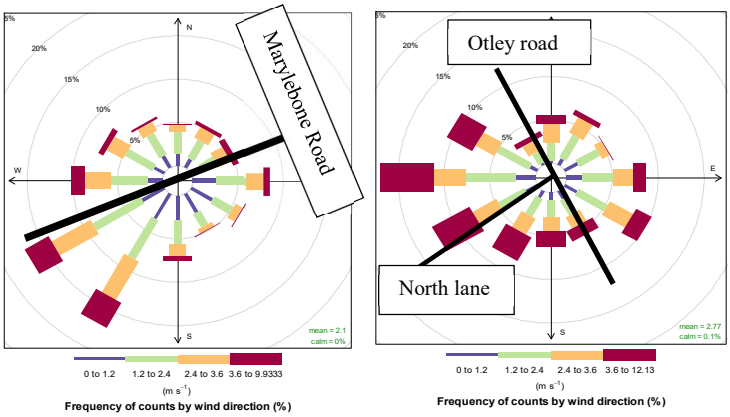


Figure 4.5 Wind characteristics at London Heathrow (left) and Instrumented Junction in Leeds (right)

In London, the dominant winds were from the Southwest and West directions. The directions of the dominant winds at the sites govern the location of the air quality monitoring sites. For example, at Marylebone Road, the air quality monitoring site is located to the south of the road to take the advantage of the effect of the cross Canyon vortex usually caused by the prevailing wind. The prevailing wind makes the flow circulate within the street canyon and deliver most of the pollutants to the leeward side of the street canyon (Tomlin et al., 2009). At the Instrumented Junction, the prevailing winds were mostly coming from West,

Southwest, Northwest and Southeast directions. ENV2 and ENV3 sites were located to the West of Otley Road and the south of north lane respectively. Considering the dominant wind directions in Leeds, these positions are the leeward sides of the street canyons where ENV2 and ENV3 are located. However, ENV1 was situated on the Windward side of the Otley Road. The temperature in both London and Leeds fluctuated between -6.4 in the winter and reached up to 37^0C in summer while the average temperature was 10^0C in Leeds and 12^0C in London. Other meteorological variables considered were solar radiation, rainfall, and relative humidity and barometric pressure.

4.5 Pollutant Data

4.5.1 Particle Concentrations

Figures 4.6 and 4.7 show the summary of the particles data collected at one representative site (MY1) in London and the sites at Instrumented Junctions Leeds. The blue and red colours on the rectangular bar at the bottom of the plots indicate the availability and non-availability of the data respectively. The percentage of the data captured for every year is written in green on the upper part of each year data plot. The minimum, maximum, number and percent of missing data, mean, median and the 95th percentile for each variable plotted are shown in black. The panel to the right of the time series plots is the density plots indicating the distribution of the data over the selected periods.

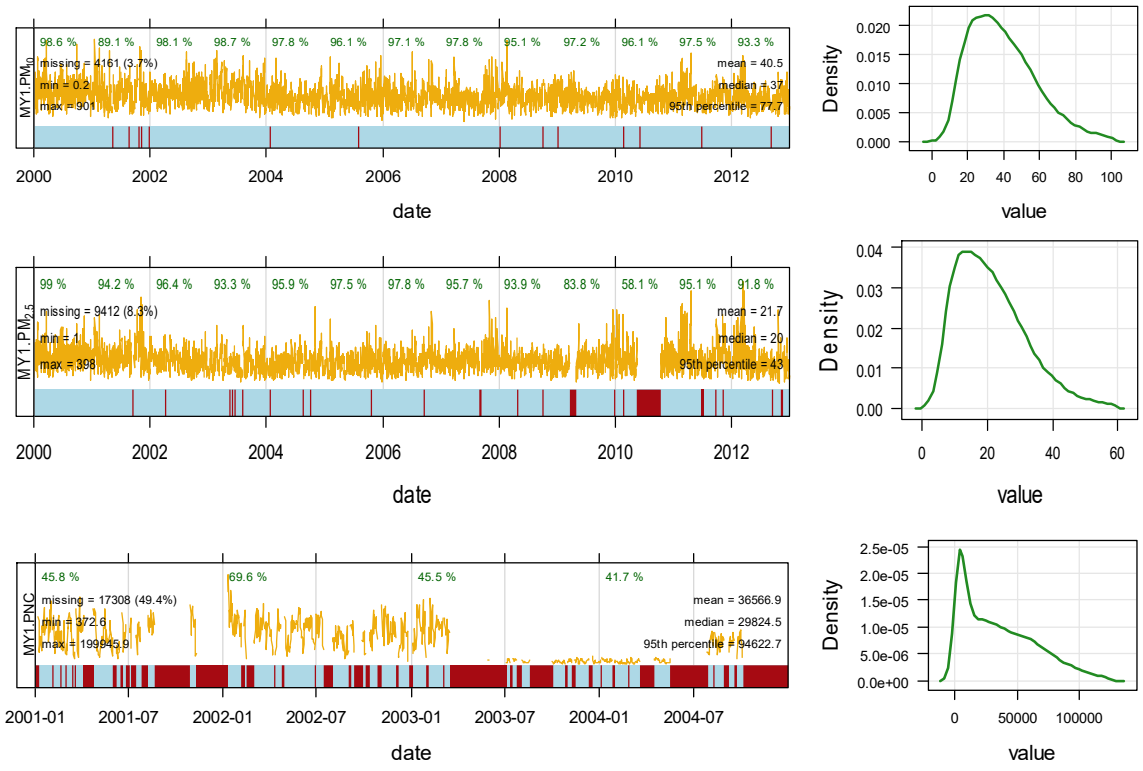


Figure 4.6 Summary plots of the particles concentrations (µg/m3) data at Marylebone road

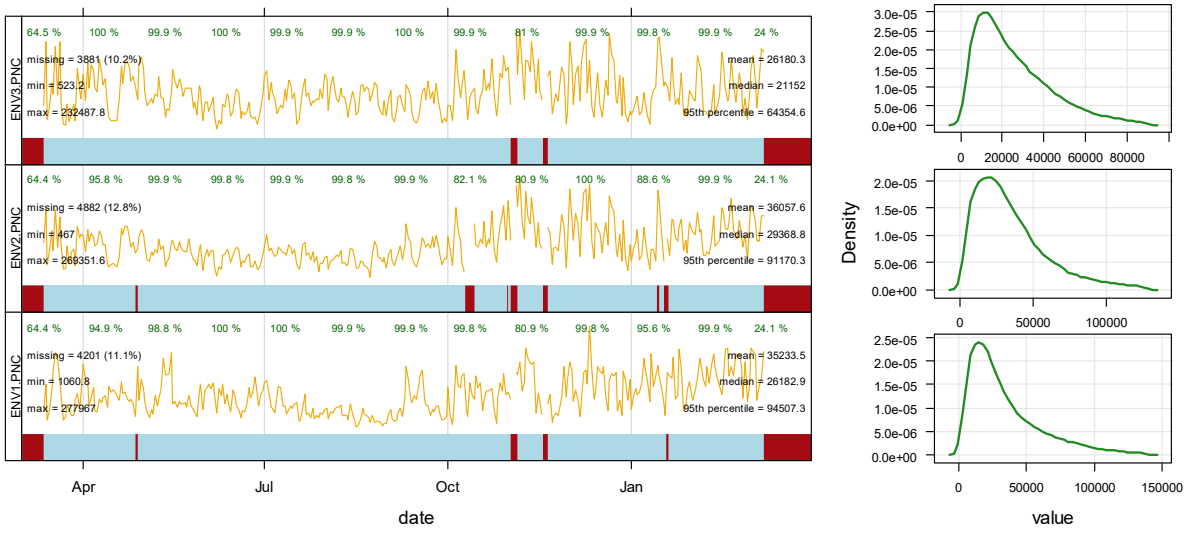


Figure 4.6 Summary plots of the particles concentrations (µg/m^3) data at Instrumented Junction in Leeds

4.5.2 Description of Hourly Particle Concentrations.

The PM_{10} hourly concentrations have been collected from twelve roadsides and six background monitoring sites as shown in Table 4.1. The average mean PM_{10} concentrations at the sites were between 22.7μg/m^3 at GR5 and 43.33 μg/m^3 at MY1. In most of the sites, the average PM_{10} concentrations were under the EU limit value of 40 μg/m^3 annual mean. The 95th percentile of the PM_{10} concentrations at the sites ranged from 49.7μg/m^3 at GR5 to 73.8μg/m^3 at GR8. The percentages of missing data in all the sites selected were less than 10% except at KC2 where the missing data was up to 19%. The average $PM_{2.5}$ concentrations at all the sites were also below the EU target value of 25 μg/m^3 annual mean. It ranged between 14μg/m^3 and 18 μg/m^3 at the five of the six roadside sites while it was 22.4 μg/m^3 at MY1. The 95^{th} percentiles of the $PM_{2.5}$ concentrations range from 29 μg/m^3 to 47 μg/m^3 at MY1. The percentages of $PM_{2.5}$ missing data at five sites were less than 10% while it was up to 18% at MY1. The levels of PM concentrations were higher from 2000 to 2004 and decreased continuously from 2004 to 2008 and maintained similar levels up to 2012. The PNC data collected at MY1 constitute only 50% of data between 2001 and 2004 while those collected at Instrumented Junction Leeds was for the period between March 2009 and March 2010. The average concentration at MY1 was 36566.9 number/cm^3 while, at ENV1, ENV2 and ENV3 were 35233.5 number/cm^3, 36096.7 number/cm^3 and 26205.6 number/cm^3 respectively.

4.5.3 Long-Term Trends of Roadside Particles in London

The long-term trends of the particle pollutants were estimated to provide a general view of the concentration levels at the sites where there was sufficient data. Moreover, to assess the significance of the changes in the concentration levels where they occur, the trends were calculated by fitting a smooth line to the monthly mean of roadside particulate matter using generalise additive modelling. The method was used such that the amount of smoothness in the estimated trend was optimised (Carslaw and Ropkins, 2012). More information about this method can be found in (Carslaw et al., 2007). A nonparametric slope estimator i.e. Theil-Sen function (Sen, 1968, Theil, 1992) was used to assess the significance of the trends. The results of the trend estimation are shown in Figures 4.8 and 4.9. The average trends in µg/m^3/year are displayed in the upper left corner of the panels in Figure 4.9. The 95% confidence intervals in the estimated slopes are shown in the parenthesis. The thick red line indicates the trends while the dashed lines indicate the 95% confidence intervals for the estimate of the slopes. The stars attached to the values indicate the levels of significance of the trends measured by *p-values*.

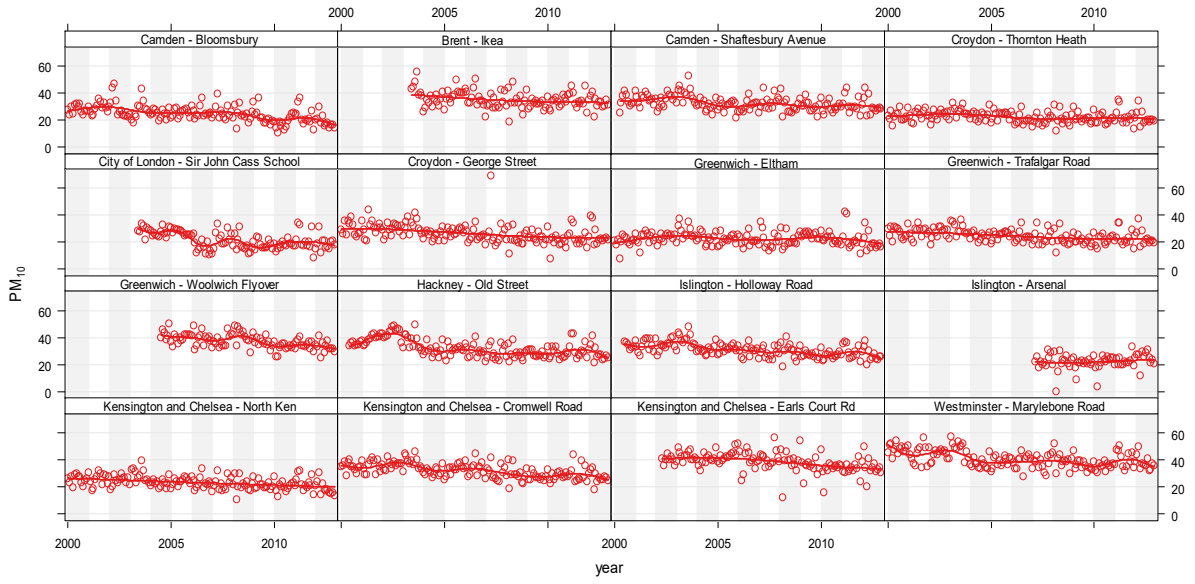


Figure 4.7 Long-term trends of PM_{10} (µg/m^3) estimated from monthly mean concentrations

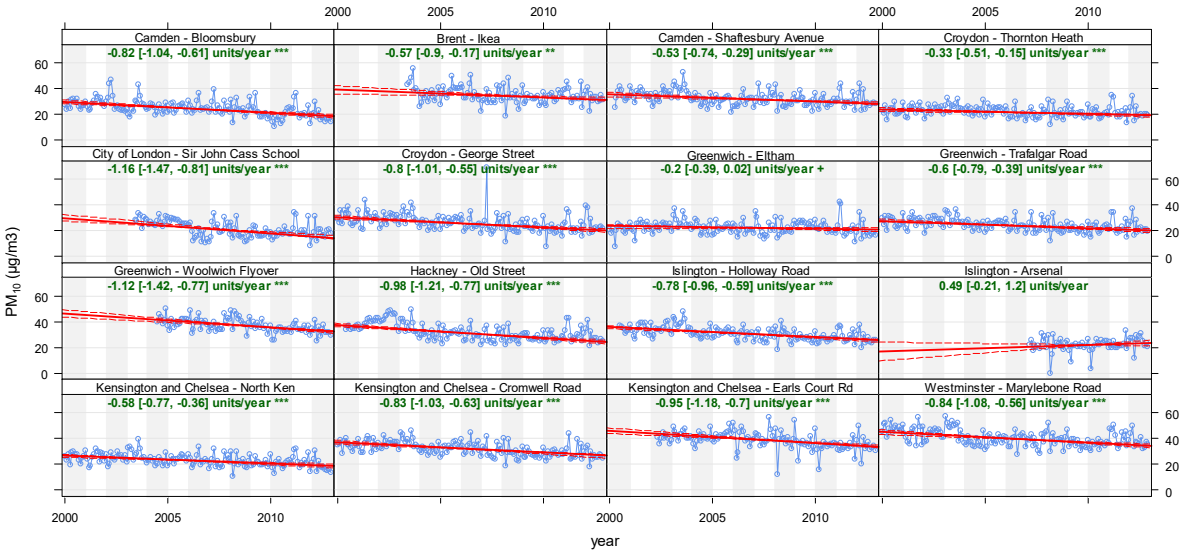


Figure 4.8 Significance of the trends in PM_{10} at London monitoring sites

The PM_{10} concentrations decrease slightly from 2000 to 2008 at most of the sites, and after that, it became somewhat constant as shown in Figures 4.8 and 4.9. There was no distinction in the trends between the background sites and the roadside or kerb sites they all seem to have closely related trends. However, the trends at Croydon - Thornton Heath (CR3) and Greenwich - Eltham (GR4) were shown to be less significant compared to the rest of the sites. There was an increase in the levels of particulates at some sites in inner and central London (i.e. MY1, HK6, BL0 and IS2. see Table 4.1) in 2011 and then decreased in 2012.

4.6 Correlation Between the Traffic, Meteorological and Pollutant Variables

The correlation between the air quality variables will help in identifying the most important variables that can be used as predictors of the particle concentrations. The correlation matrix of the variables in the data collected was derived. The Matrix explores the correlation between the roadside particles concentrations, background particles concentrations, roadside increments (i.e. Roadside – Background), SO_2, NOx, and traffic volume. Moreover, the correlation between speed, wind speeds and wind directions and the variables above at London MY1 and Instrumented Junction.

In Figure 4.10, it can be seen that the wind speeds and wind directions have more negative correlations with the background than the roadside pollutants. The negative correlation with the wind speed indicates the effect of atmospheric turbulence which enhances the ventilation of the urban environment and, in turn, reduces the pollutant concentrations. The roadside particle concentrations and their corresponding road increments have also shown a negative correlation with the traffic speeds signalling the possibility of low concentrations at higher speed and vice versa. The traffic induced turbulence is high at higher speeds, thus, affecting the concentration levels in the street canyons by enhancing the mixing of the pollutants released by the vehicles. This enhancement reduces the levels of the particle concentrations in the street. The HDV traffic has a higher correlation with the pollutants than the LDV traffic except for PNC where it has shown more correlation with the LDV traffic. It could also be observed that the HDV traffic is more associated with the particles with higher diameter than the smaller particles.

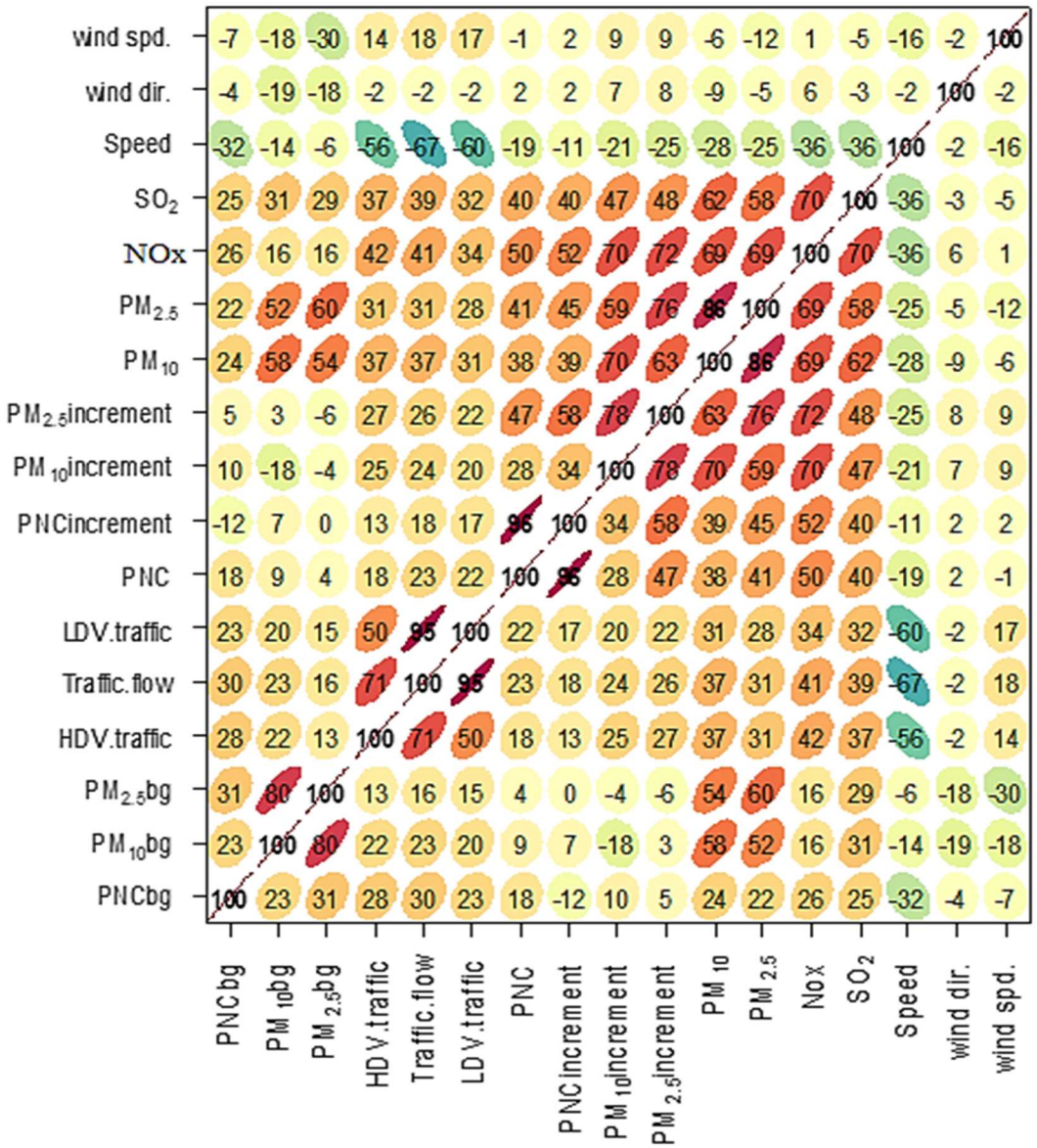


Figure 4.9 Correlation between particle concentrations, traffic variables, and other pollutants at MY1 site

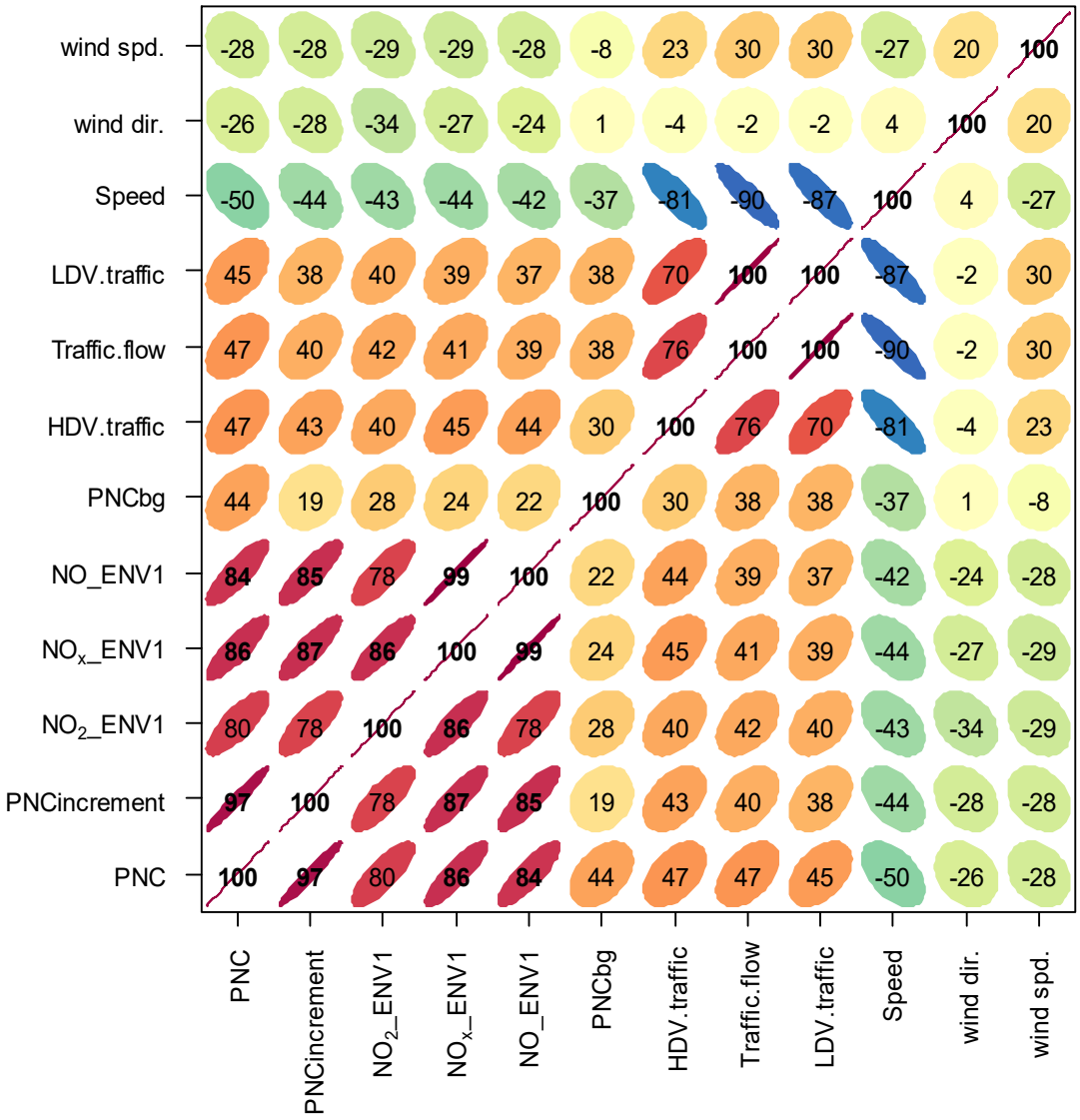


Figure 4.10 Correlation between PNC concentrations, traffic variables, and other pollutants at Instrumented Junction Leeds

There is also a strong relationship between the particles concentrations and the roadside NOx concentrations indicating a strong association between the traffic and the roadside particles since NOx is said to be mainly from road traffic in urban areas. Another strong relationship observed was between the particles and the SO_2 pointing to the contribution of the secondary particles that are formed as the results of chemical reactions after being released from the vehicles. The above analysis could help us to assume that the variables mentioned above can be used as the predictor variables for the roadside particles. However,

the cost of providing these data and their availability for long enough durations to be used for training the machine learning or statistical models with the required accuracy and reliability is a source of concern. Therefore, it is imperative to devise a means of selecting the most appropriate and most important variables to the accuracy of the models to be developed.

4.7 Validation of The Missing Data Imputation

Before the final use of the missing data algorithm, it is imperative to evaluate its performance and to examine how the imputation affects the observed data. To achieve this goal PM_{10} data from nineteen London air quality monitoring sites was prepared.

The data consists of all the predictor variables and the response variables for the prediction of PM_{10} where the columns and the rows in the data represent partially observed variables and individual observations respectively. The data was then divided into the rows with complete observations for all the variables and the rows with at least one missing observation in one or more of the variables. For the validation, the data subset with the complete observation was first perturbed with the 5%, 10%, 20% and 30% randomly inserted missing values respectively. Each of the perturbed data was then imputed using random forests method of MICE software. During the imputation process, the algorithm repeats the imputation five times for each variable. The distributions of the imputed data (red lines) and the observed data (blue lines) are shown in Figures 4.12 for the 5%, and Figures C.1 - C.3 in Appendix C for 10%, 20% and 30% missing data respectively. In Figure 4.12, it could be observed that there is little difference between the five imputations for each variable which shows the consistency of the imputation method. Also, the distribution of the imputed values compares well with the observed values with a slight underestimation of the higher values.

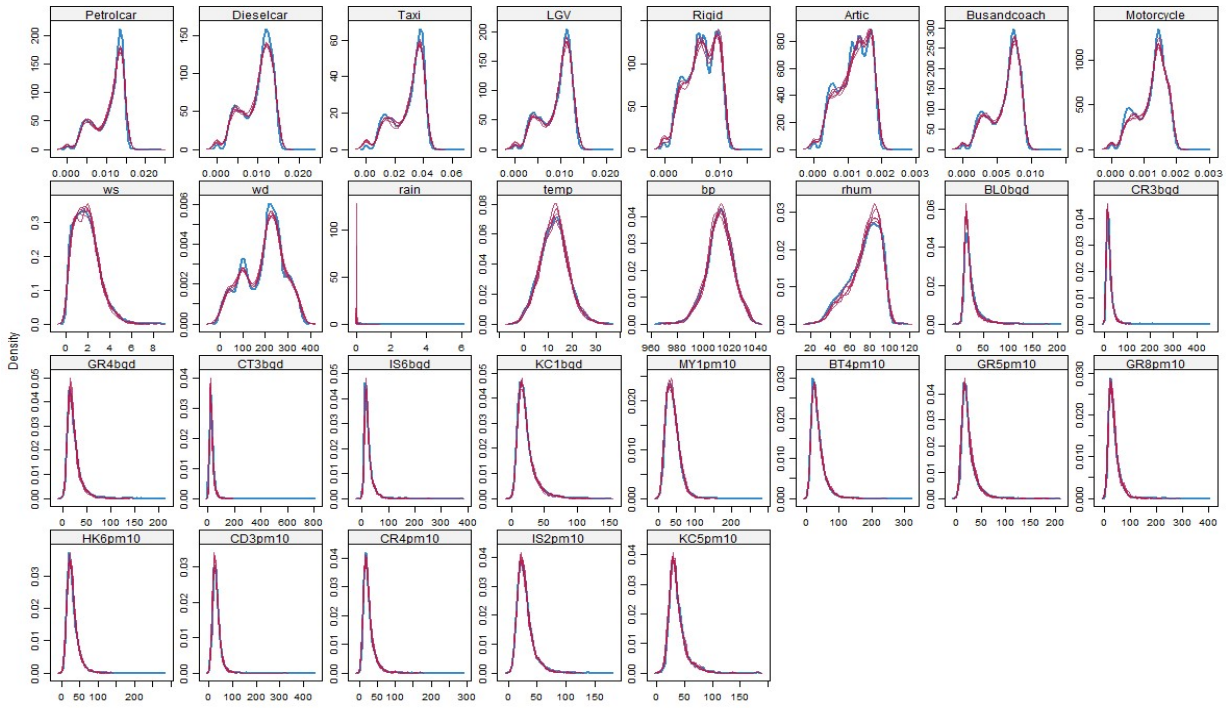


Figure 4.11 Density plots of the imputed data with 5% missing

The accuracy of the imputation decreases with the corresponding increase in the percentage of missing values. To explore the effect of the imputation on the observed data and the predictions their corresponding basic air quality statistics were estimated for ten road monitoring sites as shown in Figure 4.16.

Figure 4.13 shows that both the observed and the imputed data have similar air quality statistics. However, the number of days when PM_{10} concentrations is greater than $50\mu g/m^3$ have been overestimated estimated by 1 to 4 days in most of the sites. But in KC2 where the imputed dataset constitutes up to 27%, the number of days with PM_{10} concentrations greater than $50\mu g/m^3$ was four times the number of days found in the observed data. It is important to note that KC2 was the only site with more than 15% missing data.

Figure 4.12 Air quality statistics for the ANN predicted, observed, and the imputed observed PM_{10} ($\mu g/m^3$) data collected from 10 monitoring sites.

Note: in Figure 4.13 Obs is observed data, Obsimp is imputed observed data. BT4, CR4, KC2 and IS2 are the air quality monitoring sites.

Having satisfied that the imputation has less impact on the observed PM_{10}, it is also important to examine its effect on the accuracy of the models. For this reason, we used ANN method to train three different models with different imputation pattern. The first pattern is that no imputation was carried out, the second pattern is that the imputation was carried out for both the training and testing data sets. The third pattern is that the imputation was only made on the test data sets. Figures 4.14 and 4.15 show the normalised mean biases and the normalised mean gross errors for the models.

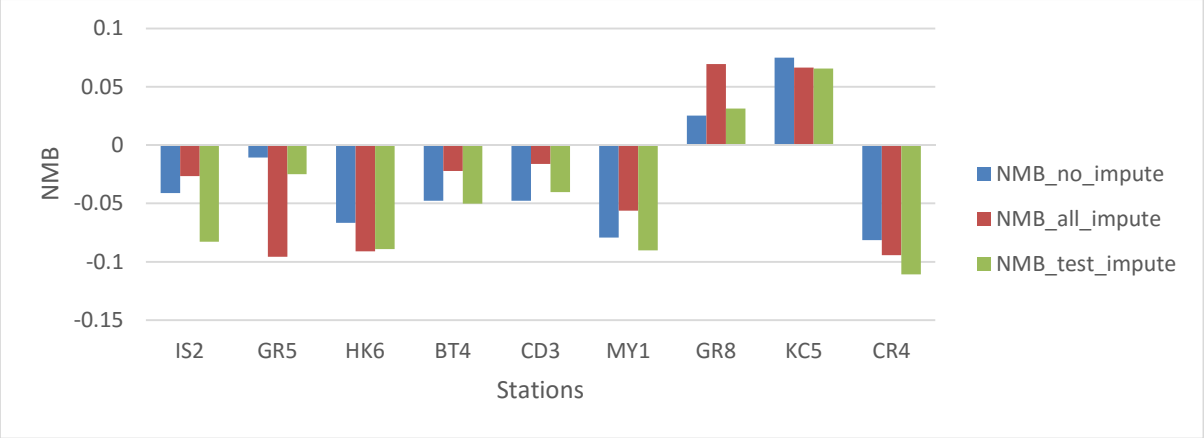


Figure 4.13 Bar charts comparing the Normalised mean bias of the ANN models trained with different missing data imputation pattern.

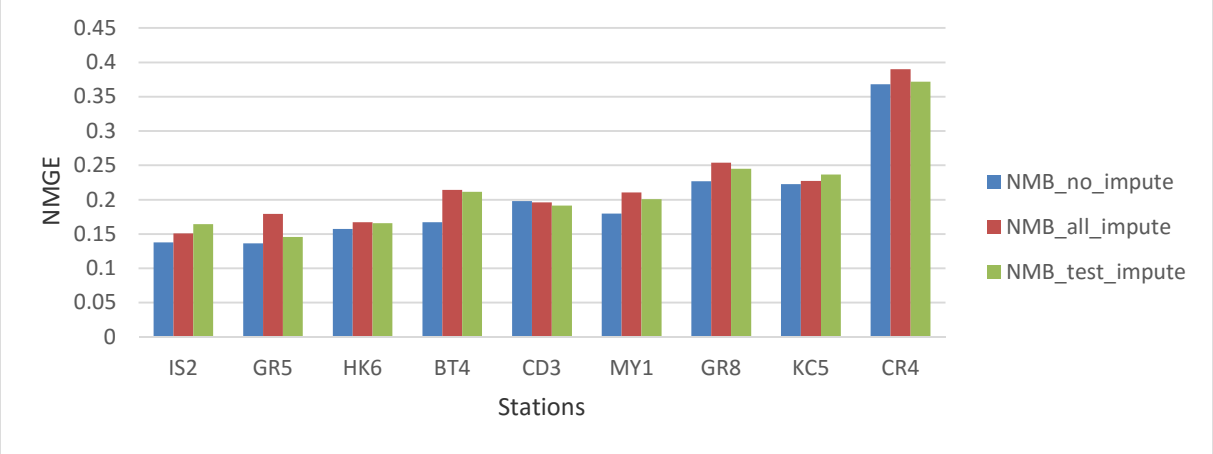


Figure 4.14 Bar charts comparing the Normalised mean gross errors of the ANN models trained with different missing data imputation pattern.

The models showed little difference in their estimations which could be seen from their normalised mean biases and gross errors shown in Figure 4.14 and 4.15. They showed similar errors on all the sites but with varied mean biases. This shows that whichever imputation pattern is adopted, the results of the predictions might be similar. Therefore, the imputation was carried out for both the training and the test data to maximise the use of the available data for the PM_{10}, $PM_{2.5}$ and PNC predictions.

After obtaining satisfactory results from the imputation validation, data covering the period between 2007 and 2012 was collected from ten PM_{10} and six $PM_{2.5}$ roadside sites in London. The sites were selected based on the availability of either PM_{10} or $PM_{2.5}$ data for the specified period and the percentage of data that is missing. All the sites selected have more than 90% of the data available except KC2 site where only 73% of the data was available. The PNC data collected from the Instrumented Junction in Leeds covering March 2009 to March 2010 was also prepared. The imputation software was then used to impute the missing data, and the imputed data was then used to train ANN, BRT, and SVM models.

4.8 Summary

This chapter presents an overview of the data collected for the purpose of this study. The description of the air quality monitoring sites used, brief description and descriptive statistics of the traffic, meteorological and pollutant variables were given and the following observations were made.

The monitoring sites with higher traffic volume also have higher concentrations. Most of the sites have met EU annual limits of PM_{10} Concentrations. However, there are sites where the levels of the particles concentrations are often high. It was also found out that there was no appreciable decrease in the concentration levels from 2008 to 2012. Some of the sites in inner London were found to have increased concentrations between 2011 and 2012 putting so many questions on the various air quality control measures put in place.

Traffic volume, gaseous pollutant concentrations and background particle concentrations were found to have good correlation with the roadside particle concentrations. These correlations make them perfect predictor variable candidates for the roadside particle concentrations.

The dominants wind directions were identified to be South, Southwest and East in London while West, Southwest, Northwest and Southeast directions in Leeds.

A detailed analysis of the relationships between the particle concentrations, the winds and the location of the monitoring units will be given in chapter five. In this manner, combining the conclusions drawn from Chapters 4 and 5 will allow for the selection of possible predictor variables for the modelling and also determine the dominant sources of the particle concentrations at the sites.

Chapter 5

Temporal and Spatial Analysis of the Roadside Particles

5.1 Introduction

In this chapter, the temporal and spatial relationships between road traffic and particles concentrations collected at the monitoring sites have been analysed. Furthermore, a bivariate polar plot (Carslaw and Ropkins, 2012) of the particle concentrations was obtained for each monitoring site with a view to discriminating between the contribution of the road traffic sources and other sources of the particles. A bivariate polar plot describes the joint variation of pollutant concentrations, wind speeds and wind direction on a continuous surface using polar coordinates. The results of the polar plot analysis were used to estimate the upper limit of road contributions to the roadside particle concentrations. The meteorological, traffic and pollutant data associated with the upper limit of road contributions were extracted in preparation for the development and training of models for predicting the road traffic contributions. The chapter concludes with the summary of the findings discussed in the chapter.

5.2 Temporal Variation of Traffic Volume and The Particles Concentrations

The particle concentrations collected from the monitoring sites located near roads. are assumed to be dominated by road traffic. Therefore, it is imperative to establish this relationship and to examine the significance of the contribution of the traffic and other sources. This analysis will help determine the factors that affect the particle concentrations and hence determine the appropriate predictor variables for their predictions. Since road traffic follows definite hourly, daily and weekly patterns, the relationship between the hourly, daily and weekly variation of the roadside particles and the traffic volume would

provide more information on how much of the particulate matter derives from the traffic. Figure 5.1 shows the time variation plots for the traffic and particle concentrations collected at a street canyon sites MY1, BT4, KC2 and the Instrumented Junction sites. From the figure, the particle concentrations follow the hourly, daily and weekly variations of the traffic closely. The concentrations are high when the traffic volume is also high and vice versa indicating a strong relationship. The traffic volume does not vary much monthly. However, the particle concentrations vary widely with the months showing that there are other factors such as meteorology that affect the monthly and seasonal variations of the particle concentrations which has less or no effects on the traffic. At the street canyons, the levels of $PM_{2.5}$ and PNC are lower early in the morning than PM_{10} concentrations indicating that they are more related to traffic sources than PM_{10}.

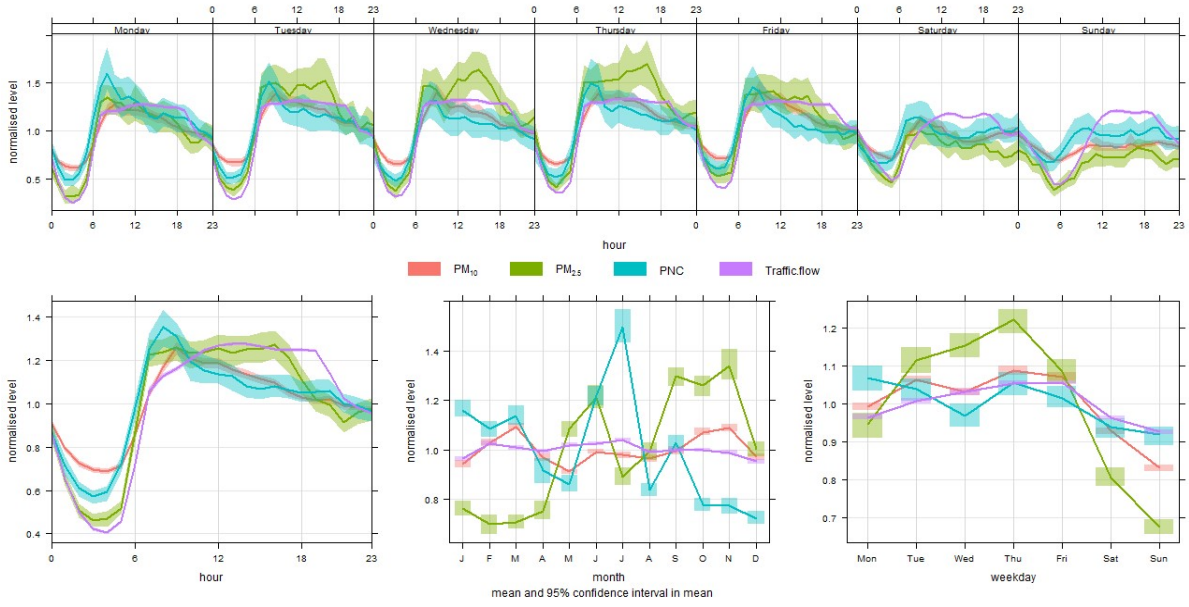


Figure 5.1 Temporal variation of traffic volume and particle concentrations for Marylebone Road (MY1)

The particle concentration levels increase as the traffic volume increases in the morning until around noon when the traffic volume remains somewhat constant. The traffic volume starts to decrease at around 8:00 pm till the next morning. At the same time, the particle

concentrations only remain high until noon when it will start decreasing slowly until the next morning. The level of the particle concentrations on weekends is fairly the same throughout the day at most of the London sites. At the Instrumented Junction in Leeds, the PNC levels increase at around 6:00 pm till 8:00 pm showing some increase in activities at the junction around that time. Although there are some minor differences between the hourly weekly and daily variations of the particle concentrations and the traffic volume, the relationship between them is so strong that it could be suggested that most of the particles are emitted by the vehicles on the road adjacent to the monitoring stations. Therefore, given the strong relationship, it is argued that traffic variables are appropriate to be used as the predictor variables in the modelling of roadside particles.

5.3 Analysis of The Relationship Between the Particles, Traffic Volume and Wind Directions

In section 5.2, it has been established that the levels of roadside particles vary temporally with traffic volume, however, there it needs to be determined whether it was the same traffic volume on the road that emitted the particles, or the particles are just responding to the general traffic profile in the area. Therefore, the following analysis tries to explore the relationship between the increasing traffic, the particle concentrations and the wind directions at some of the monitoring sites.

The trend level plots in Figure 5.2 shows the annual mean particle concentrations for each quantile of the traffic volume and the eight wind sectors. The relationship between the traffic volume, the location of the road and the particle concentrations can be established if, for each quantile of the traffic volume, the higher particle concentrations are more associated with the winds coming from the directions related to the road. These winds could carry the pollutants along the road, across the road, or from the same direction as the location of a

monitoring unit in the case of street canyons. In Figure 5.2, it could be seen that all the particle metrics showed strong association with the winds from the west, south-west, south, south-east and east. The monitoring unit at MY1 is located to the southern side of Marylebone Road in a street canyon. Therefore, the flows along the street and the canyon recirculation vortex can be said to be in effect where they carried most of the concentrations to the leeward side of the canyon. Moreover, this property is shown at all the levels of the traffic flow. Therefore, the road traffic is likely to have been responsible for the particle concentrations at this site.

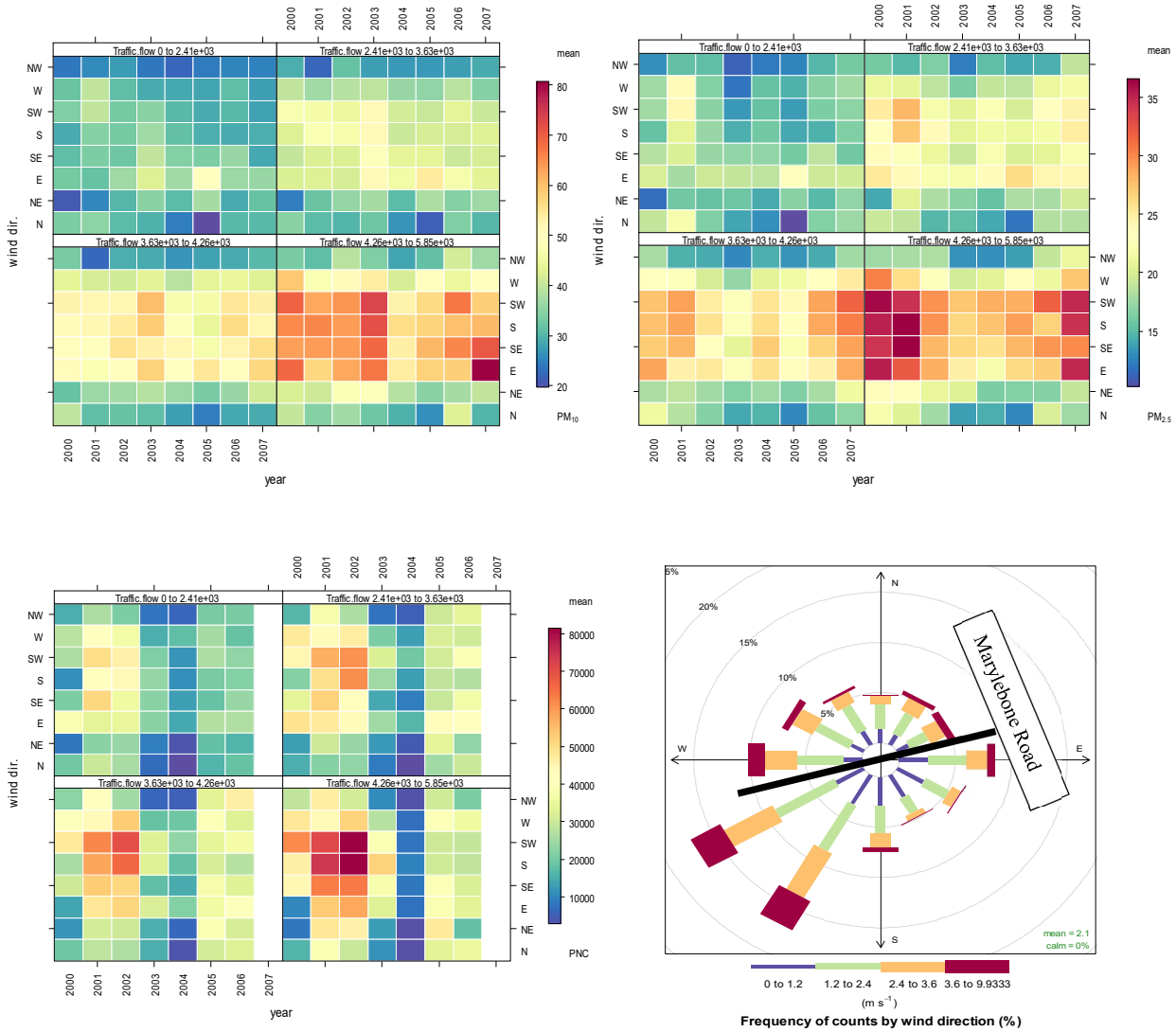


Figure 5.2 Trend level plot showing the relationship between traffic volume, particle concentrations and wind directions at Marylebone Road (MY1)

ENV1 is located on the eastern side of Otley Road in Leeds, which runs through the north-west and south-east axes and approximately 50m northwards from the Instrumented Junction. In Appendix C - Figure C.4, it could be seen that (top panel), it could be observed that the higher PNC at ENV1 is associated with the north, north – east, and east winds. The elevated concentration from the northeast might have occurred due to the downward flow caused by the interaction between the reference winds in the range 165^0 to 260^0 and façade of the building (Tate et al., 2009). However, the elevated concentrations associated with the

north-west winds represents the concentrations channelled through the street canyons. ENV2 is located to the western side of Otley Road at the heart of the Instrumented Junction formed by Otley Road and North-Lane. The higher PNC concentrations at this site are more concentrated on the southern link of Otley Road and North-Lane (Southwest). Therefore, it could be said that the channelling flow is dominant in this case (see Figure C.4 middle panel). ENV3 is located on the southern side of the North - Lane approximately 25m from the junction. The trend level plot (Figure C.4 bottom panel) shows that the elevated PNC concentrations at this site are associated with south and south-east winds. Although at ENV3, the flow along the street canyon dominates, Tate et al. (2009) discovered that there was evidence of across Canyon vortex or helical flow regimes associated with north-west and northeast winds that will contribute to the higher concentrations related to the south and south-east. This phenomenon also shows that the roads contributed most to the particle concentrations at the site.

5.4 Analysis of Spatial Distribution of Total and Road Increment PM_{10} Concentrations

In section 5.1 and 5.2, it was shown that the particle concentrations have strong temporal and spatial relationships with road traffic. However, it was not possible to isolate other pollution sources that might have contributed to the overall roadside concentrations. It is also important to estimate the percentage contribution of the roads near the monitoring stations to establish their impact on the overall air quality of their immediate environment. Moreover, such estimates can be used to train a machine learning model to predict the future contribution of these roads should there be any changes. For the purpose of these objectives, the ten PM_{10} sites in London were selected because of the diversity of the locations of the monitoring sites and the availability of long-term data. The method adopted for the

estimation of the road contributions derives from the method developed by Carslaw et al. (2006) to quantify the contribution of aircraft and other on airport sources to ambient oxides of nitrogen. The same approach was also followed by Masiol and Harrison (2015) to estimate the impact of Heathrow Airport and the M25 and M4 motorways on the surrounding air quality. The method involves estimation of the road increment by subtracting the background concentration upwind of the road site and the use of bivariate polar plots to locate the wind sectors related to the source in question. The contribution of a source will then be estimated by isolating the data related to the time of the activities at the source, the wind sector and the wind speed. In their separate studies, Carslaw et al. (2006) and Masiol and Harrison (2015) estimated the upper limit of airport contributions considering the relevant wind sectors and the wind speeds greater than 3m/s to eliminate the influence of local sources such as roads. However this study is interested in the contributions of the local sources, hence, data covering 6:00 am to 22:00 pm associated with the wind sectors related to the roads and with wind speeds less than 3m/s are used. This segregation is done to isolate the influence of other sources far away from the monitoring units. The estimates obtained will also be used in the subsequent chapters in training machine learning models for the prediction of the roads contribution to the roadside particle concentrations. The roadside increments considered were only for the period between 6:00 am to 22:00 pm to capture the time limit within which the traffic activity is high.

5.4.1 Spatial Analysis of the PM_{10} Concentrations Using Bivariate Polar Plots (BPP)

Bivariate polar plots (BPP) for PM_{10} concentrations have been drawn for each of the ten roadside monitoring sites. For each monitoring site, four BPPs were derived, each for the total concentrations, the roadside increment, the roadside increment associated with wind speed less than 3m/s and the roadside increment related to wind speed greater than 3m/s.

The BPP of the PM_{10} concentrations at BT4 site (Figure 5.3a) shows that the higher mean PM_{10} concentrations of between 40 to 70 µg/m^3 are associated with the winds coming from the northeast and east directions. These are the winds flowing along and across the road, indicating that the road traffic caused the elevated concentrations at the monitoring station. There are also traces of higher concentrations associated with the winds coming from the north and west directions that might have been caused by the parking area behind the monitoring station. The BPP of the mean PM_{10} increment shown in Figure 5.3b also indicates that the higher mean PM_{10} increment range from 20 - 35 µg/m^3 are associated with the east and northeast winds. the elevated increments fade with increasing wind speeds showing that they were emitted from the ground level sources such as traffic.

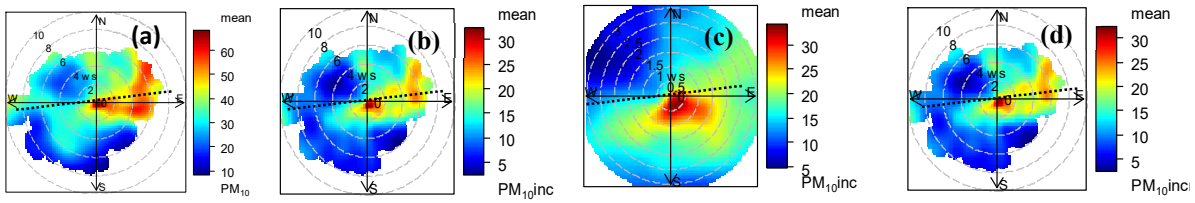


Figure 5.3 Bivariate polar plot of PM_{10} (µg/m^3) concentration at BT4 monitoring site

The BPPs were also obtained for the roadside increments associated with higher (> 3m/s) and lower (< 3m/s) wind speeds (see Figure 5.3c and 5.3d) in order to isolate the influence of the local sources (e.g. road traffic) from the long distance sources. The lower wind BPP shows that much of the road contributions are associated with the winds crossing the road from the south and south-east directions. However, the BPP for the higher winds shows that, the higher concentrations were carried by the winds flowing along the road from the northeast and east directions, and also, the concentration reduces as the wind speed increases.

The BPP of the PM_{10} concentrations at Camden - Shaftesbury Avenue (CD3) monitoring unit, shown in Figure 5.4a revealed that the higher concentrations ranging from 35 to 55 $\mu g/m^3$ are associated with the winds coming from all directions but more importantly east and west. The east and west directions coincide with the approximate directions of the traffic along St Giles High Street. However, the BPP of the mean PM_{10} increments (Figure 5.4b) shows that the higher concentrations are mostly associated with the north, north-west, west, south-west and northeast directions.

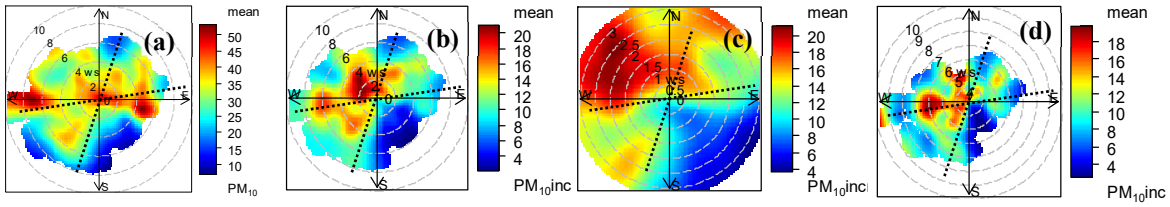


Figure 5.4 Bivariate polar plot of PM_{10} ($\mu g/m^3$) concentration at CD3 monitoring site

Therefore, the elevated concentrations at the site might have occurred due to road traffic at the junction since the monitoring unit is located to the south and west of St Giles High Street and Shaftesbury Avenue respectively. It is also possible that the prevailing winds coming from the south and the south-west come along with the high concentrations, and the recirculation flows delivered most of the concentrations from the southern and south-west links of the road. The BPP of the lower and upper wind speeds shown in Figures 5.4c and 5.5d indicates that the elevated concentrations are more associated with the channelling and recirculation flow at the junction. One important point to note here is that even at higher winds the increments seem to be relatively higher compared to the increments at lower winds.

Figure 5.5 shows the BPP of the PM_{10} concentrations at Croydon George Street (CR4) monitoring unit. It could be seen that the higher mean PM_{10} concentrations (30 - 45 $\mu g/m^3$) are associated with the winds coming from the west, north and north-east (Figure 5.5a).

These winds are the channelling flows along George Street and the northern link of Wellesley Road. The monitoring unit at this site is located to the north of the junction. Therefore, the higher concentrations might be delivered to the monitoring unit by the recirculating flow that carries the concentrations from the intersection. It could also be transported by the winds flowing parallel to George Street from the east and west directions. However, the BPP of the road increment shown in Figure 5.5b indicates that the elevated concentrations are associated with the junction, and they occur in lower winds.

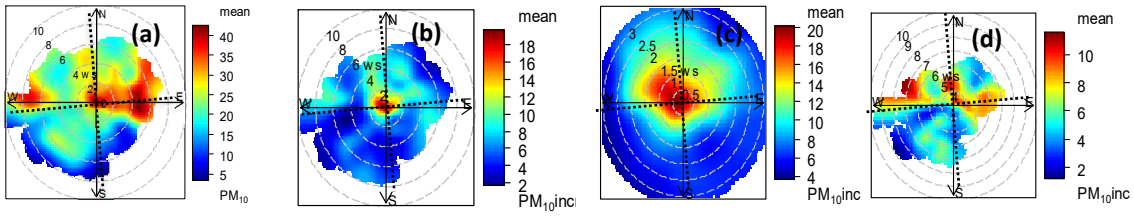


Figure 5.5 Bivariate polar plot of PM_{10} (µg/m3) concentration at CR4 monitoring site

The BPP of the road increments related to the lower wind speeds shows that the high concentrations are mostly delivered to the monitoring unit by the recirculating flows due to southern winds. However, the BPP for the higher winds indicates that the high concentrations are also channelled to the monitoring unit through George Street in addition to the recirculating flows.

At Greenwich – Trafalgar Road (GR5) monitoring site, the BPP of the mean PM_{10} concentrations (Figure 5.6a) shows that the higher concentrations ranging from 25 to 45 µg/m^3 are more associated with the winds coming from the west, northeast and east. These are the winds flowing along Trafalgar Road. However, the BPP of the mean PM_{10} increments in Figure 5.6b shows that the higher increments are related to the higher winds from the west which are the flows along the western link of Trafalgar Road. When the BPP of the PM_{10} increments related to the lower winds was separately plotted, the effect of Greenwich Park Street was clearly shown as the elevated mean increments are more associated with the

winds coming from the north (see Figure 5.6c). This elevation points to the concentration that is being accumulated at the junction as the vehicles are waiting to enter Trafalgar Road.

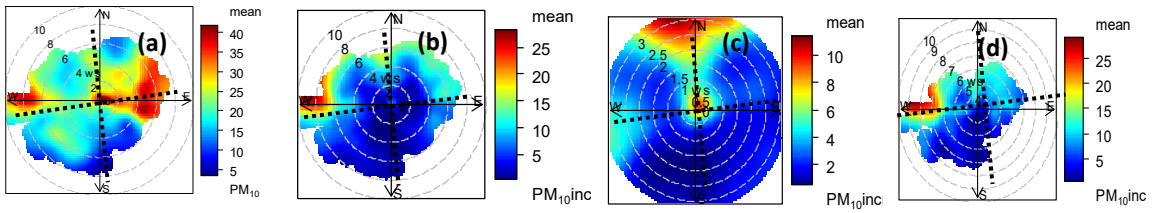


Figure 5.6 Bivariate polar plot of PM_{10} ($\mu g/m^3$) concentration at GR5 monitoring site

It is also worth noting however that the mean increments associated with the higher winds shown in Figure 5.6d are 2 - 3 times higher than what was observed with the low winds and are more related to the channelling flows along the western link of Trafalgar Road.

The BPP of the mean PM_{10} concentrations (Figure 5.7a) shows that the elevated mean concentrations ranging from 50 to 85 $\mu g/m^3$ at Greenwich – Woolwich flyover (GR8) monitoring site, are associated with the winds coming from the west. Moreover, the BPP of the average PM_{10} increment in Figure 5.7b also shows that the higher concentrations ranging from 40 to 75 $\mu g/m^3$, are associated with the winds coming from the west pointing to the contribution of the Woolwich and A102 roads. At lower winds, the higher roadside increment is associated with the winds coming from the south, south-west, west, and north-west (see Figure 5.7c).

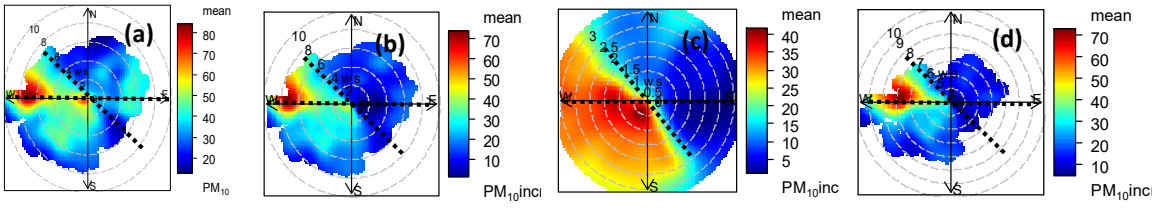


Figure 5.7 Bivariate polar plot of PM_{10} ($\mu g/m^3$) concentration at GR8 monitoring site

This association indicates that the higher roadside increments ranging 25 – 45 $\mu g/m^3$ are much more related to the Woolwich and A102 roads. However at higher winds, the

increments are more associated with the winds coming from the west which also points to the contribution of Woolwich Road especially from a far distance as shown in Figure 5.7d.

The BPP of the PM_{10} concentrations at Hackney Old Street (HK6) monitoring site shows that most of the higher concentrations ranging from 35 – 55 µg/m^3 are associated with the winds from the east, northeast, and west directions. This demonstrates that most of the concentrations were transported to the monitoring unit by the winds flowing along the street from both directions. There is also a sign of high concentrations from the north-west at higher winds which could be transported to the monitoring unit from a far distance through a wide opening at the back of the monitoring unit. However, the BPP of the PM_{10} increments indicates that the higher concentrations might have been delivered to the site from a junction located a few metres to the west of the monitoring unit and also from the western link of the road.

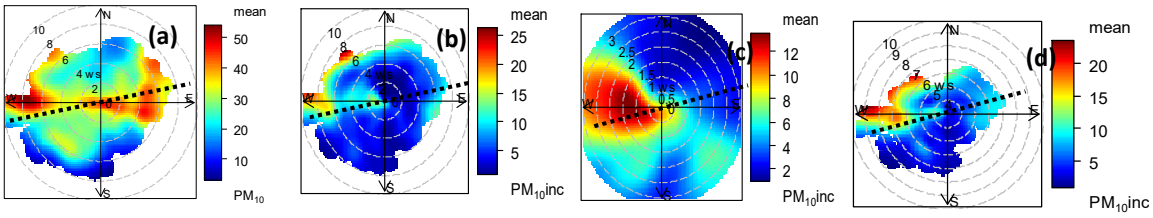


Figure 5.8 Bivariate polar plot of PM_{10} (µg/m^3) concentration at HK6 monitoring site

At lower winds, the higher road increments are mostly associated with the westerly winds. However, at higher winds, the BPP shows that most of the higher concentration increments were carried from the roundabout located a few metres west of the monitoring site and through the opening at the back of the monitoring unit.

The higher concentrations at Islington Holloway (IS2) monitoring site ranging from 35 - 55 µg/m^3, are associated with the northeast, east, south-west and westerly winds as shown in Figure 5.9a. The monitoring unit is located on the western side of the road. Therefore, the concentrations might be assumed to have been elevated by the road traffic emissions through

cross winds from northeast and east and also through the recirculated winds from west and south-west.

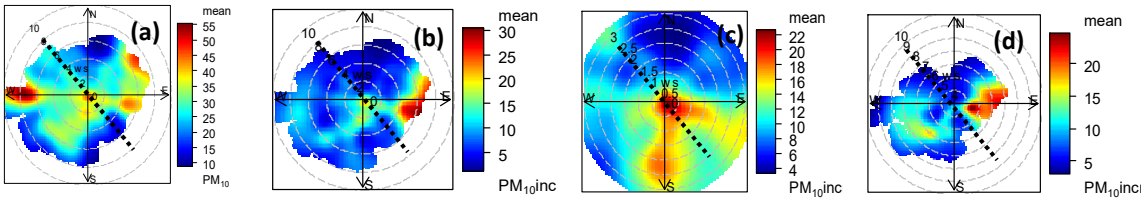


Figure 5.9 Bivariate polar plot of PM_{10} ($\mu g/m^3$) concentration at IS2 monitoring site

However, the BPP of the concentration increments in Figures 5.9b and d indicates that the higher concentrations are only associated with the higher winds from the east. Moreover, the BPP of the increments related to the lower winds shows that the canyon recirculation and the channelling flows are the dominant flows contributing to the elevated concentrations as they are associated more with the winds coming from the south and south-east as shown in Figure 5.9c.

The higher concentrations ranging from 50 to 70 $\mu g/m^3$ at Kensington and Chelsea – Cromwell Road (KC2) are more related to the higher winds from west and east. Showing that they were transported to the monitoring unit through Cromwell Road and with only little contribution from the southern link of Queen's Gate (see Figure 5.10a). The same pattern is also shown by the BPP of the concentration increment shown in Figure 5.10b, but the elevated concentrations are more related to the western link of Cromwell Road.

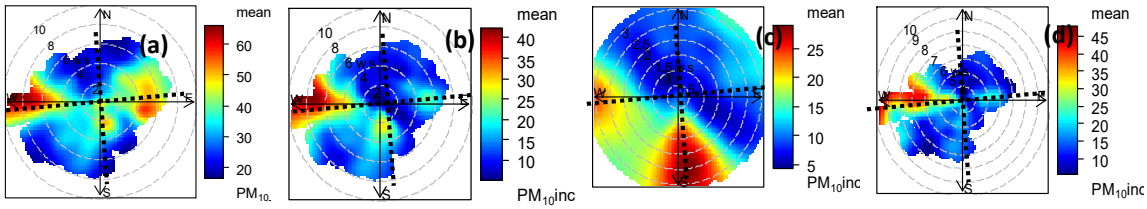


Figure 5.10 Bivariate polar plot of PM_{10} ($\mu g/m^3$) concentration at KC2 monitoring site

However, at lower winds the higher concentration increments are more related to the southern winds showing that most of it might have been emanated from the Instrumented Junction as shown in Figure 5.10c. The BPP for the higher winds in Figure 5.10d also emphasised the relationship of the higher concentration increments with the western link of the Cromwell Road.

The higher mean PM_{10} concentrations Kensington and Chelsea – Earls Court Road (KC5) monitoring unit were more associated with, the higher winds coming from north – east as shown by the BPPs in Figures 5.11a, b and c. Since the monitoring unit is located on the western side of the Earls Court Road which runs along north-west and south-east axes, it is safe to assume that the winds flowing from these directions carried the concentrations from the road to the monitoring unit.

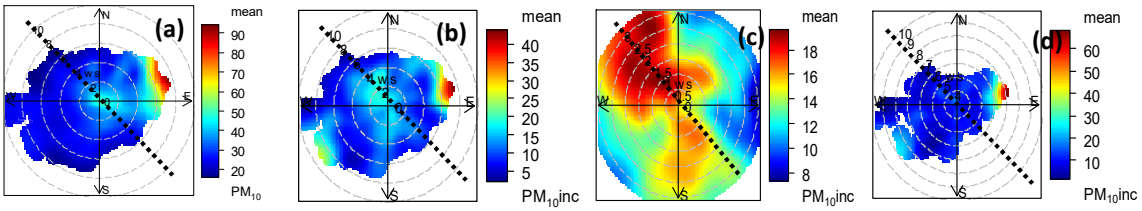


Figure 5.11 Bivariate polar plot of PM_{10} ($\mu g/m^3$) concentration at KC5 monitoring site

However, at lower winds the elevated concentration increments recorded at the monitoring site are more associated with the flows along the street from the north-west direction and also with the recirculated flows from the south.

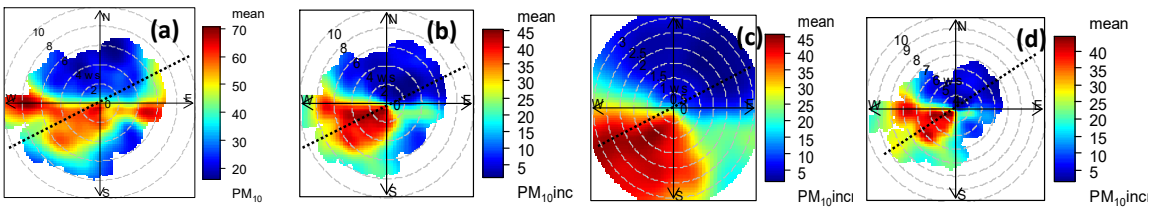


Figure 5.12 Bivariate polar plot of PM_{10} ($\mu g/m^3$) concentration at MY1 monitoring site

The higher mean PM_{10} concentrations ranging from 45 to 70 $\mu g/m^3$ at Westminster – Marylebone Road (MY1) are more related to the winds along the road and the recirculating flows within the canyon as shown in Figure 5.12a. The BPP of the total PM_{10} increments illustrated in Figure 5.12b indicates that the higher PM_{10} increment ranging from 25 to 45 $\mu g/m^3$ are more related to the winds coming from the west and south-west indicating the prevalence of the canyon channelling flows. The BPP of the increments related to the lower winds shown in figure 5.12c indicates that the elevated concentrations at the site are more related to the winds along the canyon from the west and south-west and across canyon flows from the south and south-east. However, at higher winds the BPP in Figure 5.12c shows that flows along the canyon from the south-west might carry the higher concentrations from the Baker Street junction along the road down to the monitoring unit.

5.5 Quantification of the Road Traffic Contribution to the Roadside Pm_{10} Concentrations

The bivariate polar plot analysis of the PM_{10} concentrations presented in section 5.3 provided a qualitative assessment of the likely effect on the roads near the monitoring stations. This section went further to provide a quantitative estimate of the upper limit of road traffic contribution using the reverse of the approach employed by Carslaw et al. (2006) and Masiol and Harrison (2015) for quantifying the upper limit contribution of Heathrow airport.

The method involved identifying the appropriate site pairs i.e. roadside and background sites and then estimating the roadside increment by subtracting the background concentrations from the roadside concentrations. The concentrations collected between 06:00 and 22:00 was then extracted and divided according to those associated with wind speeds greater than 3m/s and less than 3m/s. Furthermore, the BPP of the PM_{10} increment for each wind speed

class and the combined data for the period was derived. The BPPs were then used to identify the wind sectors related to the roads that contribute most to the elevated concentrations at the sites. The data for the selected wind sectors were then extracted, and their mean and frequency obtained. The upper limit road contribution was estimated as the average PM_{10} increment related to the selected wind sectors and wind speed less than 3m/s. The period between 6:00 and 22:00 with wind speeds less than 3m/s were chosen to maximise the road contribution and eliminate the influence of long distance sources respectively (Carslaw et al., 2006). The results of the estimates are shown in Table 5.1.

The first column shows the site pairs; the second column displays the wind speed classes while the third column shows the selected wind sectors for each wind speed class. The overall mean PM_{10} concentrations and the average roadside increment for wind speed/wind sector are shown in third and fourth columns respectively. The fifth column indicates the percentage of the average increment for each wind speed/wind sector as a proportion of the average PM_{10} concentrations for the whole observation over the study period. The sixth column is the percentage of the observation for each wind speed/wind sector as a proportion of the total observations at the site.

Table 5.1 Estimates of the contribution of pollution sources to roadside particulate matter (PM_{10}) between 06:00 and 22:00

Site Pairs	Wind Speed (m/s)	Wind Sector (degrees)	Total Mean PM_{10} ($\mu g/m^3$)	Mean PM_{10} increment ($\mu g/m^3$)	Percent road contribution (%)	Percentage of observations (%)
BT4 – KC1	Combined	0 – 140	37.04	22.79	62%	9%
BT4 – KC1	> 3m/s	30 – 130	37.04	21.76	59%	2%
BT4 – KC1	< 3m/s	40 – 230	37.04	22.07	60%	27%
CD3 – KC1	Combined	75 – 200	34.22	12.16	36%	49%
CD3 – KC1	> 3m/s	65 – 200	34.22	10.52	31%	12%
CD3 – KC1	< 3m/s	90 – 255	34.22	14.86	43%	21%
CR4 – HA1	Combined	70 – 255	25.62	12.16	47%	26%
CR4 – HA1	> 3m/s	70 – 240	25.62	09.27	36%	5%
CR4 – HA1	< 3m/s	90 – 180	25.62	12.81	50%	22%
GR5 – GR4	Combined	60 – 90	23.37	03.69	16%	18%
GR5 – GR4	> 3m/s	60 – 90	23.37	03.37	14%	2%
GR5 – GR4	< 3m/s	340 – 90	23.37	05.50	24%	7%
GR8 – GR4	Combined	150 – 310	40.63	22.86	56%	27%
GR8 – GR4	> 3m/s	250 – 330	40.63	24.36	60%	4%
GR8 – GR4	< 3m/s	150 – 310	40.63	24.15	59%	35%
HK6 – CT3	Combined	250 – 340	31.83	10.82	34%	18%
HK6 – CT3	> 3m/s	90 – 250	31.83	09.21	29%	4%
HK6 – CT3	< 3m/s	90 – 320	31.83	12.48	39%	17%
IS2 – IS6	Combined	140 – 330	30.73	03.04	10%	11%
IS2 – IS6	> 3m/s	180 – 335	30.73	01.66	5%	2%
IS2 – IS6	< 3m/s	145 – 330	30.73	07.76	25%	21%
KC2 – KC1	Combined	60 – 300	33.71	12.52	37%	39%
KC2 – KC1	> 3m/s	60 – 300	33.71	15.59	46%	2%
KC2 – KC1	< 3m/s	140 – 280	33.71	12.71	38%	31%
KC5 – KC1	Combined	190 – 270	35.83	13.62	38%	7%
KC5 – KC1	> 3m/s	0 – 90	35.83	11.95	33%	2%
KC5 – KC1	< 3m/s	50 – 140	35.83	16.39	46%	44%
MY1 – BL0	Combined	140 – 270	43.33	26.89	62%	35%
MY1 – BL0	> 3m/s	170 – 270	43.33	24.34	56%	4%
MY1 – BL0	< 3m/s	150 – 270	43.33	25.20	58%	35%

The upper limit of the road contribution constitutes between 24% and 62% of the mean PM_{10} concentrations at the sites. The sites with the higher average traffic volume seem to have higher contributions irrespective of their locations. For example, BT4, GR8 and MY1 contributed about 58 – 60% of the mean PM_{10} concentrations recorded at their respective locations as shown in Figure 5.13. The frequency of observations associated with the upper limit estimation constitutes about 21 to 44% of the total observations at the sites. There was not much difference between the average contribution of other sources and the roads regarding the level of concentrations. However, there is a huge difference in the frequency of their respective observations. It was observed that the frequency of observations associated with the higher wind speeds at wind sectors related to the roads constitute only about 2 – 4% of the total observations as shown in Figure 5.14.

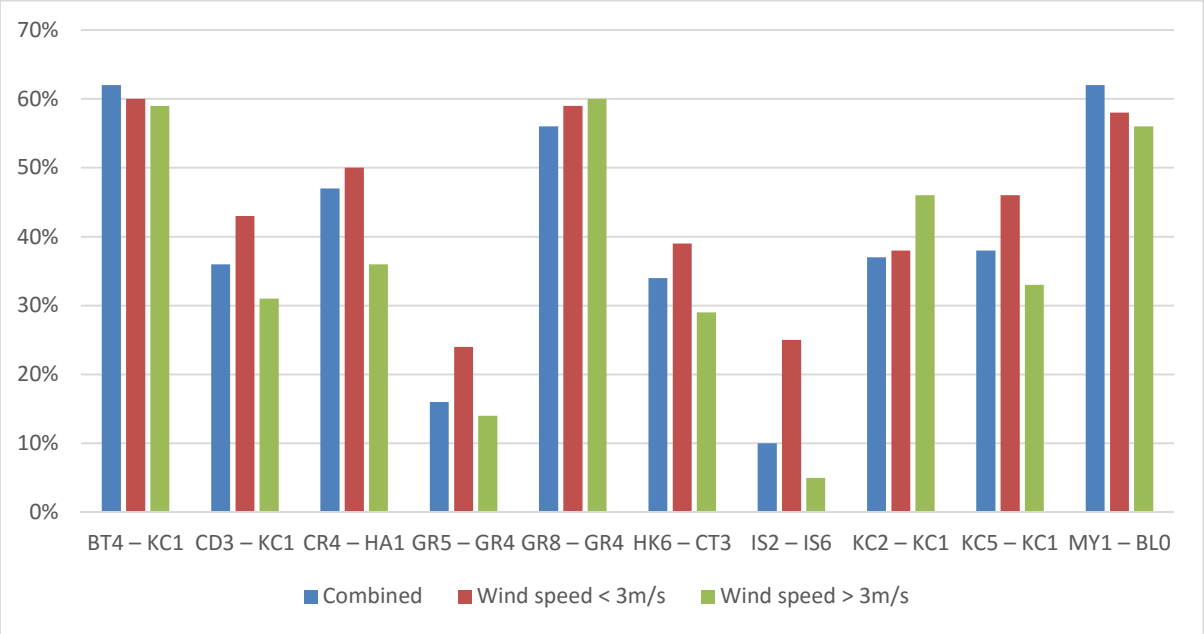


Figure 5.13 Percentage of the source contributions of PM_{10} ($\mu g/m^3$) increments by wind speed - wind direction cells (see Table 5.1 for wind directions)

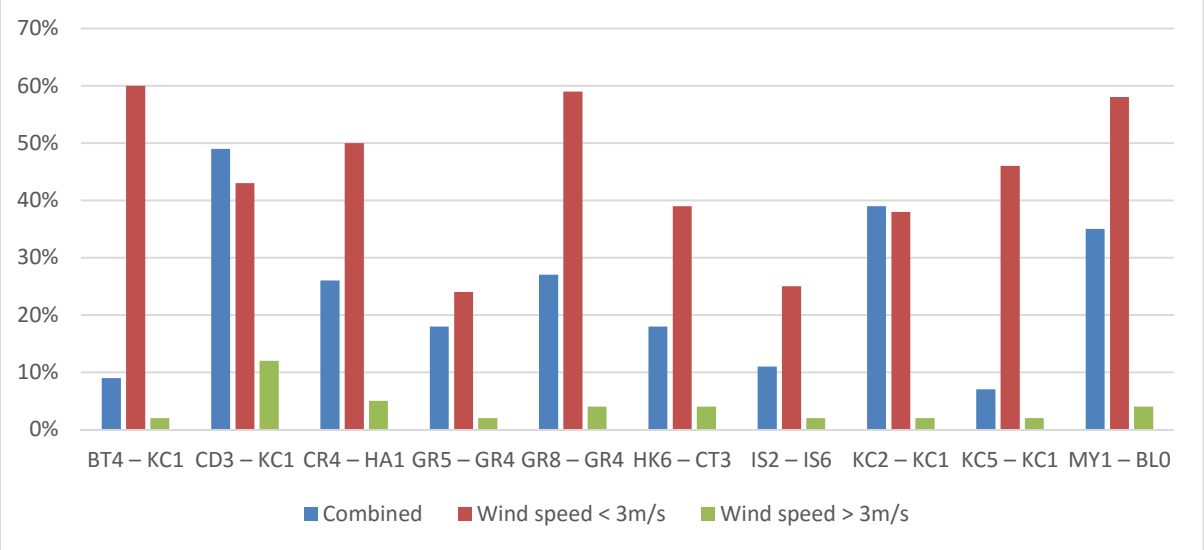


Figure 5.14 Percentage of the frequency of PM_{10} observations by wind speed - wind direction cells (see Table 5.1 for wind directions)

In the UK, the average road transport contribution to the overall PM_{10} emission is about 27% (DEFRA, 2013), however in London, road transport contributes about 46% of the total PM_{10} emission (TfL, 2014). The average upper limit of the PM_{10} increment was 15.39 $\mu g/m^3$ which is about 46% of the average PM_{10} concentrations observed at the sites.

5.6 Summary

In this chapter, the temporal and spatial relationships between road traffic and particles concentrations collected at the monitoring sites have been analysed. It has been observed that the particle concentrations follow the hourly, daily and weekly variations of the traffic closely.

Combining the findings of this chapter and Chapter 4 we concluded that road traffic contributes about 46% of the particle concentration data obtained at the monitoring stations selected. Also, traffic, meteorological variables and background concentrations of particles, and other gaseous pollutants can constitute good predictor variables for roadside particle concentrations (PM_{10}, $PM_{2.5}$, and PNC). However, the variables identified might be highly

correlated or might not be available at the intended site where the modelling is required. Moreover, using all the variables might make the application of the statistical and machine learning models somewhat expensive. Therefore, there is a need to select the most relevant to the reliability of the models. The effect of the feature selection on the accuracy of the statistical and machine learning methods will be analysed in Chapters six and seven.

Chapter 6

Statistical Air Quality Modelling and Feature Selection

6.1 Introduction

The choice of which predictor variables to include in a model is one of the difficult tasks air quality modellers encounter when dealing with statistical and machine learning methods. This issue becomes more apparent the larger the data and its complexity. Also, the computational efficiencies of computing machines are being multiplied making it easier to handle large data and more complex algorithms. A model that is built with fewer predictor variables can be more interpretable and less expensive compared with the one built with many input variables. The air quality predictor variables are often costly to measure, especially if required over long periods of time. Moreover, models tend to be less efficient when built with a large number of potentially correlated predictor variables. The immediate solution to this problem is to optimise the use of the predictor variables so that fewer variables are used without compromising the efficiency of the intended model.

Feature selection techniques are invoked for this purpose such that more easily interpretable and relatively cheaper models are obtained. Some modelling methods like ensemble regression trees have built-in mechanisms for feature selection. However, simpler methods like multiple linear regression and its variants and more sophisticated methods like artificial neural networks, support vector machines and lots more, require feature selection as part of their modelling process. In this chapter, the effect of the two feature selection methods namely: Genetic Algorithms (GA) and Simulated Annealing (SA) combined with Random Forests (RF) on some selected statistical models for predicting roadside particulate matter have been investigated. Also, the effect of using the overall roadside pollutant concentrations

or the roadside increment on the accuracy of the models was also investigated. Section 6.2 presents the results obtained from the application of the statistical methods in modelling roadside particles using the two data sets. In Section 6.3 the performance of the statistical models developed with and without the feature selections are compared. The discussion of the results obtained in Sections 6.2, 6.3 and 6.4 are presented in Section 6.5. Section 6.6 summarised the findings of the chapter.

6.2 Statistical modelling results

The five statistical methods including Stepwise Regression, Lasso regression, Elastic-net regression, Principal component regression (PCR), Partial least square regression (PLSR) and Multiple Linear Regression (MLR) were trained to predict three metrics of roadside particle concentrations (i.e. PM_{10}, $PM_{2.5,}$ and PNC) using two different sets of training data. The data used for the modelling consists of the time data (i.e. year, month, day and hour), meteorological variables, and traffic data. Others include background pollutants and roadside pollutants. The roadside increments were estimated by taking the difference between roadside concentrations and their corresponding background concentrations. Initially, the training data was divided into two major data sets. The first data set consists of all the variables except the roadside concentration increments while in the second category of the dataset, the roadside concentrations were substituted with the roadside increments. The statistical models were trained to predict both the roadside particle concentrations and their corresponding roadside increments separately. The rest of this section presents the results of the statistical modelling using these data sets. Each data set was divided into 80% training data and 20% test data using 10-fold cross validation. The statistical methods were then used to fit the training data using the 10-fold cross-validation resampling technique, and the training data was centred and scaled prior to the training and the resampling

performance was measured using Root Mean Squared Error (RMSE) and R-squared. The models were tested using the test data sets. The performances of the models were evaluated using FAC2, Mean Bias (MB), and Mean Gross Error (MGE). Others are the normalised forms of MB and MGE, RMSE, the coefficient of correlation (R), Coefficient of Efficiency (COE), and Index of Agreement (IOA).

6.2.1 Multiple Linear Regression (MLR) Results

Table 6.1 shows the training performance of MLR models for predictions of PM_{10}, $PM_{2.5}$ and PNC concentrations. The performance measures used are R-squared (R^2), and Root Mean Squared Error (RMSE). The MLR models for PM_{10} did not show much difference in the training performance when using the two training data sets. Also, the model training performance is similar when predicting either roadside or road increment PM_{10} concentrations. The average training R^2 values were 0.77, and the corresponding RMSE value was about 10μg/m^3. The performance of the models is reasonably good considering that linear regression models cannot estimate the non-linear relationships that exist between the variables involved in the model. Linear models also have difficulty in dealing with predictors that are correlated. Therefore, the performance of the models might be affected by the presence of the traffic variables which are highly correlated. There is also a strong correlation between the roadside pollutants and also between the background pollutants as discussed in Chapter 4.

.

Table 6.1 Training performance of MLR models

Note: the increment data refers to data set containing roadside pollutant increments (roadside – background) instead of roadside pollutant while roadside data refers to the data containing roadside pollutants without any modification

Pollutant	Model	Data type	Prediction	RMSE	R-squared
PM10	MLR	roadside data	Roadside prediction	9.94	0.79
PM10	MLR	roadside data	Increment prediction	10.07	0.77
PM10	MLR	increment data	Roadside prediction	10.19	0.78
PM10	MLR	increment data	Increment prediction	10.28	0.76
PM2.5	MLR	roadside data	Roadside prediction	4.96	0.87
PM2.5	MLR	roadside data	Increment prediction	4.93	0.74
PM2.5	MLR	increment data	Roadside prediction	4.91	0.87
PM2.5	MLR	increment data	Increment prediction	5.05	0.73
PNC	MLR	roadside data	Roadside prediction	11285	0.83
PNC	MLR	roadside data	Increment prediction	11208	0.83
PNC	MLR	increment data	Roadside prediction	10838	0.80
PNC	MLR	increment data	Increment prediction	11068	0.81

The training R^2 values range from 0.75 to 0.87 where the $PM_{2.5}$ models showed better performance followed by the MLR-PNC and MLR-PM_{10} models respectively. The RMSE values are relatively higher for all the models since they are required to be as low as possible with zero being the RMSE value for a perfect model. The models performed better when predicting the roadside particles than when predicting their corresponding increments regardless of which version of the data was used.

The variable importance for each model was estimated to ascertain their contribution in the models. Figures 6.1 to 6.3 show the importance of the predictor variables in the MLR models for predicting PM_{10}, $PM_{2.5}$, and PNC respectively. The variable importance for the models was estimated using the absolute values of their t-statistics shown on the vertical axes of the figures. The predictor variables used in the modelling are shown on the horizontal axes.

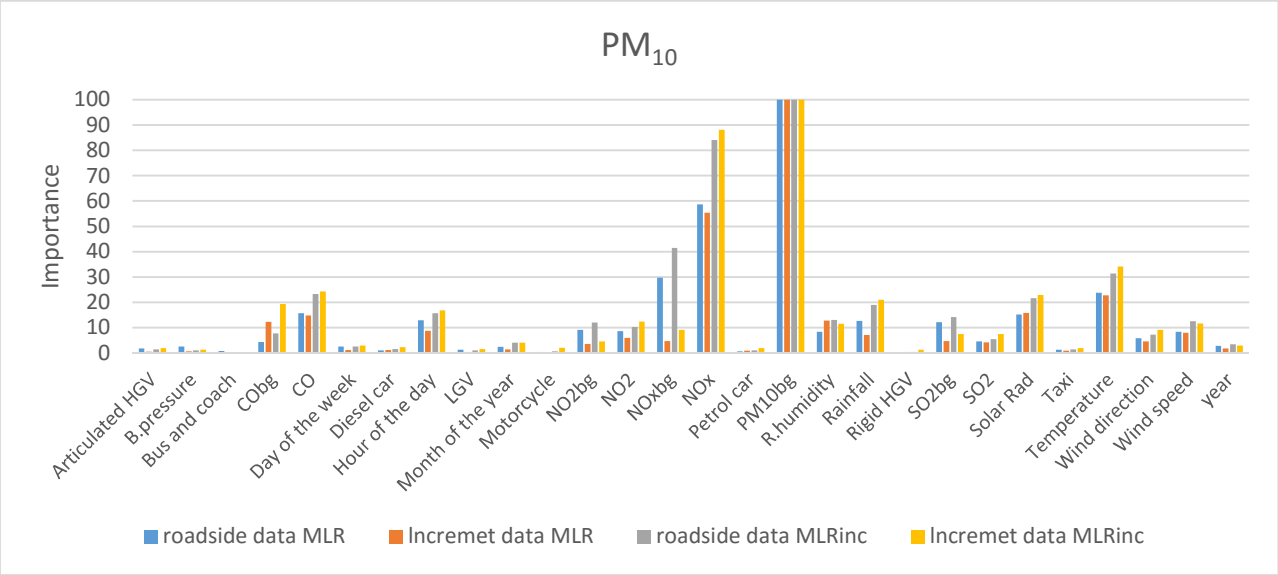


Figure 6.1 Predictor variable importance for PM_{10} ($\mu g/m^3$) models

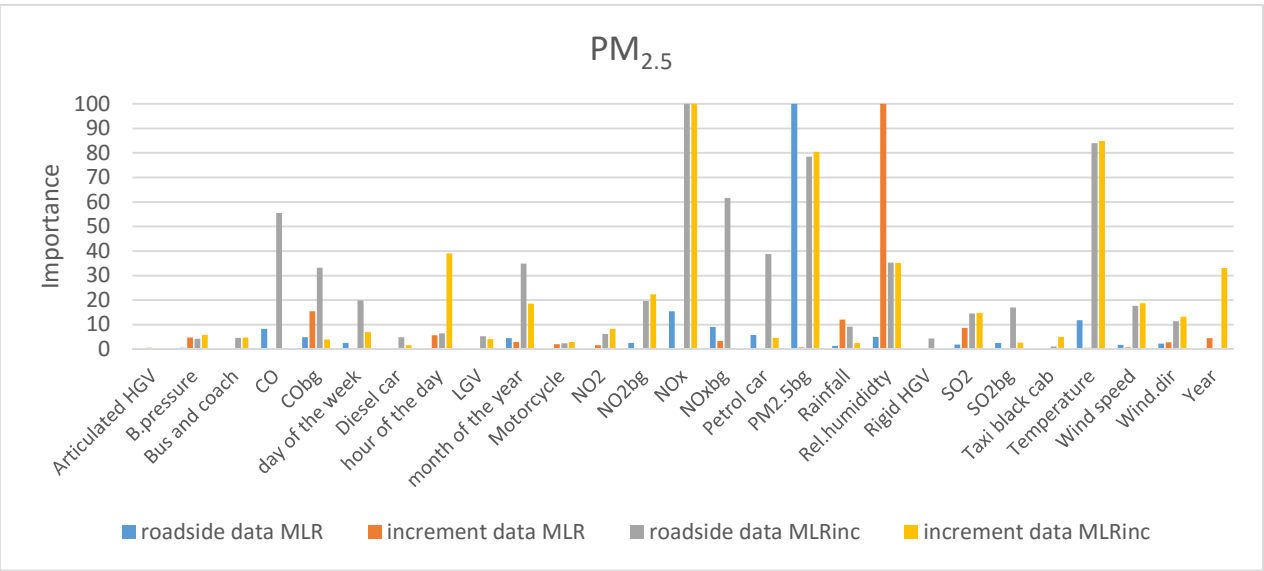


Figure 6.2 Predictor variable importance for $PM_{2.5}$ ($\mu g/m^3$) models

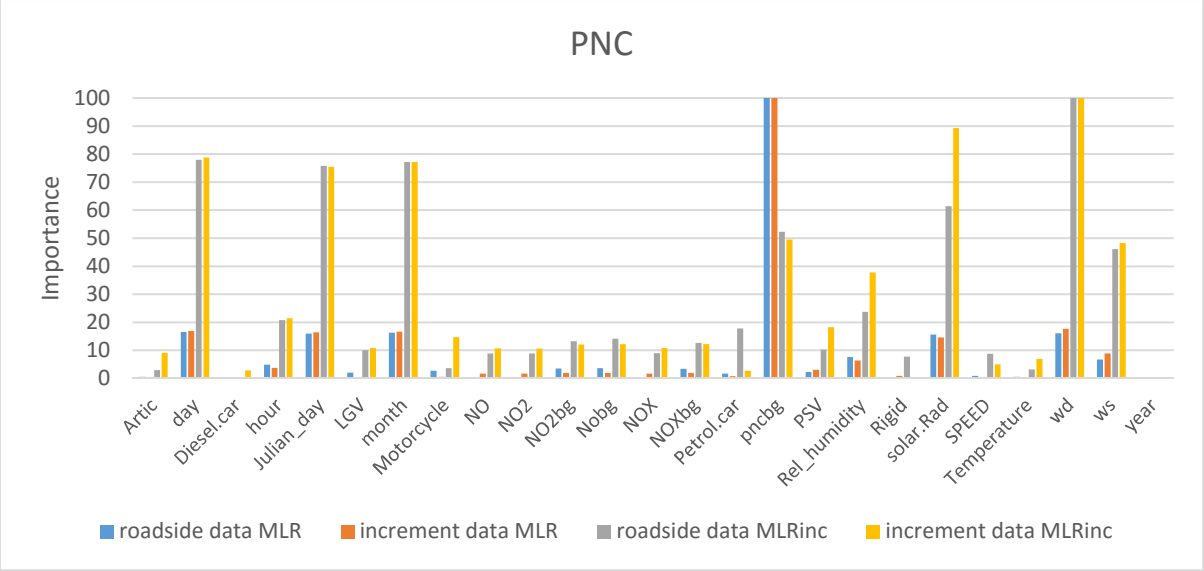


Figure 6.3 Predictor variable importance for PNC (number/cm^3) models

Note: the "inc" attached to the name of the models refers to the prediction of road increment. The values on the vertical axis represent the relative importance of the predictor variables shown on the horizontal axis. The colour codes represent the type of data used and/or variant of the response variable.

The most important variables identified by the MLR models are similar across the two datasets and the variants of the response variables in the case of MLR-PM_{10} and MLR-PNC as shown in Figures 6.1 and 6.3. However, for the variables selected by MLR-$PM_{2.5}$ models shown in Figure 6.2, it is hard to establish a pattern, and there are some differences observed with respect to the two data sets and the response variables. However, it was observed that the roadside NOx contributed more than its background concentrations in the case of MLR-PM_{10} and MLR-$PM_{2.5}$ models. Also, it was more important in predicting increment concentrations than the roadside concentrations. Overall, the most contributing variables are the roadside NOx, background particle concentrations, and meteorological variables while the traffic variables were the least contributing variables in all the MLR models. The temporal variables (time components) are more important in predicting PNC than $PM_{2.5}$ and PM_{10} concentrations.

The performance of the models on the test data shown in Table 6.2 is similar to the training performance showing that the models did not overfit the training data. Also, there was not

much difference between the performance of the models in terms of the differences in the two data sets used. However, the models performed better when predicting roadside particles than their corresponding roadside increments as indicated by the statistical performance metric shown in Table 6.2.

Table 6.2 The test performance of the MLR models

Pollutant	Data type	Type of prediction	FAC2	MB	MGE	NMB	NMGE	RMSE	R	COE	IOA
PM10	Roadside data	Roadside prediction	0.99	-0.20	6.68	0.00	0.15	10.76	0.87	0.61	0.81
PM10	Roadside data	Increment prediction	0.76	-0.15	6.68	-0.01	0.32	10.74	0.86	0.60	0.80
PM10	Increment data	Roadside prediction	0.99	-0.06	6.61	0.00	0.15	9.76	0.89	0.61	0.81
PM10	Increment data	Increment prediction	0.75	0.00	6.61	0.00	0.31	9.75	0.88	0.60	0.80
PM2.5	Roadside data	Roadside prediction	0.97	0.04	3.49	0.00	0.15	5.05	0.93	0.66	0.83
PM2.5	Roadside data	Increment prediction	0.63	0.03	3.49	0.00	0.57	5.04	0.86	0.55	0.78
PM2.5	Increment data	Roadside prediction	0.97	-0.04	3.49	0.00	0.15	5.24	0.93	0.67	0.83
PM2.5	Increment data	Increment prediction	0.62	-0.03	3.49	0.00	0.57	5.23	0.85	0.56	0.78
PNC	Roadside data	Roadside prediction	0.95	185.89	5761.36	0.01	0.18	10556.95	0.92	0.70	0.85
PNC	Roadside data	Increment prediction	0.95	-220.79	5916.16	-0.01	0.18	11084.84	0.92	0.70	0.85
PNC	Increment data	Roadside prediction	0.77	172.58	5691.15	0.01	0.28	10488.58	0.90	0.67	0.83
PNC	Increment data	Increment prediction	0.78	-181.11	5913.98	-0.01	0.28	11070.22	0.90	0.66	0.83

6.2.2 Principal Component Regression (PCR) Results

The PCR method is designed to overcome the issue of correlated predictors discussed in section 6.3.1. The method uses principal components analysis to transform the variable into new sets of uncorrelated predictors. The new variables are then used to fit the response variable. The PCR was applied to the same data used for MLR models and was trained to

predict both roadside and roadside increment particle concentrations. Table D.1 in Appendix D shows the training performance of the PCR models developed. The number of variables used was between 24 and 25 but the PCR models selected between 16 and 20 components for the models. However, this reduction does not translate into the reduction of the original variables because the components contain the linear combination of the original variables. The PNC models selected the lowest number of components than the PCR-PM_{10} and PCR-$PM_{2.5}$ models.

The importance of each predictor variable used was estimated for each PCR model, and the results are shown in Appendix D - Figures D.1 to D.3. The most important variables selected by all the models are the roadside gaseous pollutants and the background particle concentrations. The traffic variables are also shown to be more important than most of the meteorological variables except wind direction and wind speed where they have equal or more importance in the models. The temporal variables are the least contributing variables. The PCR-PNC and the PCR-PM_{10} models give more weight to oxides of nitrogen and traffic variables than the PCR-$PM_{2.5}$ models. The PCR models consider traffic variables as more important than the temporal variables where the reverse is the case in MLR models. The test performance of the PCR models is shown in Table D.2, and it was observed that the performance of the models is very much similar to that of the MLR models.

6.2.3 Partial Least Square Regression (PLSR) Results

The PLSR method is a supervised version of the PCR method. It estimates its components while taking into account the effect of the predictor variables on the response variable. In reality, the PLSR estimate search for the components that maximally summarises the variation in the feature space while at the same time having a maximum correlation with the response variable. The results of the training performance for the PLSR models are shown

in Table D.3, and their performance is similar to the performance of MLR and the PCR models with a slightly lower number of components and RMSE values for $PM_{2.5}$ models.

The importance of the predictor variables to the PLSR models is shown in Figures D.4 to D.6. The most important variables selected are similar to those nominated by the PCR models. However, the background concentrations, the traffic variables, and the meteorological variables are shown to have more weight in the PLSR models than in the PCR models, especially for PM_{10} and PNC. The test performance of the PLSR models is shown in Table D.4. Their performances are almost identical to the performance of the PCR models.

6.2.4 Stepwise Regression Results

The stepwise regression model was trained to predict the PM_{10}, $PM_{2.5}$ and PNC concentrations using the same training data explained in the previous sections. The training performance of the stepwise regression models is shown in Table D.5. The performance is similar to the MLR, PCR and PLSR models for the prediction of PM_{10}, $PM_{2.5}$ and PNC concentrations discussed in Sections 6.2.1 - 6.2.3. However, the benefit is that the stepwise regression might have used fewer predictor variables than the previous methods. The models mostly selected between 23 and 24 variables from the total of 28 variables. The variable importance for each model was estimated, and the results are shown in Figures D .7 to D.9.

The most important variables indicated by the stepwise regression models are roadside pollutants followed by wind direction and wind speed. The remaining variables have somewhat similar importance. The main difference between the important variables selected by the stepwise method and the MLR is that the stepwise method gives more weight to the traffic variables. Moreover, it also gives less weight to the background particles except for $PM_{2.5}$ models where the background $PM_{2.5}$ has the highest weight.

6.2.5 Elastic-net Regression Results

The elastic-net regression is a penalised regression method that is formulated to take advantage of Ridge and Lasso regression methods (Zou and Hastie, 2005). Ridge regression uses L_2 penalty to shrink the parameter estimates of the ordinary least square regression towards zero to minimise the sum of squared errors. The penalty introduces a bias-variance trade-off to improve the efficiency of the parameter estimates and maximise the use of correlated predictors. Lasso, on the other hand, uses L_1 penalty to force some of the parameters to absolute zero which cannot be obtained using Ridge regression. Therefore, Lasso can perform feature selection as well as improving the performance of the model. One shortcoming of Lasso is that it does not consider all the correlated predictors as does the Ridge regression (Hong and Chen, 2015). It often selects one and dumps the rest. The elastic-net regression uses both L2 and the L1 penalties to improve the parameter estimates by dealing with correlated predictors the ridge regression way and also conducts feature selection as it is obtained using Lasso regression. In this work elastic-net regression was applied to the training data to produce models for the prediction of PM_{10}, $PM_{2.5}$ and PNC concentrations. The model parameters were tuned using 10-fold cross-validation to obtain the optimum combination of the parameters that will result in the model with high performance. The performance of the models during training was measured using R^2 and RMSE as shown in Table D.7.

The training performance of the elastic-net models is similar to that of the MLR, PCR, and PLSR despite the removal of some of the predictors from the set of predictor variables (see Table D.9). The models with zero lambdas are reduced to lasso regressions. The cross-validation profiles of the elastic net models are shown in Figures 6.4 to 6.6. It could be seen that the lower lambda values yielded much higher performance than the larger values. Also,

the PNC models produced the same mixing parameters with all the lambda values. This shows that the range of the lambda values does not have an effect on the accuracy of the models. The reason for this has not been determined. Moreover, this behaviour did not affect the performance of the models since they produced models with similar performance as in the case of MLR, PCR, and PLSR as do the models for PM_{10} and $PM_{2.5}$.

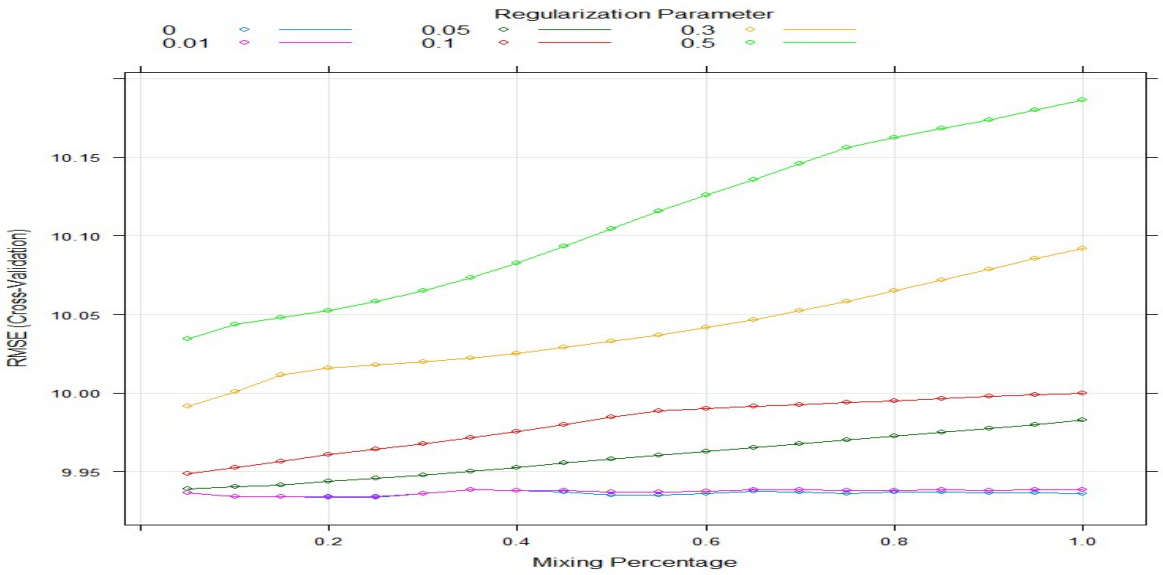


Figure 6.4 The cross-validation profile for the PM_{10} elastic –net model

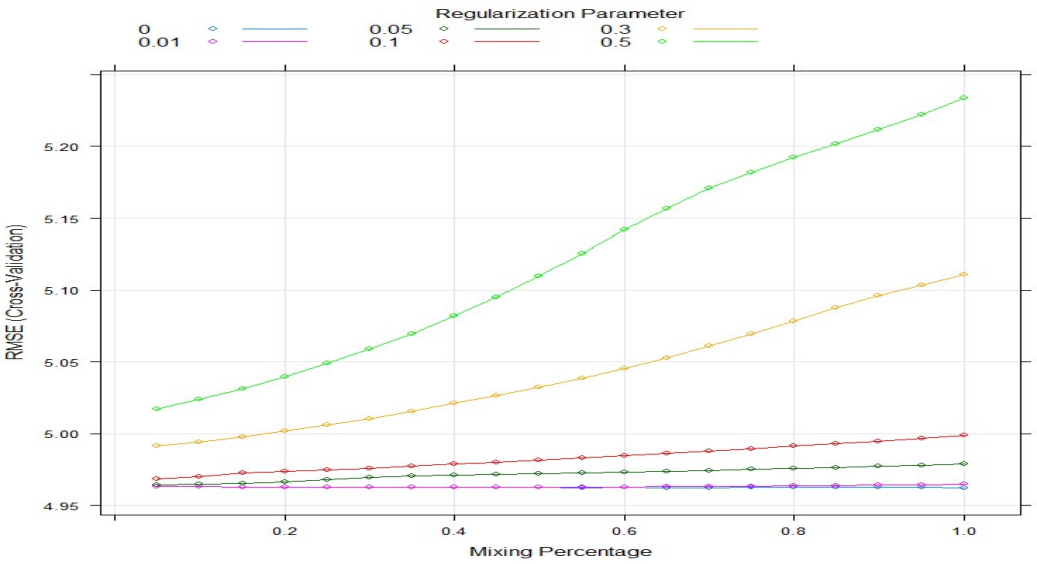


Figure 6.5 The cross-validation profile for the $PM_{2.5}$ elastic-net model

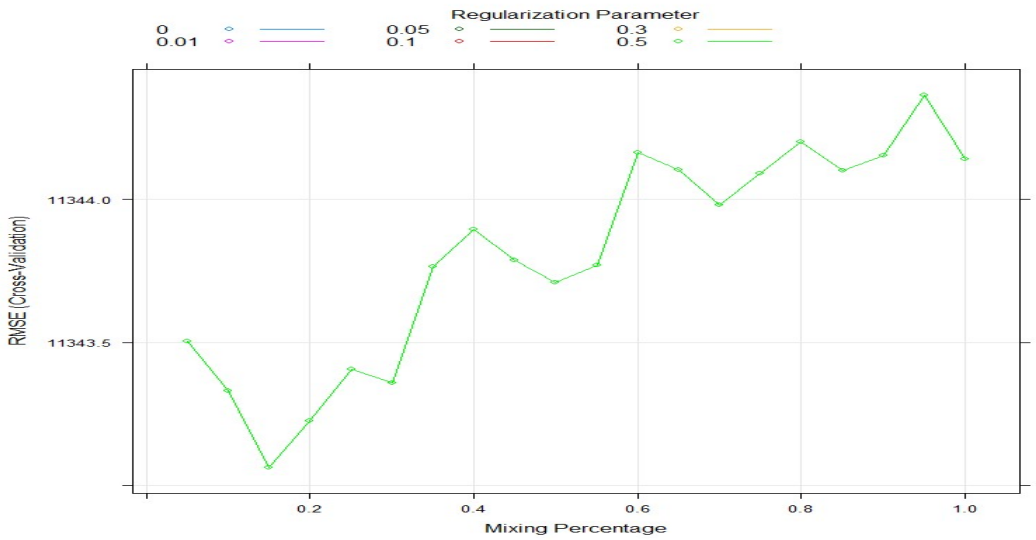


Figure 6.6 The cross-validation profile for the PNC elastic-net model

Figures 6.4 to 6.6 show the relative importance of the predictor variables to the elastic-net models for predicting PM_{10}, $PM_{2.5}$ and PNC respectively. The figures show that the most important variables selected by the elastic-net models are roadside and background NOx, and background particle concentrations. The rest of the variables have nearly the same performance. This behaviour is quite different from the other models discussed in the previous sections where the roadside pollutants were shown to be the most contributing variables, and the remaining variables have distinct contributions in most cases.

Table D.7 show that the elastic-net models performed in the same way as the previously discussed models. However, the benefit of using this method is its ability to select smaller numbers of predictor variables than the MLR, PCR, and PLSR. It could be seen that the models dump some of the variables while the coefficients of many variables are nearly zero (see Table D.7). The test performance of the elastic net regression models is shown in Table D.8. The results showed similar performance to the MLR, PCR, PLSR and the Stepwise models despite the reduction in the number of variables used by the models.

6.3 Application of Hybrid Feature Selection to the Statistical Models

This section presents the results of the two-hybrid feature selection methods discussed in section 6.2. The test performance of the statistical models developed using the predictor variables selected by the models is compared and presented. The Genetic Algorithms combine with Random Forests (GA-RF) and Simulated Annealing combined with Random Forests (SA-RF) were applied to the samples drawn from the training data for predicting PM_{10}, $PM_{2.5,}$ and PNC. The data for the selected variables were then used in the training of the same statistical models discussed in Section 6.2. The external and internal performance of the feature selection for PM_{10}, $PM_{2.5}$ and PNC are shown in Figures 6.7 to 6.9.

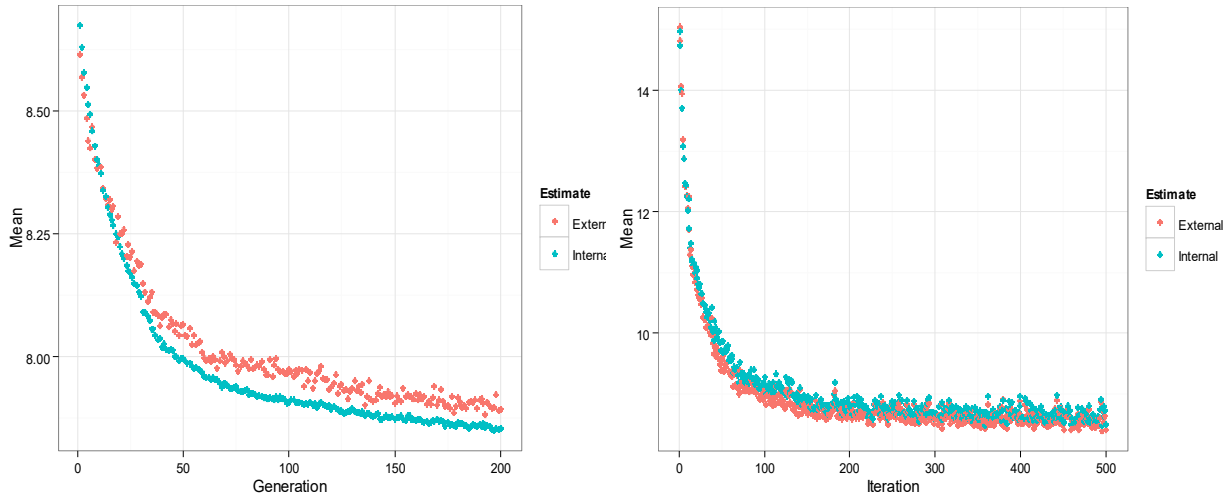


Figure 6.7 External and internal performances of GA-RF and SA-RF feature selection for PM_{10} ($\mu g/m^3$).

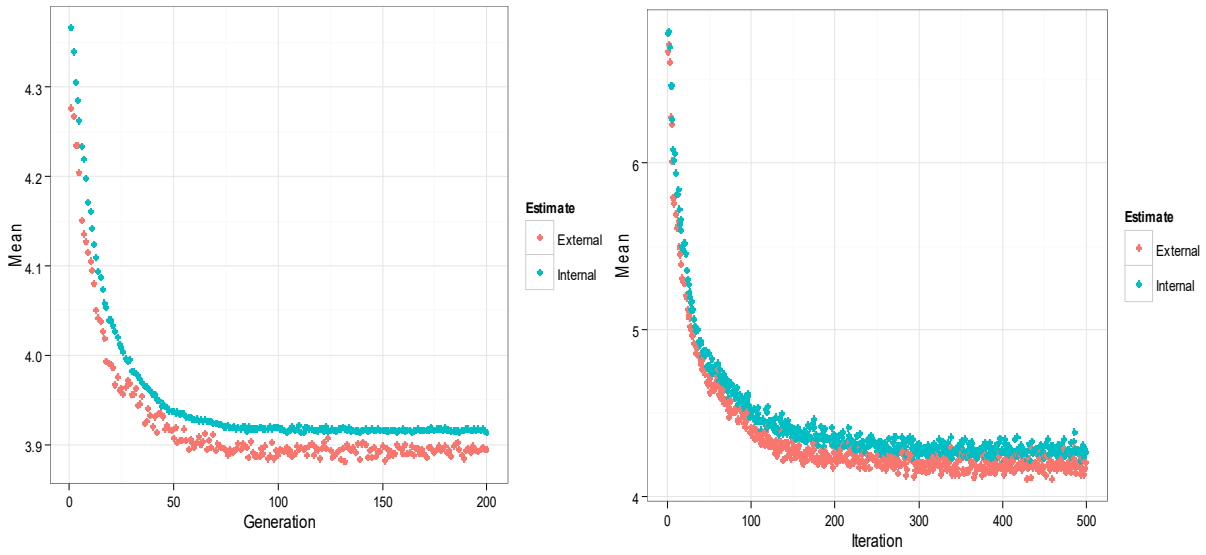


Figure 6.8 External and internal performances of GA-RF and SA-RF feature selection for $PM_{2.5}$ ($\mu g/m^3$)

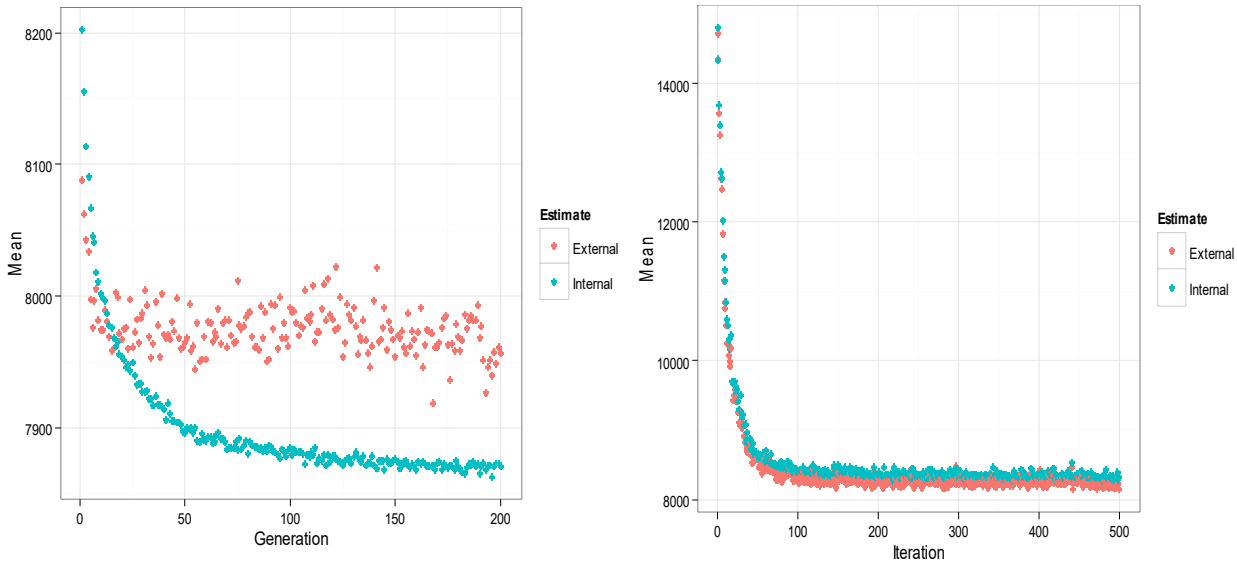


Figure 6.9 External and internal performances of GA-RF and SA-RF feature selection for PNC(number/cm^3)

For the PM_{10} data, the optimum number generation for the GA-RF was found to be 193 out of the maximum of 200 generations specified with 50 populations per generation. The crossover probability was selected to be 0.8, and the mutation probability was 0.1. The

external RMSE and R-squared were estimated to be 7.8809µg/m3 and 0.8916 respectively. Twelve variables were selected out of the 28 possible variables in the training data. For the SA-RF, the maximum number of iterations specified was 500, and the optimum number of iteration was 495, and the external RMSE and R-squared were 8.3972µg/m3 and 0.8757 respectively. The SA-RF selected ten variables out of the 28 variables.

Using the $PM_{2.5}$ training data, the optimum number generation for the GA-RF was found to be 132 out of the maximum of 200 generations specified with 50 populations per generation. The crossover probability was 0.8, and the mutation probability was 0.1. The external RMSE and R-squared were estimated to be 3.8812 µg/m3 and 0.8696 respectively. Nine variables were selected out of the 27 possible variables in the training data. For the SA-RF, the maximum number of iterations specified was 500, and the optimum number of iterations was 459 while the external RMSE and R-squared were 4.1039 µg/m3 and 0.8566 respectively. The SA-RF selected 16 variables out of the 27 variables.

The optimum number of generations when the GA-RF was applied to the PNC training data was found to be 168 out of the maximum of 200 generations specified with 50 populations per generation. The crossover probability was selected to be 0.8, and the mutation probability was 0.1. The external RMSE and R-squared were estimated to be and 7918.4297number/cm3 and 0.9399 respectively. Thirteen variables were selected out of the 25 possible variables in the training data. For the SA-RF, the maximum number of iterations specified was 500, and the optimum number of iterations was 472 while the external RMSE and R-squared were 8153 number/cm3 and 0.93 respectively. The SA-RF selected 13 variables out of the 25 variables.

The out of bag RMSE and R^2 were used as measures of the internal performance while the 10-fold cross-validation repeated five times was the resampling methods used to estimate

the RMSE and R^2 for the external performances. The two performances follow the same pattern. However, the GA-RF shows that the external performance is slightly higher than the internal performance in the case of $PM_{2.5,}$ while for the PM_{10} it is slightly lower. For the PNC data, the external performance is significantly lower than the internal performance. The trend shown in the cases of PM_{10} and PNC is expected since the internal performance procedure has some chance of overfitting the data. For the SA-RF method, the performances are nearly the same for PM_{10} while the external performance is slightly better than the internal in the case of $PM_{2.5}$ and PNC data. The external performance of the GA-RF for PNC is poorer than in the cases of PM_{10} and $PM_{2.5}$. This behaviour might be due to the amount of data used for the training since the PNC data used was only for one year. Table 6.3 show the variables selected by the hybrid feature selection methods for the PM_{10}, $PM_{2.5,}$ and PNC models.

Table 6.3 Variables selected by hybrid feature selection methods

Pollutant	**PM10**		**PM2.5**		**PNC**	
Feature Selection Methods	*GA-Random Forests*	*SA-Random Forests*	*GA-Random Forests*	*SA-Random Forests*	*GA-Random Forests*	*SA-Random Forests*
Articulated HGV						
B. pressure	✓	✓	✓	✓		
Bus and coach						
CO						
CO.bg				✓		
Day of the week	✓	✓	✓	✓	✓	✓
Diesel car				✓		
Hour of the day	✓			✓	✓	✓
Julian day	-	-	-	-	✓	✓
LGV						
Month of the year	✓	✓	✓	✓		
Motorcycle				✓		✓
NO					✓	✓
NO.bg					✓	
NO2						
NO2bg			✓		✓	✓
NOx	✓	✓	✓	✓	✓	✓
NOx.bg						
Petrol car						✓

Table 6.3 *continued*

Pollutant	**Pollutant**	**Pollutant**	**Pollutant**	**Pollutant**	**Pollutant**	**Pollutant**
Feature Selection Methods	*GA-Random Forests*	*SA-Random Forests*	*GA-Random Forests*	*SA-Random Forests*	*GA-Random Forests*	*SA-Random Forests*
PM.bg	✓	✓	✓	✓	✓	✓
R. humidity	✓		✓	✓	✓	✓
Rainfall	✓	✓		✓		
Rigid HGV				✓		
SO2						
SO2bg	✓			✓		
Solar Rad	✓	✓			✓	
Taxi						
Temperature	✓	✓	✓	✓	✓	✓
Wind direction	✓	✓			✓	✓
Wind speed			✓		✓	✓
Year				✓		✓

The methods selected the variables that are nearly the same especially for predicting PM_{10} and PNC while for the $PM_{2.5}$ models, the SA-RF selected 16 variables and the GA-RF selected only ten variables. The only difference in the variables selected for the PM_{10} is the hour of the day and roadside SO_2 selected by the GA-RF. For the $PM_{2.5,}$ the SA-RF selected background CO, Hour of the day, Motorcycle, Rainfall, Rigid HGV and the Year in addition to the variables selected by the GA-RF. The methods differ in the selection of Motorcycle, Petrol car and year by the SA-RF and Solar radiation by the GA-RF for PNC models (see Table 6.3). The general pattern in their selection is that they have eliminated most of the

correlated variables especially the background pollutants and the traffic variables. Whereas the temporal variables and the meteorological variables are selected for all the PM metrics. These variables were less significant for the linear models as discussed in section 6.2. Their inclusion here suggests that they might have a non-linear relationship with the predictor variables or their correlation with other predictors make it impossible for the linear models to discover their true relationships with the response variables. The NO was only selected for PNC models while NOx was selected in all the cases considered. The results of the test performance of the linear models developed using the feature selection methods are shown in Tables 6.4 and Tables E.1 – E.2. in Appendix E.

6.4 Comparison of the Performance of Feature Selection Methods

Tables 6.4 and Tables E.1 – E.2. in Appendix E show the comparison of the feature selection methods for PM_{10}, $PM_{2.5}$ and PNC models respectively. The performance of the models developed with the selected variables is similar to those models developed using the entire predictor variables. The differences in the performances are quite small. However, in some instances, the models with the selected variables have lower RMSE and MB values.The actual benefit derived from this exercise is the successful reduction in the number of predictor variables by more than half in most of the cases considered. The reduction in the number of variables will eventually result in the reduction of the operational and computational cost of the models without possibly compromising the predictive performance of the models.

Table 6.4 Comparison of the performance of feature selection methods for PM_{10} models

Row Labels	FAC2.	MB.	MGE.	NMB.	NMGE.	RMSE.	R.	COE.	IOA.
ENET									
GA-RF	0.99	-0.04	7.12	0.00	0.16	10.99	0.86	0.58	0.79
Linear	0.99	-0.21	6.68	0.00	0.15	10.76	0.87	0.61	0.81
SA-RF	0.99	-0.14	6.93	0.00	0.16	10.04	0.88	0.60	0.80
MLR									
GA-RF	0.99	-0.04	7.12	0.00	0.16	10.98	0.86	0.59	0.79
Linear	0.99	-0.20	6.68	0.00	0.15	10.76	0.87	0.61	0.81
SA-RF	0.99	-0.14	6.92	0.00	0.16	10.04	0.88	0.60	0.80
PCR									
GA-RF	0.98	-0.06	7.44	0.00	0.17	11.31	0.86	0.57	0.78
Linear	0.99	-0.23	6.89	-0.01	0.16	10.98	0.87	0.60	0.80
SA-RF	0.99	-0.16	7.33	0.00	0.17	10.44	0.87	0.57	0.79
PLSR									
GA-RF	0.99	-0.04	7.12	0.00	0.16	10.98	0.86	0.59	0.79
Linear	0.99	-0.21	6.69	0.00	0.15	10.76	0.87	0.61	0.81
SA-RF	0.99	-0.14	6.92	0.00	0.16	10.04	0.88	0.60	0.80
STEPWISE.REG									
GA-RF	0.99	-0.04	7.12	0.00	0.16	10.98	0.86	0.59	0.79
Linear	0.99	-0.20	6.68	0.00	0.15	10.76	0.87	0.61	0.81
SA-RF	0.99	-0.14	6.92	0.00	0.16	10.04	0.88	0.60	0.80

The performance of the models was further compared using conditional quantile plots and scatter plots shown in Figures 6.10 - 6.11 and Figures E.1- E.4 in Appendix E. The plots revealed that the PM_{10} models developed using GA-RF selected variables performed better than those developed with the SA-RF selected variables and slightly better than the linear model developed using the whole predictor variables. This performance is completely masked in the statistical performance measures. Considering Figure 6.10, the models with GA-RF variables have slightly outperformed the linear models in capturing the higher concentrations while the models with SA-RF variables failed to capture concentrations beyond 100µg/m^3 accurately. The scatter plots shown in Figure 6.11 revealed that the predictions of the linear models developed with GA-RF and SA-RF selected variables are likely to perform better in the predictions than the linear models. Because the prediction of

the linear models has more points outside the FAC2 boundaries. This feature is not highlighted by the FAC2 values of the corresponding models as they have the same values. The conditional quantile plots in Figure E.1 show that the models developed with GA-RF captured the extreme $PM_{2.5}$ values slightly better than the other two models. The same feature is also reflected in Figure E.2 and also it has fewer predictions outside the FAC2 boundaries. The PNC models developed with the features selected by the GA-RF and SA-RF have shown more data coverage than the linear model (see Figure E.3). However, the higher concentrations were poorly predicted by the models. The SA-RF linear models performed poorer than the GA-RF linear models in that respect (see Figure E.4). The scatter plots show that the SA-RF linear models have more of its prediction outside the FAC2 boundaries than the remaining two model types.

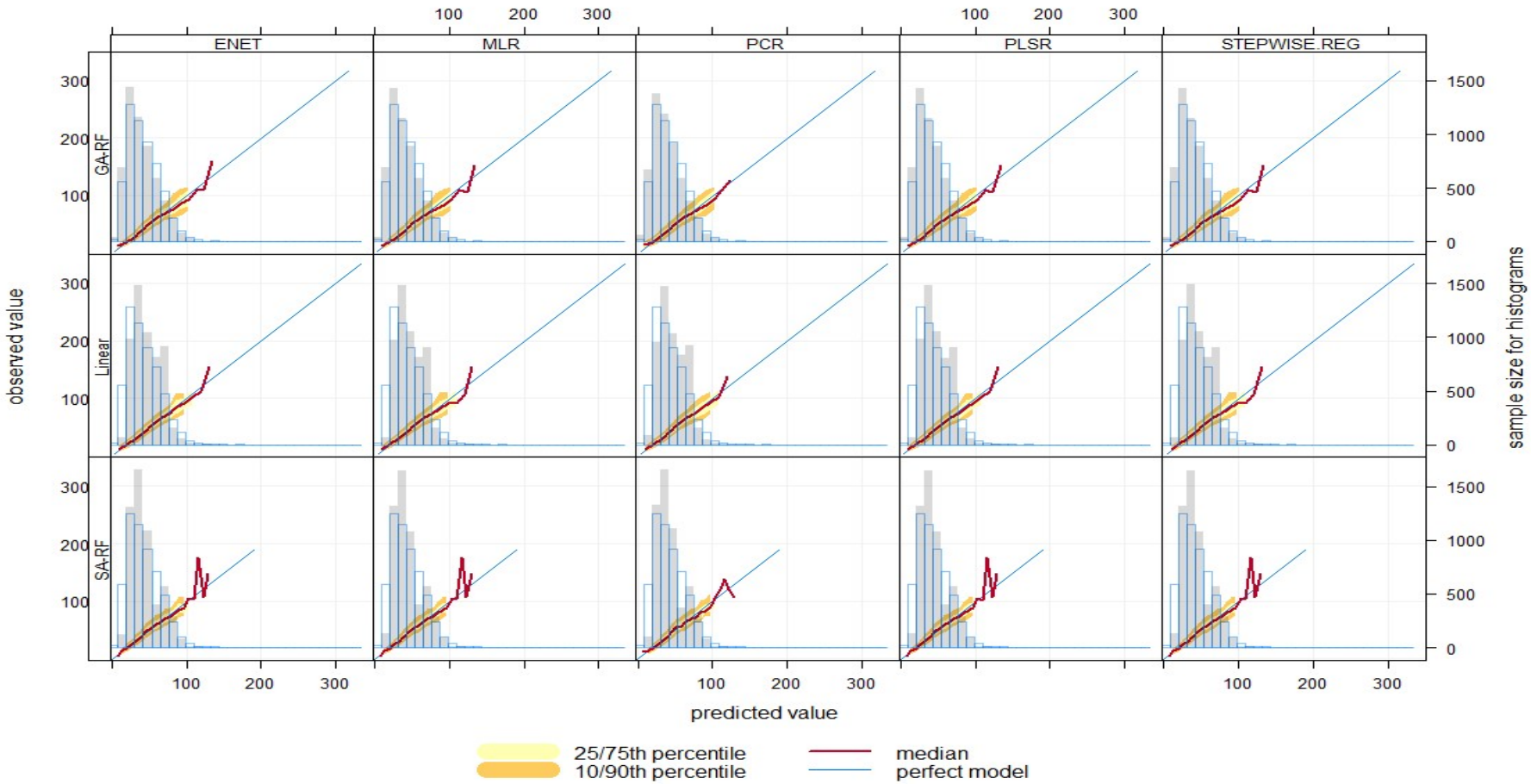


Figure 6.10 Conditional Quantile plots comparing the performance of PM_{10} ($\mu g/m^3$) models

Note: predicted value and observed value are modelled and observed PM_{10} concentrations respectively

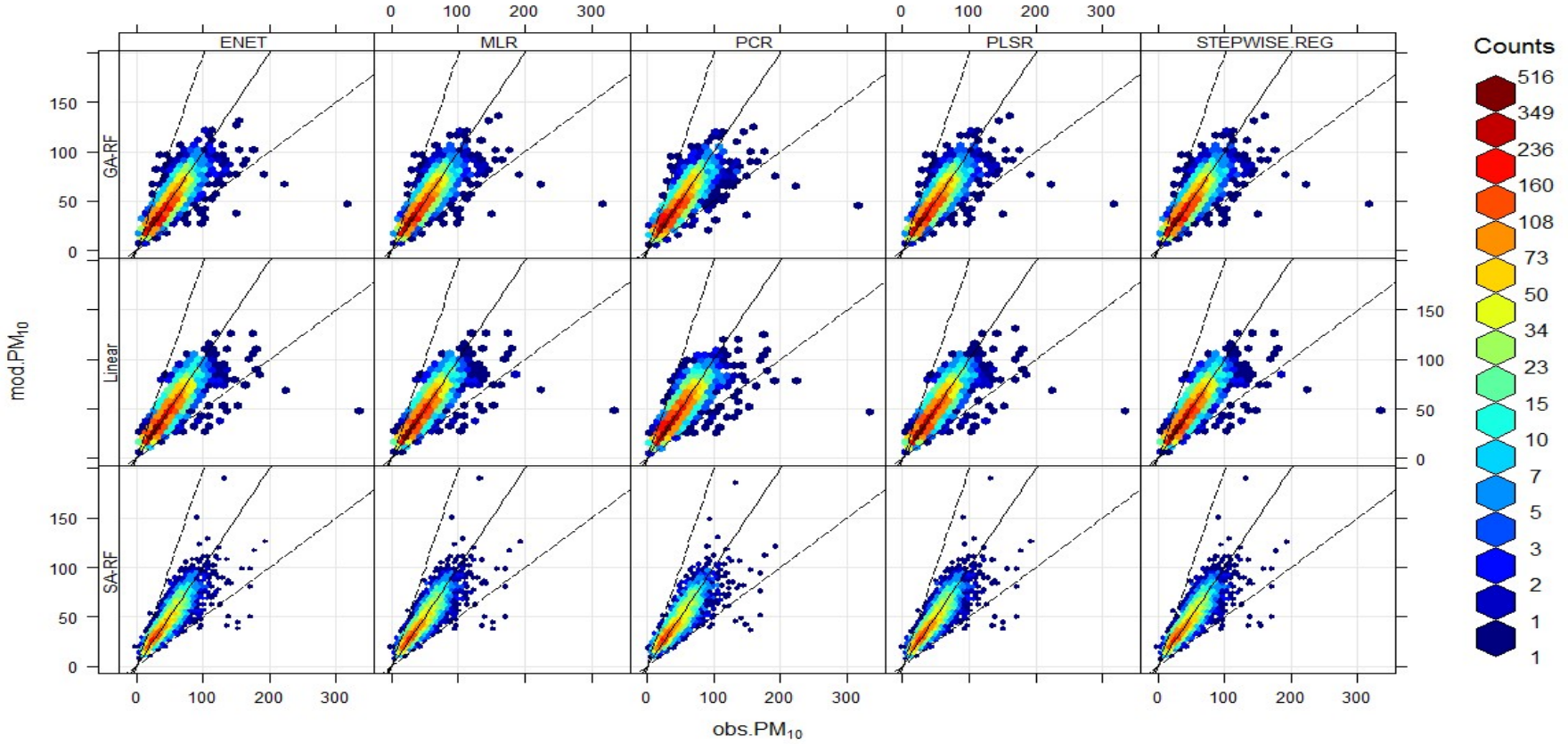


Figure 6.11 Scatter plots comparing the performance of PM_{10} ($\mu g/m^3$) models

Note: $modPM_{10}$ and $obsPM_{10}$ are modelled and observed PM_{10} concentrations respectively

6.5 Using GA-RF and SA-RF Feature Selection Methods with the Statistical Methods

This section presents a discussion of the statistical modelling results presented in section 6.2 and feature selection in section 6.3. In section 6.2, the results obtained for the training of five linear regression models using two separate datasets and two variants of the response variables for each particle concentration metric was presented. The data sets consist of emission rates for the eight vehicle traffic composition, meteorological variables, and pollutant concentrations. The pollutant concentrations consist of background and roadside concentrations. The difference in the two datasets was in the consideration of these concentrations. The first data set consists of all the variables mentioned above, while, in the second data set, the roadside concentrations were replaced with their corresponding roadside increment. The roadside increment was obtained by subtracting the background concentrations from the roadside concentrations. Each linear regression method was applied to the two data sets and was trained to predict particle concentrations (i.e. PM_{10}, $PM_{2.5}$ PNC) and their corresponding increments separately. In section 6.3, the results of the investigation on the effect of using two hybrid feature selection techniques on the efficiency of the linear regression models were presented. The linear models were selected in consideration of the differences in their formulations aimed at reducing the shortcomings of ordinary least square regression.

The results obtained show that there was not much difference in the performance of the linear models regarding the use of the two data sets. However, there was a difference in the performance of the models when predicting either the roadside particles or their corresponding roadside increments. The models performed much better in predicting the roadside concentrations than in predicting roadside increment. This difference could be

attributed to the fact that most of the urban background sites are not free from interference from local emission sources such as traffic and house heating. Therefore, the background concentrations may not necessarily represent the actual background conditions of the areas under consideration. Another problem observed during the modelling is that the difference between the two concentrations sometimes gives negative concentrations meaning that the background concentrations at those times were higher than the roadside concentrations. This condition might introduce noise into the data, and the models will try to model the noise instead of the actual relationship that might result in overfitting the data.

The average training performance of all the models across the methods was nearly the same. The R^2 values for PM_{10} models were between 0.75 and 0.79, and the corresponding RMSE values were between 9.9 and 10.6μg/m^3. The R^2 values for $PM_{2.5}$ models were between 0.73 and 0.87, and the corresponding RMSE values were between 4.9 and 5.06μg/m^3. The PNC has the R^2 values ranging from 0.8 to 0.83 with RMSE values ranging from 10826 to 11222 number/cm^3. These are good performances considering the shortcomings of the linear models since there might exist non-linear relationships between the predictor variables and the response variables. In addition, some of the variables are highly correlated. For example, the traffic variables are highly correlated with themselves, and there is also a strong correlation between the roadside pollutants and also between the background pollutants. The correlations reduce the interpretability and efficiency of the linear models.

The predictor variable importance associated with each of the models was estimated, and the four categories of each model obtained give preference to similar variables irrespective of the type of data or variant of the response variables. However, there were differences in the magnitude of the contribution of the variables across the four categories of each model. It has been observed that in some cases where the response variable is a roadside increment, the background pollutants become more important. The $PM_{2.5}$ models showed a slight

variation in the variable importance across different data sets and the different response variables. However, it was difficult to establish any pattern. The most contributing variables identified by all the models are the roadside oxides of nitrogen and background particle concentrations. The meteorological variables, background pollutants, and traffic variables are the second most contributing variables and have similar contributions in all the models except in MLR. The temporal variables are more important in predicting PNC than $PM_{2.5}$ and PM_{10} concentrations.

The performance of models on the test data is similar to the training performance showing that the models did not overfit the training data. The test performance of the models also shows not much difference between the performances of the models in terms of the differences in the two data sets. The models performed better in predicting roadside particles than their corresponding roadside increments. Also, they predicted PNC slightly more accurately than $PM_{2.5}$ and PM_{10}.

The feature selection methods discussed in section 6.3 were applied to the training samples with the aim of reducing the number of predictor variables to be used in the statistical modelling. The GA-RF selected 12 variables while the SA-RF selected ten variables out of the 28 predictor variables in the PM_{10} training data. In the case of $PM_{2.5}$ training data, GA-RF and SA-RF selected 9 and 16 variables respectively out of the 27 predictor variables. Thirteen variables were selected by the GA-RF out of the 25 possible variables in the PNC training data while the SA-RF selected 13 variables. The methods selected variables that are nearly the same especially for predicting PM_{10} and PNC while for the $PM_{2.5}$ models, the SA-RF selected 16 variables while the GA-RF selected only ten variables. The general pattern in their selection is that they have eliminated most of the correlated variables especially the background pollutants and the traffic variables. Whereas the temporal variables and the meteorological variable have been selected in all the cases considered. These variables were

shown to be less significant for the linear models as discussed in section 6.2. Their inclusion here, suggests that they might have non-linear relationships with the response variables or their correlation with the other predictors make it impossible for the linear models to discover their true relationships with the response variables.

The results of the statistical performance of the linear models developed using the variables selected by the feature selection methods are similar to those developed using the entire predictor variables. The differences in the performances are quite small. However in some instances, the models with the selected variables have lower RMSE values, and mean biases but these differences are so small to be considered as better performance than the previously developed linear models. The actual benefit derived from this exercise is the successful reduction in the number of predictor variables by more than half in most of the cases considered. The reduction in the number of variables will eventually result in the reduction of the operational and computational cost of the models without compromising the predictive performance of the models.

To explore more about the differences in the performance of the models, conditional quantile and scatter plots were used. The plots revealed that the PM_{10} models developed using GA-RF selected variables performed better than those developed with the SA-RF selected variables and slightly better than the linear model developed using the predictor variables. This performance is completely masked in the statistical performance measures. The scatter plots revealed that the predictions of the linear models developed with GA-RF and SA-RF selected variables are likely to perform better in practice than the linear model that have more points outside the FAC2 boundaries. This feature is not highlighted by the FAC2 values of the corresponding models as they have the same values. The plots show that the models developed with GA-RF try to estimate the extreme values of $PM_{2.5}$ concentrations more than the other two models. The PNC models developed with the features selected by

the GA-RF and SA-RF have shown more data coverage than the linear model. However, the higher concentrations were poorly predicted by the models.

6.6 Comparison of the Performance of Statistical methods with Other Studies

The performance of the statistical methods (MLR, PCR, PLSR,Stepwise regression and elastic net/Lasso) used in this study is similar. The R^2 values for PM_{10} models were between 0.75 and 0.79, and the corresponding RMSE values were between 9.9 and 10.6μg/m^3. The R^2 values for $PM_{2.5}$ models were between 0.73 and 0.87, and the corresponding RMSE values were between 4.9 and 5.06μg/m^3. The PNC has the R^2 values ranging from 0.8 to 0.83 with RMSE values ranging from 10826 to 11222 number/cm^3. Comparing these results with similar studies, Singh et al. (2012) used PLSR methods to train models for predictions of respirable particulate matter and found the correlation coefficient (R) of 0.84 between the measured and predicted values. Pires et al. (2008) used MLR, PCR and PLSR in the prediction of daily PM_{10} concentrations and found the correlation coefficients of 0.7, 0.76 and 0.77 between the observations and the predictions respectively. Ul-Saufie et al. (2013) chose PCA-MLR as the best method for next two-day predictions of PM_{10} concentrations based on 14.4758 (RMSE), 0.8712 (IoA), and 0.6358 (R^2). The same model was also chosen for predicting next three-day PM_{10} concentration. Performance indicators for PCA-MLR model perform the best with 18.2686 (RMSE), 0.8099 (IA) and 0.5998 (R^2). These results are quite similar to the results obtained in this study. The main difference is that hourly concentrations were predicted in the present study, while the studies mentioned above predicted daily concentrations.

6.7 Summary

This chapter investigates the use of two different data set for predicting roadside particle concentrations and their corresponding roadside increments using five statistical modelling techniques. Also, the chapter investigates the effect of using two hybrid feature selection methods on the prediction performance of the statistical models. The results obtained show that there was no difference in performance of the models when using the two different datasets. However, there was a remarkable difference in the performance of the models when predicting either the roadside particle concentrations or their corresponding roadside increments.

The feature selection methods successfully selected variables that are less correlated and are quite small in number, and that will enhance interpretability. However, where the relationship between the predictor variables and the response variables are nonlinear as is the case in air quality modelling, the models might not capture the underlying relationships. These shortcomings limit the use of the linear models to just prediction rather than to be used for analysing air quality problems based on the relationships of the variables expressed by the models. Therefore, invoking methods that are more sophisticated in handling nonlinear relationships will offer more benefit than using the linear methods if the prediction performance is the primary goal. The use of the feature selection methods on machine learning models in modelling the roadside particles will be investigated in Chapter 7 to compare their prediction performance with the linear models developed in this chapter.

Chapter 7

Machine Learning (ML) Models for predicting roadside particles

7.1 Introduction

The effects of traffic-derived air pollution can be controlled through the provision of adequate and effective air quality control and mitigation measures. These measures are designed and tested with the aid of air quality models. Environmental regulatory agencies have to complement measurements of air quality with models that can predict pollutant concentrations accurately and determine the cause and future extent of the quality problems.

This chapter examines the application of three ML methods including Artificial Neural Networks (ANN), Ensemble regression trees and Support Vector Machines (SVM) in air quality modelling. Five different ANN, formulations including Multi-Layer Perceptron with Principal Component Analysis (PCA-MLP), Multi-Layer Perceptron with Model Averaging (AVG-MLP), Bayesian Regularised Neural Network (BRNN), Extreme Learning Machine (ELM) and Deep Learning (DL) have been considered. Ensemble regression trees considered are Boosted Regression Trees (BRT) and Random Forests (RF). Two SVM kernels including radial basis kernel (SVM-Radial) and linear kernel (SVM-linear) are the variants of the SMV algorithms used. The purpose of selecting these methods is to compare their ease of application, training speed and predictive accuracy among and across similar and different methods respectively.

The selected ML methods were trained to predict roadside particles concentrations (i.e. PM_{10}, $PM_{2.5,}$ and PNC). In the rest of the chapter, Sections 7.2 - 7.6 present the results and discussion of training and test performances of the trained ML models. Section 7.7 compares

the performances of the machine learning models across the different algorithms. Also, the response of the ML models to the GA-RF feature selection procedure carried out in Chapter 6 was evaluated and compared. In section 7.8, seasonal performances of the models are evaluated. Section 7.9 summarises the main findings and conclusions of the chapter.

7.2 Selection of ANN Model Parameters

The selected ANN algorithms were tuned to find the optimum combination of the model parameters (i.e., weight decay and a number of hidden nodes) using 10-fold cross-validation repeated five times and the performance of the models with the selected parameters was evaluated using RMSE and R-squared values.

7.2.1 Multilayer Perceptron with Principal Component Analysis (PCA-MLP)

The results of the PCA-MLP model tuning for PM_{10}, $PM_{2.5,}$ and PNC prediction models are shown in Figure 7.1, from left to right panels respectively. The colour coded lines represent the RMSE values of the models for each combination of the weight decay and a number of hidden nodes. Optimum weight decay values were determined from the following values (i.e. 0,0.001,0.01,0.1,0.2, 0.5,0.7,.8,0.9,1), and the optimum number of hidden neurons was searched between 1 and 50. The range of the values was determined after several trial and error runs of the models. RMSE values for each trained network are shown on the y-axes of the figures against a number of hidden neurons on the x-axes.

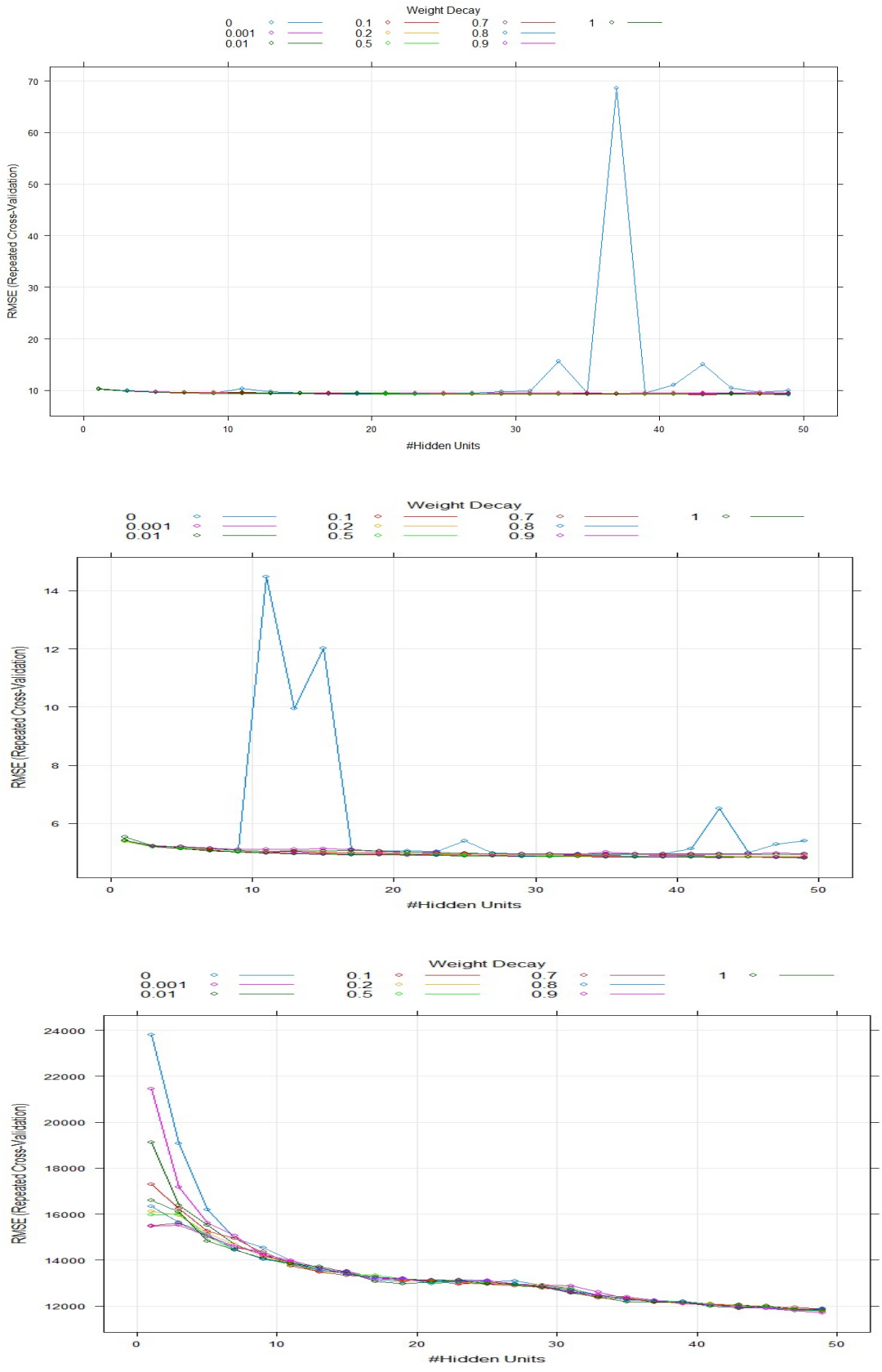


Figure 7.1 Optimisation of PCA-MLP model parameters for PM_{10}(top), $PM_{2.5}$(middle), and PNC (bottom) respectively.

Figure 7.1 shows little change in performance between the models with different weight decay values (> 0) while the models with a higher number of hidden nodes performed marginally better than the models with a lower number of hidden nodes. The effect of increasing number of hidden neurons is more pronounced in PNC models as shown in the right panel of Figure 7.1. The final models selected for the PM_{10}, $PM_{2.5}$ and PNC predictions, were the models trained with 49 hidden neurons and the weight decay values of 0.8, 1.0 and 0.5 respectively. The number of the Principal Components (PCs) which explained 99% of the variance in the data were found to be 19 for the PM_{10} and $PM_{2.5}$ models and 14 for the PNC model when trained with all the predictor variables. However 12 PCs were selected for PM_{10} and PNC models, and 9 PCs for the $PM_{2.5}$ models when trained with RF-GA selected predictor variables.

7.2.2 Neural Networks Using Model Averaging (AVG-MLP)

The AVG-MLP algorithm finds the average prediction of a given number of MLP models to regularise the final model and avoid over-fitting. Similar model tuning procedure explained in section 7.2.1 was adopted for the AVG-MLP models, and the result is shown in Figure F.1 in Appendix F. Figure F.1 shows that the AVG-MLP models developed with different weight decay values performed the same way with little difference and their performance slightly increased with an increase in the number of hidden neurons. Also, when the value of weight decay was set to zero the model performance was unstable as shown in Figure F.1. The final parameters used for the models were 49 hidden neurons and the weight decay values were 1.0, 0.9 and 0.5 for PM_{10}, $PM_{2.5,}$ and PNC prediction models respectively.

7.2.3 Bayesian Regularised Neural Networks (BRNN)

The BRNN algorithm was implemented through a model tuning function (*"train"*) of the caret package (Kuhn, 2012) of R software (R Development Core Team, 2015) where only the hidden layer size was to be optimised. The optimisation process carried out using 10-fold cross-validation repeated five times, was very slow especially when using large training data and a large range of the number of hidden neurons. Therefore, the number of hidden neurons used was determined using trial and error between 1 and 50 and the final models were trained using 10-fold cross-validation repeated five times with the selected number of hidden neurons.

7.2.4 Extreme Learning Machine (ELM)

The ELM algorithm was tuned to choose a proper activation function among six different activation functions namely: Sigmoid, Tan sigmoid, Radial basis, Hard-limit, Sine and Satlins functions(Gosso and Martinez-de-Pison, 2012). The number of hidden neurons between 1 and 100 were used to search for the optimum number of hidden neurons. The model tuning was implemented using 10 – fold cross – validation repeated five times.

Figure F.2 show the performance (RMSE) of the ELM models on the y-axes and the number of hidden neurons on the x-axes while the colour coded lines show the range of performance of the models for each activation function and a number of hidden neurons. Among the activation functions used, the ELM with the sigmoid function performed better than the remaining functions in all the cases considered. Therefore, it was adopted for the training of the final models. In the course of the training, it was discovered that the ELM algorithm needed a much higher number of hidden neurons than the remaining algorithms. Hence, the results of the model tuning were only used for the selection of the activation function, and trial and error was used to decide the number of hidden neurons for the final models. It was

also found out that the performance of the model remained fairly constant when using a number of neurons greater than 1000, and therefore, 1500 was adopted for training the final models using 10 - fold cross-validation repeated 5 times. The ELM was the fastest algorithm used so far in this study. Regardless of the numbers of hidden neurons used and the various activation functions, the results were obtained within five hours in each case depending on the volume of the training data. However, the training results of the MLP based models discussed in Sections 7.2.1 – 7.2.3 above, were obtained within 3 -10 days using HPC computer clusters with 15 cores depending on the volume of the data.

7.2.5 Deep Learning

In this study, the deep learning algorithm was first applied to do a grid search for the optimal number of hidden neurons and number of layers suitable for modelling the roadside particle concentrations. Three to four hidden layers with a varied combination of the number of neurones between 100 and 500 in each of the three or four hidden layers were tested. Also, three L1 (Lasso penalty) values 0.001, 0.0001 and 0.00001 were tested alongside the number of hidden layers and neurons mentioned above. Other regularisation parameters including input dropout and hidden neurons dropout were set to default values of the software. The grid search was implemented using 10-fold cross-validation and 10 epochs. After selecting the correct number of layers and neurons, the number of epochs was found to have a significant impact on the training performance, and 100 epochs were selected after several trials on the number of epochs between 10 and 1000. Three layers with 200 neurons each was found to be suitable for the training and with L1 = 0.00001.

7.2.6 Artificial Neural Network Training Results

In this section, the results of the training of the five ANN algorithms explained in sections 7.2.1 – 7.2.5 are presented and compared. The training results for all the ANN algorithms are presented in Table 7.1.

Table 7.1 shows that the PCA-MLP trained with the input variables selected by the RF-GA method to predict PM_{10} and PNC performed slightly better than those trained with all the input variables. While for the $PM_{2.5}$ predictions, the models trained with all the input variables performed relatively better. This is an indication that this type of algorithm needs feature reduction as well as the dimension reduction during the training. Also, the algorithm showed better training performance when trained for $PM_{2.5}$ predictions than in the case of PM_{10} and PNC predictions. The AVG-MLP models for predicting PM_{10} showed slightly better training performance than the PCA-MLP models with R^2 and RMSE values of 0.85 and 8.27 µg/m^3 as against 0.82 and 9.08 µg/m^3 for PCA-MLP respectively.

Table 7.1 Training Results for the ANN models

Row Labels	R-squared	RMSE
PM_{10}		
All variables		
PCA-MLP	0.82	9.08
AVG-MLP	0.85	8.27
BRNN	0.84	8.49
ELM	0.82	9.04
Deep Learning	**0.94**	**5.18**
RF-GA		
PCA-MLP	0.83	8.80
AVG-MLP	0.83	8.80
BRNN	0.83	8.82
ELM	0.82	9.22
Deep Learning	**0.91**	**6.34**

*Note: In Tables 7.1, the terms **PM_{10}**, **$PM_{2.5}$** and **PNC** in the first column from the left represent the response variables predicted by the models. Also, the terms **all variables** and **RF-GA** refers to the training data with all the variables and the training data with the RF-GA selected variables respectively. The acronyms for the models are shown in the first column while the first row shows the statistical performance metrics.*

Table 7.1 *continued*

Row Labels	R-squared	RMSE
PM2.5		
All variables		
PCA-MLP	0.89	4.53
AVG-MLP	0.91	4.18
BRNN	0.91	4.27
ELM	0.88	4.72
Deep Learning	**0.96**	**2.58**
RF-GA		
PCA-MLP	0.88	4.82
AVG-MLP	0.89	4.63
BRNN	0.88	4.74
ELM	0.87	4.90
Deep Learning	**0.91**	**4.24**
PNC		
All variables		
PCA-MLP	0.80	12458
AVG-MLP	0.84	11044
BRNN	0.87	10257
ELM	0.86	10442
Deep Learning	**0.92**	**7867**
RF-GA		
PCA-MLP	0.82	11765
AVG-MLP	0.83	11392
BRNN	0.86	10337
ELM	0.86	10384
Deep Learning	**0.91**	**8249**

The models trained without feature selection performed slightly better than the ones trained with the RF-GA selected variables as indicated by the R^2 and RMSE values (see Table 7.1). Although the difference in the performance was small, it shows that the method has a bias towards having more predictor variables than the small number of selected predictor variables. Also as in the case of the PCA- MLP algorithm, the AVG-MLP models for the prediction of $PM_{2.5}$ performed slightly better than those trained for the prediction of PM_{10} and PNC concentrations as indicated by their higher values of coefficient of correlation (R).

The BRNN algorithm like AVG-MLP performed slightly better when trained without prior feature selection. Also, it showed similar performance to PCA –MLP and AVG – MLP algorithms when trained with all the variables while slightly underperformed when trained

with the RF-GA selected variables. However, these differences are so small that they can hardly influence the overall performance of the models in practice.

The ELM algorithm showed a similar performance when trained with the two data sets, and its performance is comparable with the performance of the PCA-MLP and AVG-MLP while showing a slightly poorer performance than the BRNN models for the prediction of PM_{10} and $PM_{2.5}$. However, in the case of PNC prediction, its performance is similar to the performance of BRNN algorithm.

The Deep Learning algorithms showed better training performance than all the models discussed above, with R^2 values 0.91, 0.9 and 0.91 for the prediction of PM_{10}, $PM_{2.5}$ and PNC respectively. They have fewer prediction errors (RMSE) and higher R-squared values in all the cases considered. Moreover, the Deep Learning algorithms showed better performance when trained with all the predictor variables than when trained with the RF-GA selected variables. Largely, the training performance of the remaining ANN models is quite good and similar to the performance of Deep learning algorithm. However, the training performance might only be a signal to how the models will perform using the test data. Therefore, the final conclusion will be drawn from the test results presented in Section 7.2.7.

7.2.7 Comparison of The Test Performance Of ANN Models

The test data set was kept hidden from the models during the training. Therefore, it was used to test the actual performance of the models in predicting new data. The statistical and graphical performance metrics explained in Chapter 3 were applied to compare the accuracy of the models in predicting the roadside particle concentrations. Table 7.2 show the statistical performance of the PM_{10}, $PM_{2.5,}$ and PNC models.

Table 7.2 Test performance statistics for the ANN models

Row Labels	IOA	COE	R	FAC2	RMSE	NMGE.	NMB.	MGE.	MB.
PM_{10}									
all variables									
AVG-MLP	0.81	0.61	0.87	0.98	10.87	0.15	-0.01	6.71	-0.31
BRNN	0.83	0.67	0.90	0.99	9.51	0.13	-0.01	5.75	-0.24
Deep learning	0.85	0.71	0.93	0.99	7.94	0.12	0.00	5.11	-0.13
ELM	0.82	0.64	0.90	0.99	9.79	0.14	-0.01	6.15	-0.30
PCA-MLP	0.82	0.64	0.90	0.99	9.87	0.14	0.00	6.21	-0.20
RF_GA									
AVG-MLP	0.83	0.67	0.90	0.99	9.36	0.13	0.00	5.70	0.00
BRNN	0.82	0.64	0.89	0.99	9.93	0.14	0.00	6.21	-0.01
Deep learning	0.83	0.67	0.91	0.99	9.20	0.13	0.01	5.67	0.40
ELM	0.82	0.64	0.89	0.99	9.80	0.14	0.00	6.22	-0.02
PCA-MLP	0.82	0.64	0.89	0.99	9.79	0.14	0.00	6.11	0.03

All the ANN models performed well on the test data with 94 – 99% of their predictions falling within the factor of two of the observed particle concentrations as shown by the FAC2 values in Table 7.2. The percentages are well above the minimum of 50% recommended to the DEFRA UK (Derwent et al., 2010) for the acceptance of an air quality model. The models showed low bias in their prediction, however, most of them slightly under predicted the PM_{10} and $PM_{2.5}$ concentration levels as indicated by the negative sign of the MB and NMB values.

Derwent et al. (2010) recommended that for an air quality model to be accepted it should satisfy the minimum requirement of NMB values in the range between -0.2 and +0.2. The NMB values obtained here for all the ANN models are nearly zero which shows that they performed way above the minimum requirement. MB and the NMB values of the deep learning are slightly higher than the corresponding values for all the other algorithms. The FAC2 and the NMB values are the most commonly used measures to describe the agreement between the observed and the predicted values and the results presented in Table 7.2 show that the models have performed extremely well. Although the performance of all the models for predicting PM_{10} is close, the Deep learning algorithm trained with all the variables performed much better than the remaining algorithms. It has the lowest prediction error indicated by the RMSE and MGE values, and it has the highest model – observation

agreement indicated by the COE, IOA and R values. Its outstanding performance has been consistently indicated by all the performance metrics.

The AVG-MLP models showed a slight improvement in performance when trained with the RF-GA selected variables while Deep learning and BRNN models performed slightly better when trained with all the predictor variables. In most of the cases, ELM and PCA-MLP performed in the same way. The performance of the models trained to predict $PM_{2.5}$ did not vary much with the type of the training data, although most of them showed little bias towards using all the predictor variables. In the case of PNC predictions, Deep Learning, and BRNN performed similarly and better than the remaining models. They have lowest RMSE values and higher COE, IOA, and R values, but similar NMB values. PCA-MLP is the least performing model in this case, with higher RMSE, NMGE errors and lower FAC2 and COE values.

The prediction errors (RMSE) of the ANN models trained with all the variables is slightly higher than those trained with the RF-GA selected variables. The effectiveness of the feature selection process in all the three cases of prediction considered can be deduced from the fact that the performance of the models is very much similar with very little differences despite the reduction in the number of predictor variables used. Modelling the particles with fewer variables could reduce the cost of the models in terms of the cost of measuring the variables and in terms of computational cost which is higher when dealing with many variables.

7.3 Ensemble Regression Tree models

7.3.1 BRT and RF Training

The BRT algorithm has basically three tuning parameters i.e. number of trees, shrinkage parameter, and interaction depth. The formulation of the BRT in H2O R package (Malohlava and Hank, 2015) was adopted in this research. However, the BRT model parameters (i.e. learning rate and interaction depth) were first determined using grid searches over a selected

range of the parameters and using 10-fold cross-validation repeated five times to train the BRT models for each combination of the two parameters. The parameters that yielded the best performing BRT models using the grid search varies little with default parameters of the H2O software, and they did not produce models with better performance, therefore, the default parameters of the learning rate (lr = 0.1) and interaction depth (d = 5) were adopted and the number of trees was set to 1000 using the results of the grid search as shown in Figure 7.2 below.

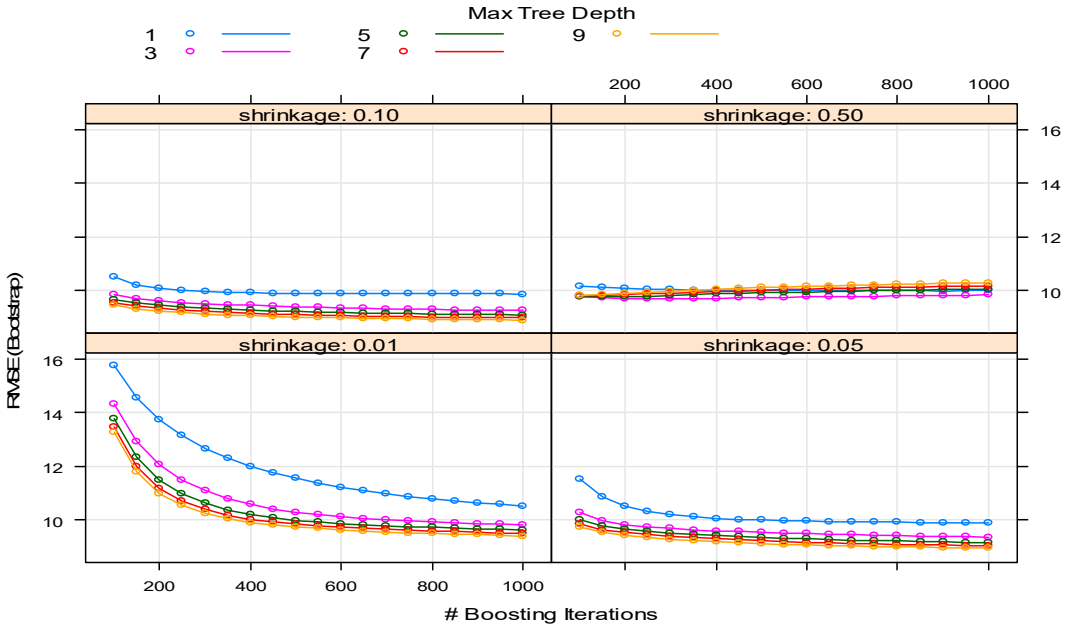


Figure 7.2 Determination of BRT model Parameters.

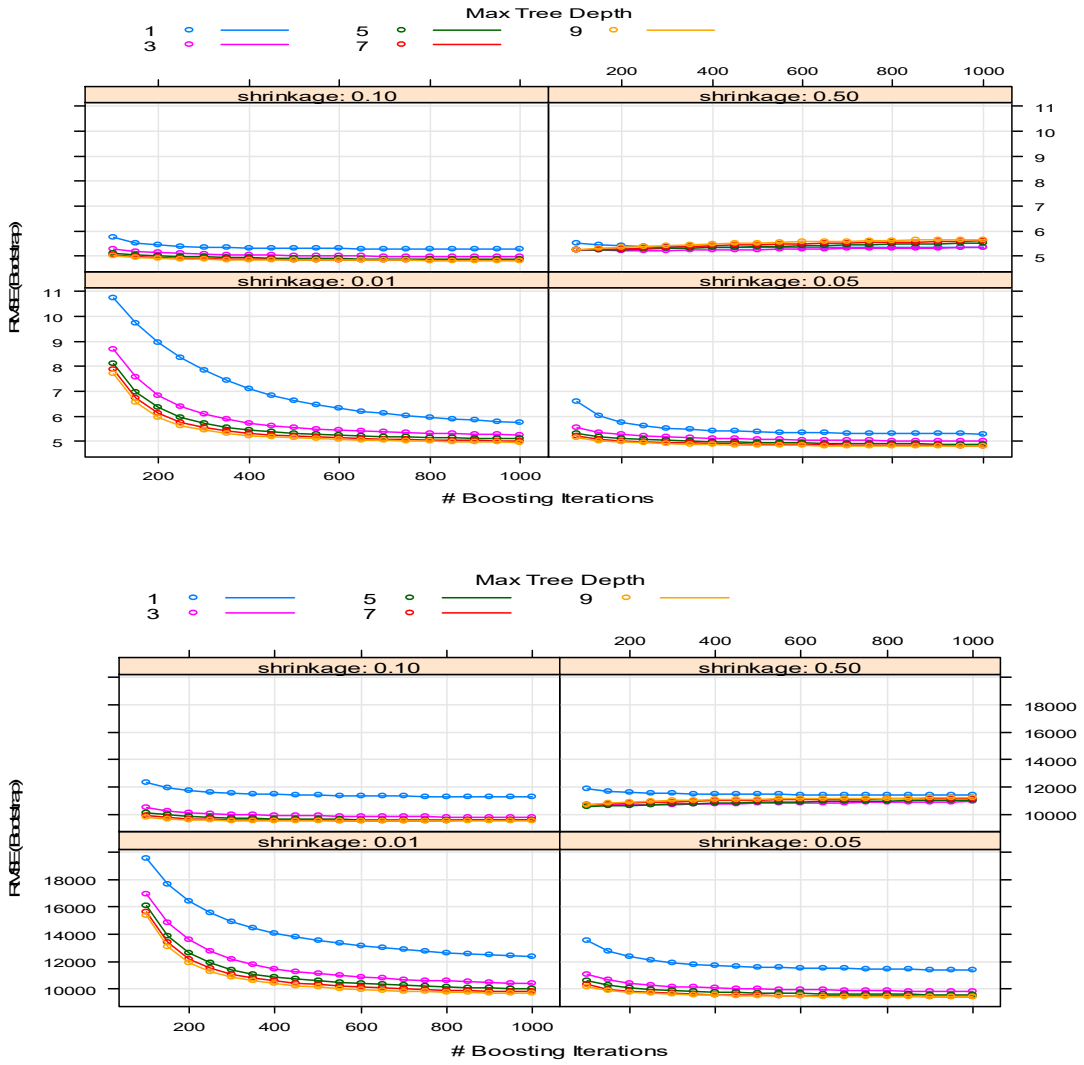


Figure 7.2 *continued*

Note in Figure 7.2; the colour coded lines show the interaction depth values and each box in the panels display the results for a particular learning rate (shrinkage) values. The x and y-axes show the number of trees and the RMSE values for the models respectively. The panels from top to bottom represent the grid parameter search results for PM_{10}, $PM_{2.5}$, and PNC prediction models respectively.

The RF method was also implemented using the formulation of the H2O R package (Malohlava and Hank, 2015). RF model parameters tuned were interaction depth and number of trees which were set to 20 and 1000 respectively after several trial-and-error runs. The algorithm was then applied to train the models for predicting PM_{10}, $PM_{2.5}$ and PNC

concentrations. The models were trained and tested using the same data used for the statistical models and ANN models to allow for comparison.

7.3.2 Training Performance of The BRT and RF Models

During the training, the BRT and RF algorithms produced two training performance results, each for the best-trained model during the cross-validation and the overall cross-validation as shown in Tables 7.3.

The results in Table 7.3 show that the BRT models might have a higher tendency for over-fitting the data than when trained using the cross-validation. This could be seen from the margin between the performance (RMSE and R-squared values) of the best performing individual BRT models and the cross- validated performance of the BRT models. Also, the performance of the BRT model is somewhat insensitive to the feature selection performed before the training of the models as the training performances for the BRT models with and without feature selection are similar. The BRT models trained for predicting PNC and $PM_{2.5}$ performed slightly better than those for predicting PM_{10}. The training performance of the RF models indicates that both the best trained RF and the cross-validated performance are similar showing that even if the models were trained without cross-validation, they might have a low tendency of over-fitting the data.

Table 7.3 Training Results for BRT models

Row Labels	R-squared	RMSE
PM_{10}		
All variables		
BRT		
Best Training incident	0.96	4.43
Cross validation	0.86	7.99
RF		
Best Training incident	0.84	8.62
Cross validation	0.84	8.72
RF-GA		
BRT		
Best Training incident	0.94	5.25
Cross validation	0.85	8.44
RF		
Best Training incident	0.84	8.62
Cross validation	0.84	8.69

RF models trained with all the input variables have shown similar performance to those trained with the RF – GA selected variables. The sensitivity of the models to the feature selection is similar to some of the ANN models where they show the same performance with the two data sets. Two advantages could be drawn from this behaviour. First, the models can choose the most important variables itself without performing feature selection separately. Second, if the feature selection becomes necessary for cost reduction purposes, it will not affect the performance of the models negatively once it's done appropriately.

7.3.3 Test Performance of the BRT and RF Models

After the training, the BRT and RF models were tested using the test data which was not used during the training and the test performance results are shown in Table 7.4.

The test performance results indicate that the BRT and RF models performed very well with very low bias and higher agreement between the predicted and the observed particle concentrations which can be seen from the NMB, COE, IOA, FAC2 and R values. The results also show that all the BRT models for the prediction of $PM_{2.5}$ and PNC performed similarly irrespective of the set of predictor variables used alluding to the insignificance of the RF-GA feature selection on the predictive accuracy of the models. However, for the PM_{10} predictions, the models trained with RF-GA selected variables performed better than those trained with all the predictor variables. The performance of the BRT models across the pollutant metrics is largely similar.

The RF models show similar performance to the BRT models especially in the prediction of $PM_{2.5}$ and PNC but in the case of PM_{10} prediction, the BRT models are slightly better than the RF models. Both BRT and RF models, unlike the ANNs, performed similarly with and without the RF-GA feature selection carried out before the training.

Table 7.4 Comparison of the test performance of the BRT and RF models

Row Labels	IOA	COE	R	FAC2	RMSE	NMGE	NMB	MGE	MB
PM_{10}									
All variables									
BRT	0.85	0.71	0.92	0.99	8.79	0.12	0.00	5.09	-0.05
RF	0.84	0.68	0.90	0.99	9.50	0.13	0.00	5.53	-0.02
RF - GA									
BRT	0.88	0.77	0.96	1.00	6.37	0.09	0.00	4.05	-0.05
RF	0.84	0.68	0.91	0.99	9.19	0.13	0.00	5.55	0.11
$PM_{2.5}$									
All variables									
BRT	0.86	0.71	0.95	0.98	4.31	0.13	0.00	2.97	0.01
RF	0.85	0.70	0.94	0.98	4.51	0.13	0.00	3.13	0.07
RF - GA									
BRT	0.85	0.70	0.95	0.98	4.27	0.13	0.00	3.11	0.04
RF	0.85	0.70	0.95	0.98	4.27	0.13	0.01	3.11	0.14
PNC									
All variables									
BRT	0.88	0.77	0.94	0.99	8787	0.14	0.00	4532	1.78
RF	0.88	0.77	0.94	0.99	9032	0.14	0.00	4494	89.45
RF - GA									
BRT	0.88	0.76	0.95	0.99	8150	0.14	0.00	4547	-127.62
RF	0.89	0.77	0.96	0.99	7955	0.13	0.00	4317	101.21

This insensitivity might be because they have built-in feature selection mechanisms and/or because the feature selection used (RF-GA) involved a tree based model (i.e. random forests). This behaviour is at variance with the behaviour of some of the ANN models where some of the algorithms appreciated the use of the RF-GA feature selection while some showed higher performance with a higher number of predictor variables. The statistical models also showed less sensitivity to feature selection in terms of the model performances as discussed in Chapter 6. However, the tree based models performed much better.

7.3 Estimation of Variable Importance

The BRT, RF, and Deep learning implemented using H2O package provide an estimate of variable importance as shown in Figure 7.3. The upper panels are for the models trained with all the variables while the bottom panels are for the models trained with GA-RF

selected variables. Considering models trained with all the variables to predict PM_{10} and $PM_{2.5}$ (see Appendix F – Figure F.3), the BRT selected NOx, NO_2, and CO and PM.bg as the most contributing predictor variables to the models while RF selected SO_2 as well. Besides these variables, the remaining variables gave nearly equal contributions to the BRT and RF models. However, all the Deep learning models showed a slightly different pattern where the meteorological variables and temporal variables are the most important variables followed by the gaseous pollutants and the least important variables were traffic variables. For the PNC prediction models, both the BRT and RF selected NOx as the most important variable while RF selected NO and NO_2 as the second most important variables while BRT gave equal weight to all the other variables (see Figure F.4).

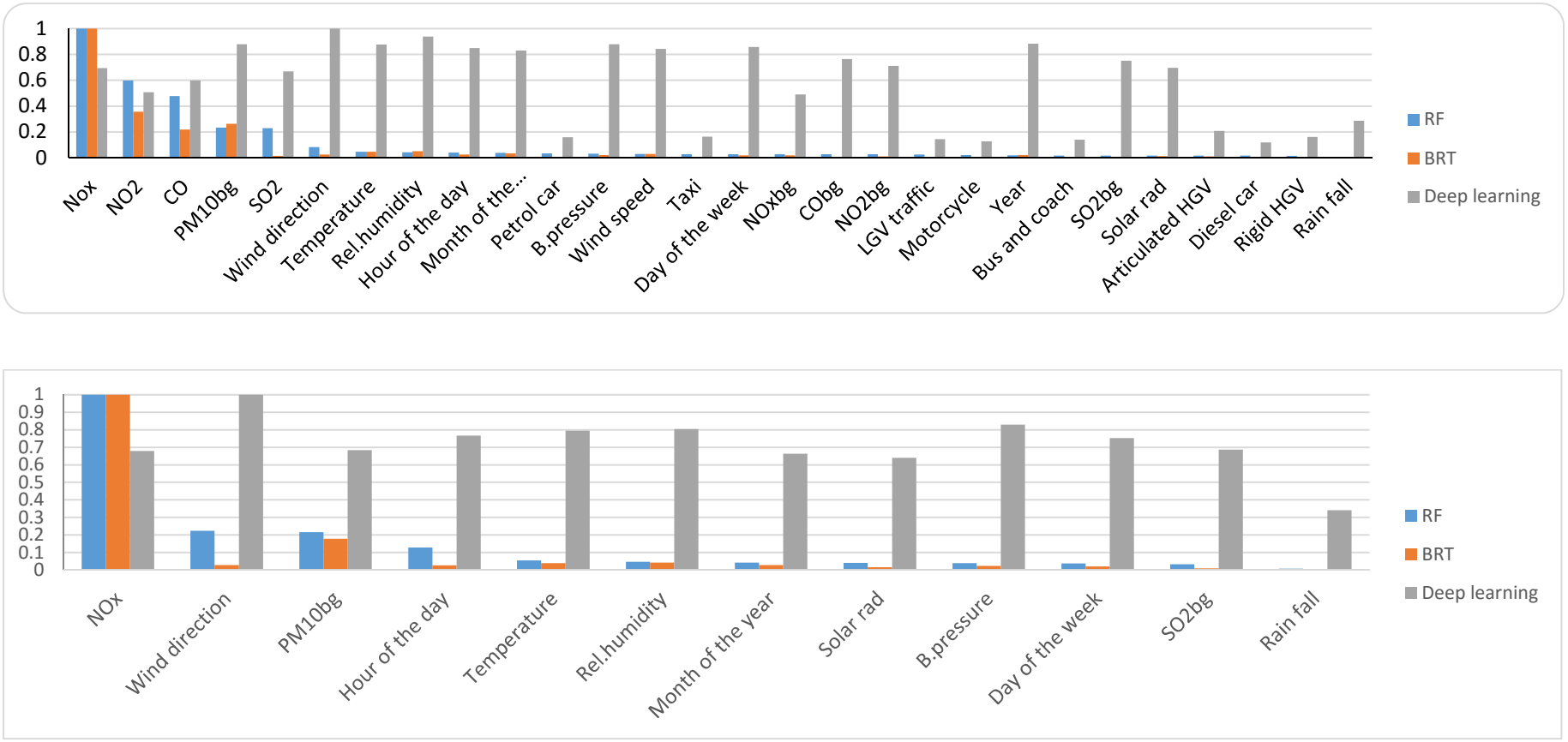


Figure 7.3 Variable importance estimated by machine learning models for the prediction of PM_{10}

Note: In Figures 7.3, the upper panel is for the model trained with all the variables while the right panel is for the model trained with GA-RF selected variables.

The BRT and RF models trained with selected predictor variables showed that NOx and background particle concentrations are the most contributing predictor variables while the remaining variables were assigned nearly similar and much lower contributions. The Deep learning models also showed similar behaviour when trained with all the predictor variables. The notable difference between the estimates of the tree based models and the deep learning is in the distribution of the contributions. While BRT and RF models gave higher weights to a small number of predictor variables, the deep learning models distribute the contribution to so many variables and the differences in the contributions are small compared with the difference estimated by BRT and RF.

7.4 Partial Dependence Plots

A partial dependence plot is another important formulation in the BRT algorithm that allows for the display of the effect of an input variable on the target variable while taking into account the average effects of all other variables in the BRT model (Carslaw and Taylor, 2009, Friedman, 2001). Although the integrity of the plots is affected by highly correlated variables, they give a useful basis for interpreting the models (Elith et al., 2008, Friedman and Meulman, 2003). Figures 7.4 – 7.6 show the partial dependence plots for the predictor variables used in the training of the BRT models for the prediction of roadside particles.

The partial dependence plots revealed that the roadside particle concentrations increase with the corresponding increase in roadside NOx concentrations. This relationship is described by the roughly linear line graphs shown in Figure 7.4. The pattern of the relationship is different with PNC concentration where the slope of the line in the plot flattened when the NOx concentration was around 300 $\mu g/m^3$ and become steeper again at around 1000 $\mu g/m^3$ where the PNC concentration increases without a corresponding increase in the NOx concentrations (see Appendix F – Figure F.5).

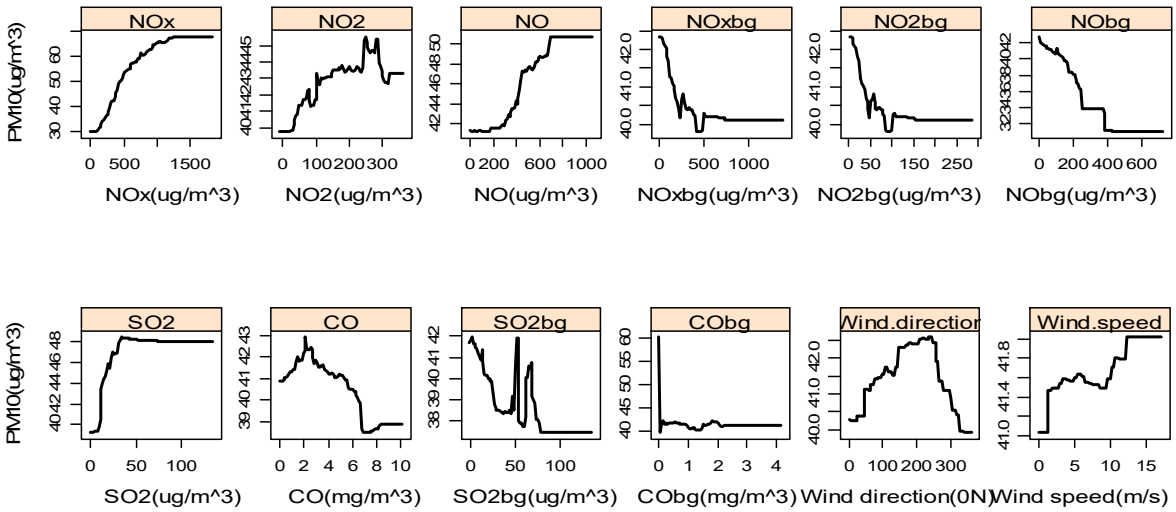


Figure 7.4**.** Partial dependence plots showing the effects of pollutants and wind variables on the BRT model predictions of the roadside particle concentrations.

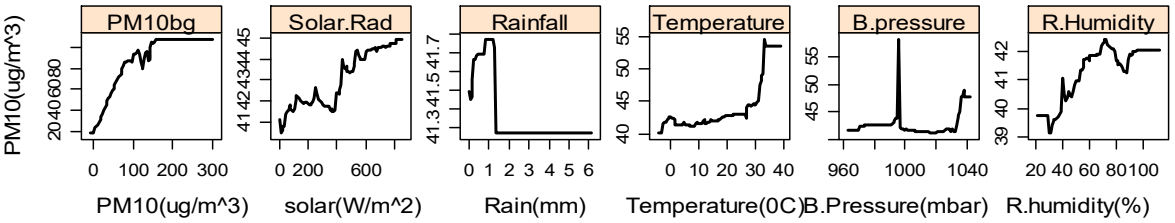


Figure 7.5. Partial dependence plots showing the effects of background particle concentrations and meteorological variables on the BRT model predictions of the roadside particle concentrations.

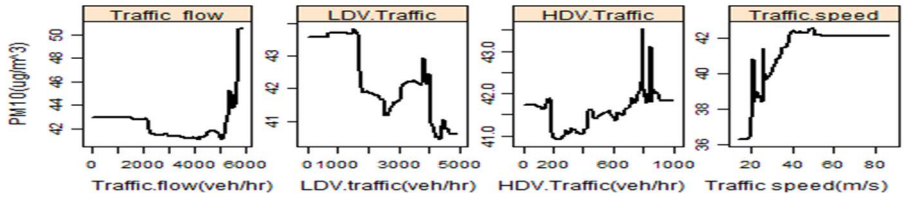


Figure 7.6. Partial dependence plots showing the effects of traffic variables on the BRT model predictions of roadside particle concentrations.

The NO_2 concentrations show a parabolic relationship with the PM_{10} and $PM_{2.5}$ concentrations, and they increased with the corresponding increase in NO_2 concentrations up to around 200 $\mu g/m^3$ of the NO_2 and then decreased with further increase in the NO_2 concentrations. However, the NO_2 concentration shows a negative linear relationship with the PNC and all the particle concentrations show a positive linear relationship with the NO and background particle concentrations. The PM_{10} and PNC concentrations decreased with corresponding increases in the background concentrations of NOx, NO_2, and NO while the $PM_{2.5}$ concentrations increase with a corresponding increase in their concentrations.

The roadside SO_2 concentrations show a linear relationship with the PM_{10} and $PM_{2.5}$ when its concentration was between 0 and 20$\mu g/m^3$ and then the relationship remained constant over the remaining range of the concentrations. On the other hand, the PNC concentrations decrease with corresponding increases in SO_2 concentrations up to 20$\mu g/m^3$ of SO_2 and then the relationship changes to positive linear up to around 35$\mu g/m^3$ and then remains constant for the rest of the values.

The BRT model shows that the PM_{10} and $PM_{2.5}$ concentrations have a negative linear relationship with the CO concentrations while having a positive linear relationship with the PNC. The positive relationships between the particles and most of the gaseous (NOx, NO_2, NO and SO_2) pollutants show that the gaseous pollutants play a vital role in the formation of the particles, or they share common sources. This information could give a clue on the intricate relationship between gaseous and particle pollutants and will help in taking an urgent decision before conducting a detail laboratory analysis of the relationships.

The BRT models also showed that the higher particle concentrations are more associated with the winds coming from the south, south-west, and south-east (see Figure 7.8). These are the directions of the dominant winds at the site where the data was collected. Also, these

directions coincide with the side of the road (southern side of Marylebone road) where the monitoring unit is located which suggests that canyon recirculation vortices delivered most of the particle concentrations to the monitoring unit.

This information is also useful as it provides a clue on whether the monitoring unit is located in the correct position or not. The relationship between the wind speeds and the concentrations of $PM_{2.5}$ and PNC was shown to be negative linear. This relationship is expected because when the wind speed is high, the ventilation in the street increases and then most of the particle concentrations are removed from the street. However, the PM_{10} concentrations show the opposite where they increased with the corresponding increase in wind speed. The possible explanation for this relationship is that the higher winds might carry dust and other larger size particles especially non-exhaust particles which could have raised the concentrations of the PM_{10}.

Temperature and relative humidity show a nearly positive linear relationship with the PM_{10} and $PM_{2.5}$ while the PNC concentrations showed a nearly linear relation with temperature and nearly constant relationship with relative humidity. The positive association between the particle concentrations and the temperature did not agree with the findings of an earlier study (Dos Santos-Juusela et al. (2013). However, Barmpadimos et al. (2011) reported a positive relationship between temperature and PM_{10} in the summer and Tai et al. (2010) found a positive correlation between most of the components of $PM_{2.5}$ except for nitrate which shows a negative association.

To investigate further, a linear correlation between the temperature and the particle concentrations was estimated, and the coefficients of correlations between them were found to be 0.15, 0.14 and 0.26 for PM_{10}, $PM_{2.5,}$ and PNC respectively. Moreover, the elastic net models also show positive relationships with temperature. This relationship needs to be

further investigated particularly to find out the seasonal relationship between the temperature and the particles and the levels at which the relationship changes.

The traffic flow and the HDV traffic show a negative linear relationship with the concentrations of $PM_{2.5}$ which is not in agreement with the fact that the concentration increases with the corresponding increase in traffic flow. However, it might explain the stop and go situation at the site, where the emissions are high when the vehicles are not moving and during acceleration and then reduces as the flow becomes normal. However, in the case of PM_{10} and PNC, the concentrations remained fairly constant when the traffic flow was between 2000 and 4000 veh/hr and then suddenly increased to higher concentrations and then remained constant as shown in Figure 7.7. This could explain the situation when the road reaches its capacity where the concentration increases as a result of a high number of vehicles. HDV and LDV traffic captured the hourly variation of the concentrations of PM_{10} and PNC. The HDV traffic shows strong associations with the average PNC concentration, and it shows a bimodal distribution which suggests that it keeps track of the temporal variation of the PNC in the model.

The positive linear relationship shown by most of the variables indicates the sign of their contribution in determining the suitable prediction. For example, the positive relationship might be excitatory while the negative relationship might be inhibitory in deciding the final predicted value, therefore, both the input variables with the positive and negative relationships are vital in determining the final predictions of the model.

The analysis of the partial dependence plots could help the model user to have a fair understanding of the relationship between the predictor variables and the particle concentrations. The information gained could inform several management decisions related to the control of air quality. For example, any control measure taken to reduce the roadside

oxides of nitrogen will have a significant impact on the particle concentrations due to their strong relationship explained by the BRT models. Also, accurate determination of the levels of oxides of nitrogen could yield better BRT models for the prediction of roadside particles.

7.5 Support Vector Machines (SVM)

The SVM algorithm was applied to the same training and testing data used for the ANN and tree based models to allow for performance comparison of the trained models. The training was applied using 10-fold cross-validation repeated five times to the two data sets and using linear and radial basis kernels (See Chapter 2). SVM parameters to be determined during the training were the kernels to be used and the cost and sigma values. Therefore, linear and radial basis kernels were selected for the training, and the sigma values were determined empirically by a function in the software while the SVM models were tuned to 15 cost values between 0.25 and 4096 on a log scale. The cost value 1 was selected by all the linear kernels while the formulation for the linear kernel used does not need tuning of sigma values. The SVM models with the combination of the cost and sigma values that yields best-performing models were selected as the final models.

7.5.1 Results of the Training Performance for SVM Models

The training result shown in Table 7.5 revealed that the SVM with radial basis kernels performed much better than those with linear kernels and all the models performed slightly better when trained with all the variables than with the RF-GA selected variables. The cost and sigma values selected for PM_{10} and $PM_{2.5}$ are lower with an increasing number of variables while the reverse was observed for the cost values of the models for PNC prediction. The training performances vary little with the performances of the ANN, BRT, and RF models.

Table 7.5 Training performance for SVM models

Row Labels	C	Sigma	R-squared	RMSE
PM_{10}				
All variables				
SVM Linear	1.0	-	0.81	9.91
SVM Radial	16	0.03	0.88	7.72
RF-GA				
SVM Linear	1.0	-	0.80	10.21
SVM Radial	32	0.07	0.88	7.96
$PM_{2.5}$				
All variables				
SVM Linear	1.0	-	0.87	4.99
SVM Radial	8	0.03	0.89	4.61
RF-GA				
SVM Linear	1.0	-	0.83	5.64
SVM Radial	64	0.09	0.89	4.64
PNC				
All variables				
SVM Linear	1.0	-	0.75	16656
SVM Radial.	4096	0.04	0.82	11982
RF-GA				
SVM Linear	1.0	-	0.82	11725
SVM Radial	32	0.07	0.88	9468

The terms ***PM_{10}, $PM_{2.5}$,*** and ***PNC*** in the first column from the left represent the response variables predicted by the models. Also, the terms ***all variables*** and ***RF-GA*** in Table 7.1 refers to the training data with all the variables and the training data with the RF-GA selected variables respectively. The acronyms for the models are ***SVM Linear*** and ***SVM Radial*** in the first column while the first row shows the statistical performance metrics.

variables while the reverse was observed for the cost values of the models for PNC prediction. The training performances vary little with the performances of the ANN, BRT, and RF models.

7.5.2 Test Performance of the SVM Models

The test performance of the SVM models with linear and radial basis kernels trained using the two data sets are shown in Table 7.6. The SVM models trained with RF-GA selected variables performed better than those trained with all the variables considering the RMSE, NMGE, and the COE values. More than 95% of the prediction of the models are within the factor of two of the observed concentrations, and they show very low bias towards underestimation of the actual observations (see FAC2 and NMB values).

Table 7.6. Test performance statistics for the SVM models

Row Labels	NMGE	NMB	IOA	COE	R	RMSE	MGE	MB	FAC2
PM_{10}									
All variables									
SVM linear	0.16	0.01	0.80	0.6	0.87	11.01	6.93	0.48	0.99
SVM radial	0.19	-0.08	0.76	0.52	0.82	13.14	8.33	-3.54	0.94
RF-GA									
SVM linear	0.17	-0.01	0.79	0.58	0.86	11.06	7.18	-0.31	0.99
SVM radial	0.15	-0.02	0.81	0.62	0.88	10.33	6.46	-0.67	0.99
$PM_{2.5}$									
All variables									
SVM linear	0.15	0.00	0.83	0.66	0.93	5.12	3.49	0.00	0.97
SVM radial	0.13	-0.01	0.86	0.72	0.95	4.25	2.91	-0.14	0.98
RF-GA									
SVM linear	0.15	-0.01	0.83	0.66	0.93	5.00	3.56	-0.23	0.98
SVM radial	0.13	0.00	0.85	0.7	0.95	4.33	3.12	-0.03	0.98
PNC									
All variables									
SVM linear	0.25	-0.12	0.79	0.58	0.88	15120	8193	-4058	0.96
SVM radial	0.14	-0.04	0.88	0.76	0.93	9889	4572	-1269	0.99
RF-GA									
SVM linear	0.16	-0.04	0.86	0.72	0.93	10397	5283	-1205	0.98
SVM radial	0.13	-0.01	0.89	0.77	0.95	8310	4392	-431	0.99

The prediction errors of the SVM models for the prediction of PM_{10} are relatively higher compared with those of ANN, BRT and RF models as indicated by RMSE and NMGE values. The performance of the SVM models with a linear kernel is similar to the performance of the linear models presented in Chapter 6.

7.6 Comparison of the Test Performances of the Machine Learning Models

In this section, the performance of all the machine learning models discussed in the previous sections are compared using statistical performance metrics (Table 7.7), time variation plots (Figure 7.8), conditional quantile plots (Figure 7.9), scatter plots (Figure 7.10) and Taylor's diagrams (Figure 7.11). The performance is compared in terms of the agreement between the observed and the predicted concentrations and the response of the models to the use of feature selection before the model training.

Table 7.7. Comparison of the performance statistics of the Machine learning models.

Row Labels	IOA	COE	R	RMSE	NMGE	NMB	MGE	MB	FAC2
PM_{10}									
all variables									
AVG-MLP	0.81	0.61	0.87	10.87	0.15	-0.01	6.71	-0.31	0.98
BRNN	0.83	0.67	0.90	9.51	0.13	-0.01	5.75	-0.24	0.99
BRT	0.85	0.71	0.92	8.79	0.12	0.00	5.09	-0.05	0.99
Deep learning	0.85	0.70	0.93	7.94	0.12	0.00	5.11	-0.13	0.99
ELM	0.82	0.64	0.90	9.79	0.14	-0.01	6.15	-0.30	0.99
PCA-MLP	0.82	0.64	0.90	9.87	0.14	0.00	6.21	-0.20	0.99
RF	0.84	0.68	0.90	9.50	0.13	0.00	5.53	-0.02	0.99
SVM linear	0.80	0.60	0.87	11.01	0.16	0.01	6.93	0.48	0.99
SVM radial	0.76	0.52	0.82	13.14	0.19	-0.08	8.33	-3.54	0.94
RF-GA									
AVG-MLP	0.83	0.67	0.90	9.36	0.13	0.00	5.70	0.00	0.99
BRNN	0.82	0.64	0.89	9.93	0.14	0.00	6.21	-0.01	0.99
BRT	0.88	0.77	0.96	6.37	0.09	0.00	4.05	-0.05	1.00
Deep learning	0.83	0.67	0.91	9.20	0.13	0.01	5.67	0.40	0.99
ELM	0.82	0.64	0.89	9.80	0.14	0.00	6.22	-0.02	0.99
PCA-MLP	0.82	0.64	0.89	9.79	0.14	0.00	6.11	0.03	0.99
RF	0.84	0.68	0.91	9.19	0.13	0.00	5.55	0.11	0.99
SVM linear	0.79	0.58	0.86	11.06	0.17	-0.01	7.18	-0.31	0.99
SVM radial	0.81	0.62	0.88	10.33	0.15	-0.02	6.46	-0.67	0.99

The statistical performance of the machine learning models is presented in Table 7.7. The performance of the models measured by FAC2, R, NMB, NMGE and IOA values rounded to one decimal place is largely similar across the different training data and the response variables. However, using RMSE, MGE and COE values some important differences could be deducted from the performance of the models. The RMSE, MGE and COE values for the models showed that BRT is the best performing model when feature selection is applied before the training while it has similar performance with the Deep learning when trained with all the variables and they performed better than the remaining models in the case of PM_{10} prediction. For the models trained with all the predictor variables to predict $PM_{2.5}$, the BRNN, BRT, RF, and BRNN performed similarly and slightly better than all the remaining algorithms considering their low prediction errors measured by the RMSE and MGE and they also show higher observed-predicted agreement as indicated by the higher values of IOA, COE, and R values.

This trend is also the same in the case of using RF-GA selected predictor variables with AVG-MLP joining the BRNN, BRT, RF, and SVM radial in performance. In the case of PNC prediction, BRT, RF and SVM radial also performed similarly and slightly better than all the ANN algorithms. However, they are closely followed by Deep learning and BRNN algorithms as shown by their respective IOA, COE and R values.

From this analysis, it could be seen that the BRT and RF models have consistently shown higher performance throughout closely followed by SVM radial, BRNN, and Deep learning algorithms while AVG-MLP, PCA-MLP and ELM algorithms performed similarly in most of the cases and with slightly less performance than those mentioned above.

Considering that the performance statistics did not give much insight into the differences in the performance of the models some visual performance evaluation mechanisms are used to explore further the predictive performance of the models.

The hourly time variation plots in Figures 7.7, and Figures F.6 - F.7 in Appendix F show the plots of the observed particle concentrations and their corresponding predictions by the ML models. The prediction of the models captured the hourly pattern of the observed particle concentrations very well. Although the original nature of the particle concentration profile has been slightly distorted by the random division of the original data, the models reflected the changes in their predictions.

The plots show that the deep learning models have slightly underestimated $PM_{2.5}$ concentrations and slightly overestimated the PNC concentrations as equally indicated by their respective MB and NMB values. Most of the ANN models for predicting PM_{10} trained with RF-GA selected variables show a slight tendency of overestimating the concentrations between10:00 and 16:00. Moreover, those trained with all the variables have shown the tendency for underestimating the concentrations between midnight and 09:00. The most accurate predictions are those of the BRT models in all the three cases while the least accurate predictions were the predictions of the PM_{10} and PNC by the SVM models trained with all the variables and with radial and linear kernels respectively. The plots also showed that most of the ANN and SVM models trained with RF-GA selected variables performed better than those trained with all the variables. The predictions of all the models are within the 95% confidence level of the observed concentrations as indicated by the shaded portion of the plots except in the case of the two SVM models mentioned.

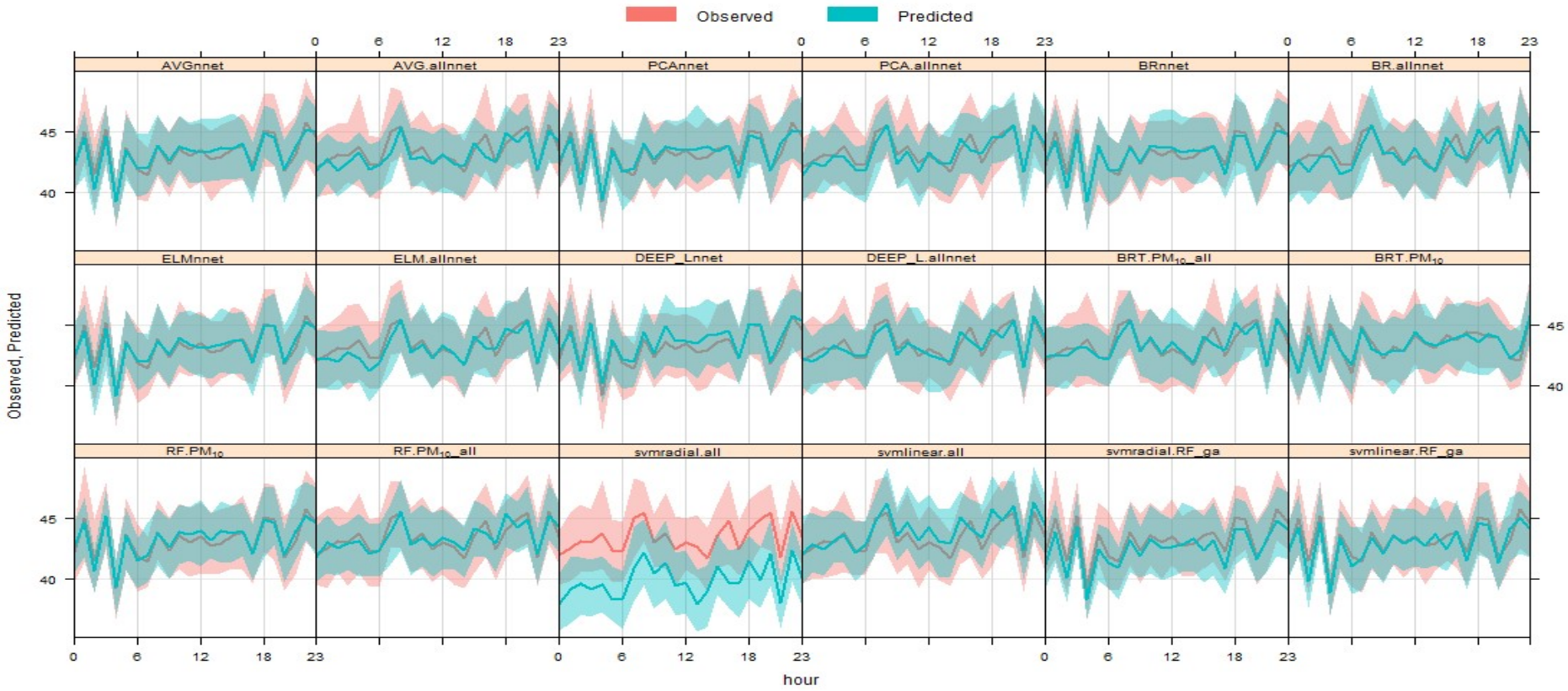


Figure 7.7 Hourly variation plots comparing the pattern of the PM_{10} ($\mu g/m^3$) prediction of the ML models and the observed PM_{10} ($\mu g/m^3$) concentrations.

Note: in Figure 7.7 the model names with an extension *"all"* indicates those that were trained with all the predictor variables while the rest were the ones trained with RF-GA selected variables. The *"nnet"* indicates that the model is trained with neural network algorithms.

The conditional quantile plots (Figure 7.8) show that the prediction of the ML models agreed very well with the observed particle concentrations. Among the PM_{10} prediction models, the performances of the models are nearly the same except for the AVG-MLP trained with RF-GA selected variables which show better data coverage than when trained with all the predictor variables. The distribution of the models' predictions also matched the distribution of the observed particle concentrations as indicated by the histograms in the figure. Most of the models predicted the higher concentrations (>100 $\mu g/m^3$) accurately with some showing less accurate prediction as indicated by the bumpy lines towards the region of higher concentrations. The spread of the prediction is also narrow in most cases showing good agreement with observations.

The AVG –MLP and PCA-MLP, BRT and RF models show slight improvement with the feature selection while BRNN, ELM, and deep learning showed slightly better prediction when trained with all the predictor variables. In Figure F.8, the $PM_{2.5}$ models showed much better data coverage than the PM_{10} models where only AVG – MLP shows significant improvement with the feature selection. The model performances in terms of accurate prediction of higher concentrations up to 100 $\mu g/m^3$ and prediction spread are largely the same.

The PNC models also performed well but with higher uncertainty in the prediction of higher concentrations (Figure F.9). The data coverage is not as good as in the case of $PM_{2.5}$ but better than the predictions of $PM_{10,}$ and they all have comparable performance in terms of prediction spread. The predictions of $PM_{2.5}$ and PNC have better data coverage than PM_{10}. However, the PNC models predicted the higher concentrations less accurately. The plots also reaffirm the superiority of the radial basis kernel over the linear kernels for the SVM models and the feature selection using RF-GA have positively affected the performance of the models. All the models showed less accuracy in the prediction of higher concentrations

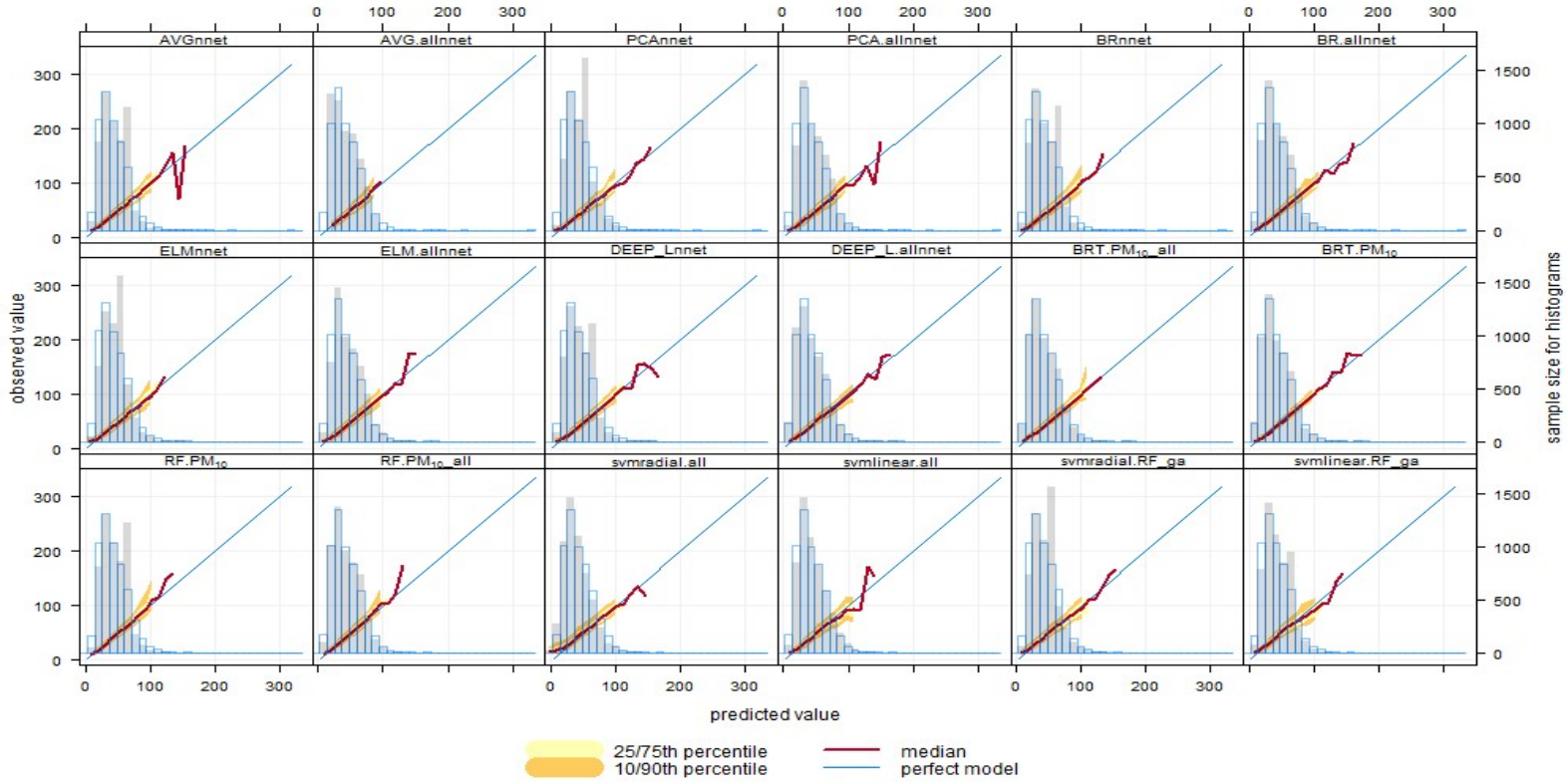


Figure 7.8 conditional quantile plots showing the agreement between the observed and Machine learning predictions of the PM_{10} ($\mu g/m^3$) concentrations.

Note: predicted value and observed value are modelled and observed PM_{10} concentrations respectively

especially in the case of the PNC models. The spread of the predictions around the perfect model line (blue line) shown by the shaded portions around the lines is narrow indicating high precision in the predictions of the models.

The scatter plots shown in Figures 7.9 and Figures F. 7.10 – F.11 show that all the ANN models performed extremely well in capturing the behaviour of the particles observations, and most of their predictions fall within the factor of two of the observations as indicated earlier by their FAC2 values in Table 7.8. However, the PM_{10} and $PM_{2.5}$ models trained with the RF-GA selected variables have fewer points outside the FAC2 boundaries than those trained with all the predictor variables except for deep learning algorithm where the reverse is the case.

The plots also showed that in the case of PM_{10} predictions, the BRT, Deep learning, AVG-MLP, PCA-MLP and ELM models trained with RF – GA variables have predicted the higher concentrations more accurately than the remaining models. The $PM_{2.5}$ models trained with the RF-GA selected variables also predicted the higher concentrations more accurately than those trained with all the variables. The disparity between the performances of the ML models for the prediction of PNC is very slim. However, it could be observed that the BRT, RF, SVM radial, BRNN, ELM and deep learning models trained with RF – GA variables show the more accurate prediction of the higher concentrations than the remaining models.

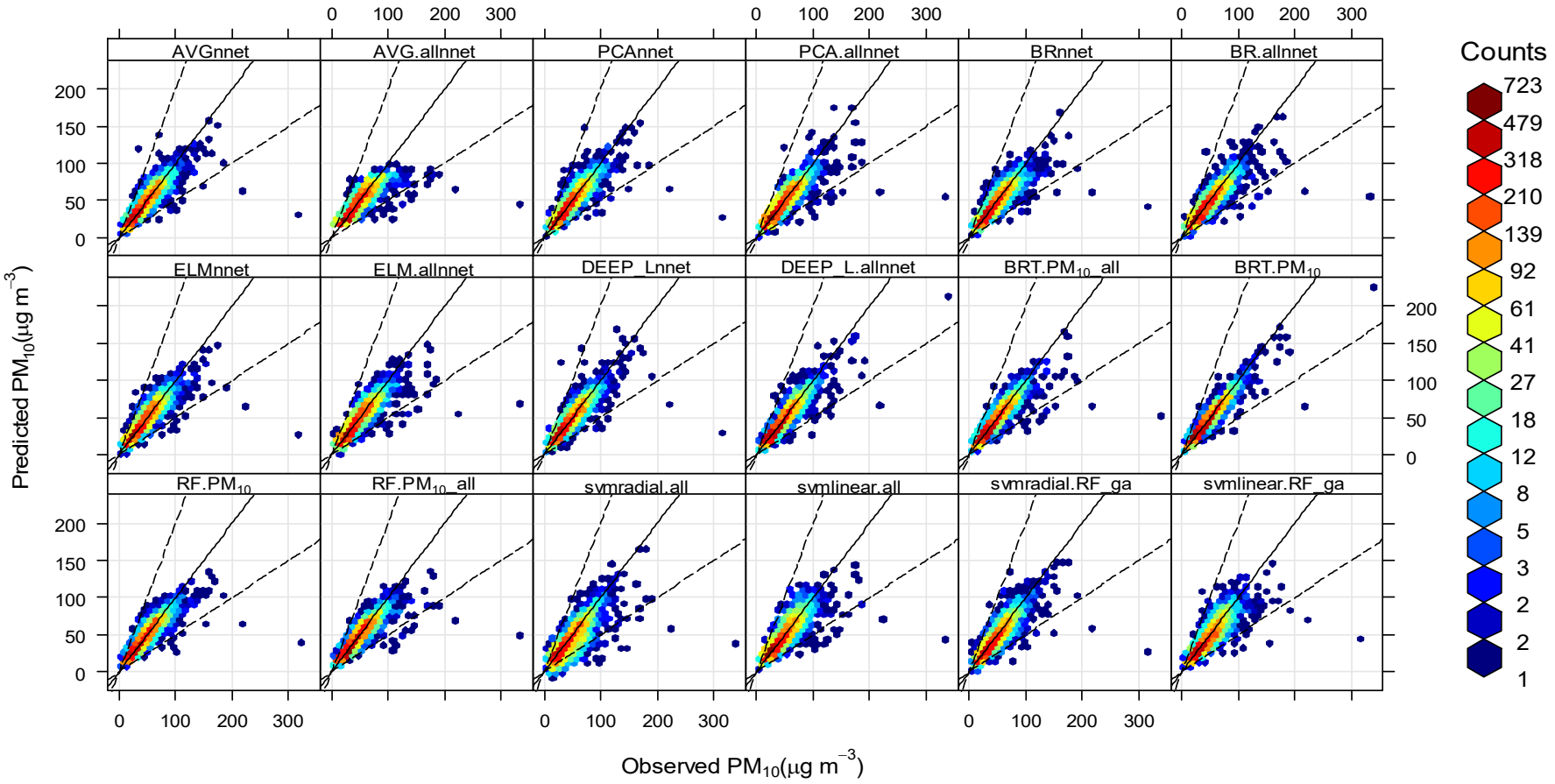


Figure 7.9 Scatter plots comparing the prediction of the ML models and the observed PM_{10} concentrations.

7.7 Seasonal Evaluation of the Performance of the ML Models Using Taylor's Diagrams

The analysis of the performance of the machine learning models so far focused on the overall performance of the models, and it is important to examine their seasonal performance as most pollutants are known to have seasonal variations. The seasonal evaluation was carried out using Taylor's diagram. This is a useful tool for comparing the performance of various models graphically. It shows three model performance metrics; the correlation coefficient, standard deviation, and centred RMSE. Taylor (2001) showed that it was possible to relate these statistics through the use of the law of cosines on a 2D graph.

The standard deviation measures the variations in the prediction of the models and the observed concentrations. It can be read from Taylor's diagram by taking the radial distance from the origin of the plots to the individual dots representing the models. The centred RMSE are represented by the concentric dashed lines starting from the point marked "observed". The correlation is represented by the graduated arcs located at the rightmost end of each plot. Each panel compared a seasonal performance of the models. The Taylor's diagram shown in Figure 7.10 displays the seasonal performance of all the machine learning models used in this work trained for the prediction of PM_{10}. The models slightly underestimated the variation in the observed PM_{10} concentrations in all the seasons most especially in winter where all the models are further away from the black dashed line passing through the standard deviation of the observed concentrations. The BRT and deep learning models have their variation closer to the observed than the other models in all the seasons.

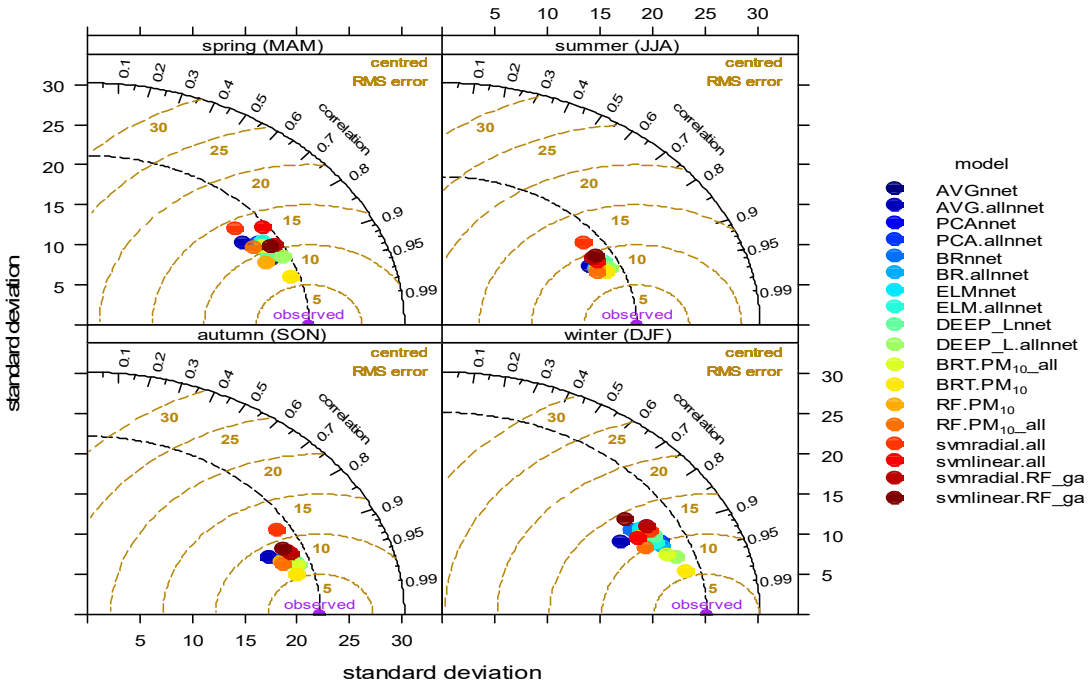


Figure 7.10 Taylor's plot comparing the performance of Machine learning models for predicting PM_{10}

The prediction error of most of the models measured by the centred RMSE is less in Summer and Autumn while a bit higher in Spring and Winter. Also, in all the seasons but Summer the BRT and deep learning models showed lesser errors than the other models. All the models also have higher correlation coefficients in Summer and Autumn than in the Spring and Winter with BRT and Deep learning leading in performance in all the seasons. The poorest performing models are the SVM models with linear kernels.

The Taylor's diagrams in Figure F.12 are drawn from the $PM_{2.5}$ observations and predictions. The plots show that the variation in the predictions of the models is more accurate than the variations estimated by the PM_{10} prediction models and the models performed similarly in all the seasons. The AVG – MLP trained with all the variables have grossly underestimated the variation in the $PM_{2.5}$ observations. Both the prediction errors and the correlations are similar in all the seasons.

The seasonal performance of the PNC prediction ML models is shown in Figure F.13. In summer and autumn, all the models except SVM linear with all the input variables have estimated the variations in the observed PNC concentrations very well. They also showed good and similar correlation and prediction errors in all the seasons except in winter where all the models performed poorly. The best performing models are also deep learning and BRT models.

7.8 Comparison The Performance of Machine Learning Models with Other Studies

The machine learning models used in this study including AVG-MLP, BRNN, BRT, Deep learning, ELM, PCA-MLP, RF, SVM radial, have shown superior performance to the statistical models discussed in chapter 6 having 99% (FAC2), 0.90, 0.89, 0.96, 0.91, 0.89, 0.89, 0.91, 0.88 (R) and NMB values between -0.02 and 0.01 for the prediction of PM_{10}. The average performance of the machine learning models was found to be 0.90,0.94,0.94 (R); 99% ,98% and 97% (FAC2); 9.2,4.5 and 10400 (RMSE) and 0.83,0.84,0.7 (IOA) for the prediction of PM_{10}, $PM_{2.5}$ and PNC. Comparing these results with similar studies, He et al. (2015) combined PCA with MLP for the prediction of PM_{10} concentrations in spring and discovered that PCA-MLP model performs poorer than ordinary MLP model (R = 0.80 vs R = 0.88). However, the PCA-MLP performs better on the testing data than MLP model (R = 0.66 vs R = 0.59). This is probably due to the fact that the PCA-MLP model combines the merits of PCA to overcome the overfitting problem and avoid problem the model becoming trapped in local optima, respectively. Ul-Saufie et al. (2013) evaluated the performance MLR and Feedforward backpropagation (FFBP) combined with PCA for predicting future (next day, next two-day and next three day) PM_{10} concentration in Negeri Sembilan, Malaysia. The best prediction model for next day PM_{10} concentration was the PCA-ANN with RMSE 11.1071, with accuracy of 0.9315 (IA), and 0.7812 (R^2). Singh et al. (2012) used linear and nonlinear modelling techniques to predict the urban air quality of the

Lucknow city (India). The methods include PLSR, Multivariate Polynomial Regression (MPR), and three ANN methods. The performance of all the three ANN models were comparable with the R values of 0.890 for the best ANN model.

Mishra et al. (2015) trained ANN, MLR and Neuro-Fuzzy (NF) techniques models for the haze hour forecasting in terms of $PM_{2.5}$ concentrations (more than 50 mg/m3) and relative humidity (less than 90%). The correlation coefficient between the predictions and the observations were 0.25, 0.53, and 0.72 for the ANN, MLR and NF models respectively. Vlachogianni et al. (2011) Forecasting models based on stepwise multiple linear regression (MLR) have been developed for Athens and Helsinki. The study concluded that the MLR is a useful tool for forecasting particulate matter and ANN models performed only slightly better. These studies have shown that the machine learning methods especially ANN and its hybrid versions performed better than the linear models pointing to the inability of the linear models to estimate the nonlinear relationships that exist in air quality variables. Taspinar and Bozkurt (2014) selected predictor variables for the training of ANN using stepwise regression approach and the resulting ANN-MLP appeared to be promising with R2 up to 0.69 and index-of-agreement up to 0.79. The study concluded that local monitoring systems associated with ANN model predictions may be a sound way to develop embedded online systems for public health. Although some different versions of the hybrid ANNs are used in this study the results obtained are largely better than the aforementioned studies despite using smaller time units (hourly).

7.9 Summary

The application of three major machine learning techniques in the modelling of roadside particulate matter have been examined and the results obtained showed that all the methods can be used for training models for the prediction of the PM_{10}, $PM_{2.5}$ and PNC

concentrations. The BRT, RF, ELM and Deep learning algorithms were found to be most suitable for this purpose due to their predictive accuracy and faster training speed. The application of feature selection procedure before the training of the models was found to be desirable in reducing the cost and complexity of the models to be developed. The feature selection is most suitable for the traditional MLP neural networks because they show some improvement in performance when trained with the selected variables. The models performed slightly better in predicting $PM_{2.5}$ and PNC than the PM_{10} concentrations.

Among the machine learning methods considered, BRT shows consistent and outstanding performance for all the particle metrics used and has the additional capability of producing partial dependence plots which provide more information about the interactions between the predictor variables and response variables during the modelling process. Both BRT and RF can do feature selection themselves without using external algorithms. This property gave them an advantage over traditional neural network algorithms. Deep learning and the ELM algorithms though under active development are purported to be more sophisticated algorithms that can handle a variety of cases including image processing, character recognition, and speech recognition and even in language translation. However, in this case, they did not show much difference in performance than the traditional neural network and the tree-based models.

In Chapter 8, the algorithms with better prediction accuracy, popularity, ease of application and/or easily accessible through open source software will be used in predicting the effect of a hypothetical air quality management scenario. Also, their predictive performance will be compared with the performance of the ADMS – Roads (operational model). Such comparison will give a clue on whether the machine learning models can be accepted as operational models.

Chapter 8

Application and Evaluation of Machine Learning Models for Air Quality Management

8.1 Introduction

Urban air pollution is increasingly becoming the major environmental concern in major cities around the world. The ever increasing population of the major urban areas has resulted in an increase in activities and higher demands for energy and transportation. These factors contribute significantly to urban air pollution emanating from major roads that are often congested because of high traffic demand. The urban air pollution can be managed through careful planning and execution of urban air quality management (UAQM) which hinges on several elements. The key components of UAQM consist of clear definition of objectives and standards, a well-designed air quality monitoring network and effective air quality modelling. These components help in designing air quality control strategies and evaluating their effectiveness. Air quality modelling, in particular, is an important aspect of the UAQM as it helps in taking a decision on major issues relating to the budget for the UAQM and predicting the likely effects of the control strategies to be implemented.

The major aim of this chapter is to evaluate the application of the machine learning methods discussed in chapter 7 in air quality management and compare their performance with that of the ADMS-Roads. First, the temporal and spatial prediction capabilities of the machine learning models in comparison with the ADMS-Roads were evaluated. Also, the application of the models in the management of roadside PM_{10} and $PM_{2.5}$ concentrations. To achieve these objectives, some of the machine learning models were used to predict the concentrations at various roadside monitoring stations. Also, a hypothetical air quality management scenario called Euro4/VI was conceptualised, and its future effect was

predicted using the models. The scenario suggests that only petrol and diesel vehicles meeting EuroIV/4 and EuroVI/6 design specifications respectively would be allowed to enter the study area in 2011 and 2015 in the case of PM_{10} and 2012 and 2015 in the case of $PM_{2.5}$. The EuroIV/4 and EuroVI/6 represent the European emission standard for vehicles. The Roman numerals indicate standards for heavy-duty vehicles while the Arabic numbers indicate a standard for light-duty vehicles.

The study area consists of the Westminster City and the major roads around the selected monitoring stations outside Westminster City (see Chapter 4). The selected machine learning methods and the most widely used operational air quality model in the UK, i.e. "ADMS-Roads" were used to test this scenario in 2011 and 2015 for PM_{10} concentrations, and 2012 and 2015 for $PM_{2.5}$ concentrations. The performance of the selected machine learning models was then compared with the performance of the ADMS-Roads model.

In the rest of the chapter, Section 8.2 discusses the data preparation to allow for missing data. Section 8.3 briefly describes the hypothetical air quality management scenario. In Section 8.4 the performance comparison of the spatial and temporal predictions of the machine learning and the ADMS-Roads models is presented. Section 8.5 shows the performance of the models in predicting a daily pollutant and air quality indexes. The comparison of the performance of the models in predicting the effects of the Euro4/VI scenario on the PM_{10} and $PM_{2.5}$ concentrations is presented in Section.8.6. Sections 8.4 and 8.5 were presented to establish the performance of the machine learning models in comparison with the ADMS-Roads before their application in the evaluation of the hypothetical scenario since there is no actual data on the scenario to evaluate the performance of the models. Section 8.7 presents a summary of the findings of the chapter.

8.2 Data Preparation for Air Quality Management Study

Air quality monitoring is a complex exercise that involves both human and machine interactions, therefore, it is far from being perfect. These imperfections usually result in incomplete data. There are generally two ways of handling missing data in machine learning modelling, (1) to ignore any row in the data associated with the missing value or (2) to use one of the missing data imputation methods to impute the missing values. The first option causes significant loss of data while the second option maximises the use of the available data. However, the disadvantage of the imputation is that it adds noise to the data thereby making it difficult to achieve good generalisation. To avoid this shortcoming, a data imputation algorithm involving powerful machine learning algorithms was used for the missing data imputation in this study. Before the imputation, the hourly average traffic volume was disaggregated into eight traffic categories. The categories were Petrol car, Diesel car, Taxi, LGV, Rigid, Artic, Bus and coach, and Motorcycle. They were estimated based on the UK traffic composition projections. Moreover, their corresponding emission rates were estimated using LAQM Emission Factor Toolkit (EFT) version 6.0.1(DEFRA, 2015b). The emission rates were then used as part of the inputs of all the three model types. The reason for the traffic volume disaggregation was to serve as a medium through which changes to emissions can be conveyed to the models, and therefore, the response of the models to the emission changes will allow them to be used as management tools for measuring traffic-related air quality control scenarios.

8.3 Euro 4/VI Air Quality Management Scenario

An air quality management scenario is required here to test the performance of the models in predicting the impact of such management options in future. A hypothetical scenario which is called here as Euro4/VI scenario was conceptualised to test the use of the machine learning models in real life situation and compare their performance with that of the ADMS-

Roads. The name of the scenario was selected according to the Euro vehicle emission standard notations for heavy - duty vehicles standards (Euro I, Euro II, Euro III, Euro IV, Euro V and Euro VI) and the notation for light - duty vehicle standards are Euro 1, Euro 2, Euro 3, Euro 4, Euro 5 and Euro 6.

The Euro4/VI scenario was conceived from the proposal of Ultra Low Emission Zone (ULEZ) in London which will take effect from September 2020 (TfL, 2016). All vehicles entering the ULEZ will need to meet the proposed emission standards (ULEZ standards) or pay a certain amount of money for travelling in the ULEZ area. The ULEZ area will cover the same area as the current London Congestion Charge Zone (CCZ). Coincidentally, the ULEZ contain most of the air quality monitoring stations used in this research. The ULEZ when implemented, is expected to reduce exhaust emissions of NO_2 and particulate matter PM_{10} and $PM_{2.5}$ in central London. The ULEZ proposal provides that only vehicles are meeting Euro4 for petrol vehicles and Euro 6/VI for diesel vehicles can travel the area without paying a daily charge. However, the standards for London taxis will be covered as part of the licensing regime. There are also specific rules for buses (London Assembly, 2014, TfL, 2016). Therefore, it is on the basis of the provisions of ULEZ, Euro4/VI scenario suggests that only vehicles are meeting Euro4 for petrol vehicles and Euro6/VI for diesel vehicles would be allowed to enter the study area (Westminster City).

This hypothetical scenario was aimed at testing the use of the machine learning models in real life applications and compare their performance with the performance of an operational air quality model (i.e. ADMS-Roads). The emission standard restriction proposed in the scenario was implemented through Emission Factor Toolkit (EFT) version 6.0.1(DEFRA, 2015b).

The projected Euro composition in the emission factor toolkit was altered to reflect the restriction on the minimum vehicle standard of Euro6/VI and Euro4 for diesel and petrol vehicles respectively. The scenario assumed that there would be no any changes in the number of vehicles entering the study area and the vehicles outside the study area will have no effect on the estimated emissions for simplicity. Therefore, the emission rates were estimated based on the assumption that all the vehicles have met the minimum standard imposed. The estimated emission rates were then used in the ADMS-Roads for the prediction of the particle pollutants (PM_{10} and $PM_{2.5}$).

For the machine learning models, the difference between the emission rates estimated with and without the scenario was obtained. The difference was then subtracted from the hourly emission rates in the predictor variables data to reflect the changes due to the Euro4/VI scenario assumptions.

8.3.1 Estimation of Emission Rates

The emission standard restriction proposed in the scenario was implemented through Emission Factor Toolkit (EFT) version 6.0.1(DEFRA, 2015b). The EFT requires vehicle counts (veh/hr), average speed (km/hr), link length, road type, road name, the projection of vehicle composition and Euro traffic composition (Euro1, Euro2, Euro III, Euro6, etc.) as inputs. The fleet composition data in London for motorways, central, inner and outer areas was used in the EFT to estimate the emission rates as g/km/s, g/km, or kg/year or tonnes/year from the total traffic for PM_{10} and $PM_{2.5}$ including PM_{10} and $PM_{2.5}$ from tyre and brake wear and road abrasion emission sources.

The road traffic projections in Table 8.1 show that the percentage of petrol car is decreasing while the percentage of a diesel car is increasing. Also, the percentage of electric vehicles are increasing. An increase in the percentage of diesel cars might have negative impacts on

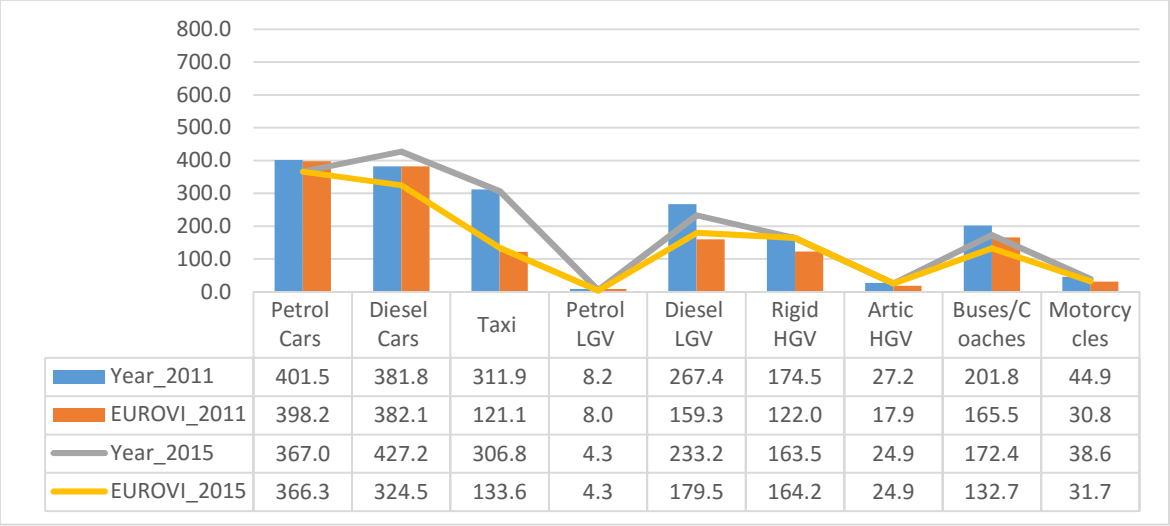
the particle emissions since they have higher particle emissions (Lewis et al., 2015)There were no significant changes in the percentage of the other vehicles between 2011 and 2015.

Table 8.1 Projected traffic composition for central London (NAEI, 2014)

Year	Year_2011	Year_2012	Year_2015
Electric car	0.0%	0.0%	0.1%
Petrol car	40.1%	38.5%	34.0%
Diesel car	23.0%	24.6%	29.0%
Taxi (black cab)	12.4%	12.4%	12.4%
Electric LGV	0.0%	0.0%	0.1%
Petrol LGV	0.6%	0.4%	0.3%
Diesel LGV	11.2%	11.4%	11.4%
Rigid	3.1%	3.1%	3.1%
Artic	0.4%	0.4%	0.4%
Bus and coach	4.2%	4.2%	4.2%
Motorcycle	5.1%	5.1%	5.1%

The implementation of the scenario in 2011 resulted in the PM_{10} emission reduction of 3.3 kg/yr and a slight increase of 0.3kg/yr for petrol and diesel cars respectively as shown in Figure 8.1. The decrease in the PM_{10} emission for petrol and diesel LGV were 1kg/yr and 108.1kg/yr respectively. The Taxi emissions were reduced by 190.8 kg/yr. The reduction in Rigid and Articulated HDVs was 52.5 kg/yr and 9.3 kg/yr respectively. The PM_{10} emission for the buses and coaches was estimated to be reduced by 14.1 kg/yr. The total PM_{10} emission reduction due to the scenario in 2011 was 414.7 kg/yr (see Figure 8.1). In the year 2015, there was no decrease in the PM_{10} emission of petrol LGV, Rigid HGV and Articulated HGV. A slight decrease of in the 0.7 kg/yr in the PM_{10} emissions of the petrol cars and rigid HGV due to the scenario was observed. The PM_{10} emissions from the London Taxi show the highest PM_{10} emissions reduction of 173.2 kg/yr due to the scenario. The

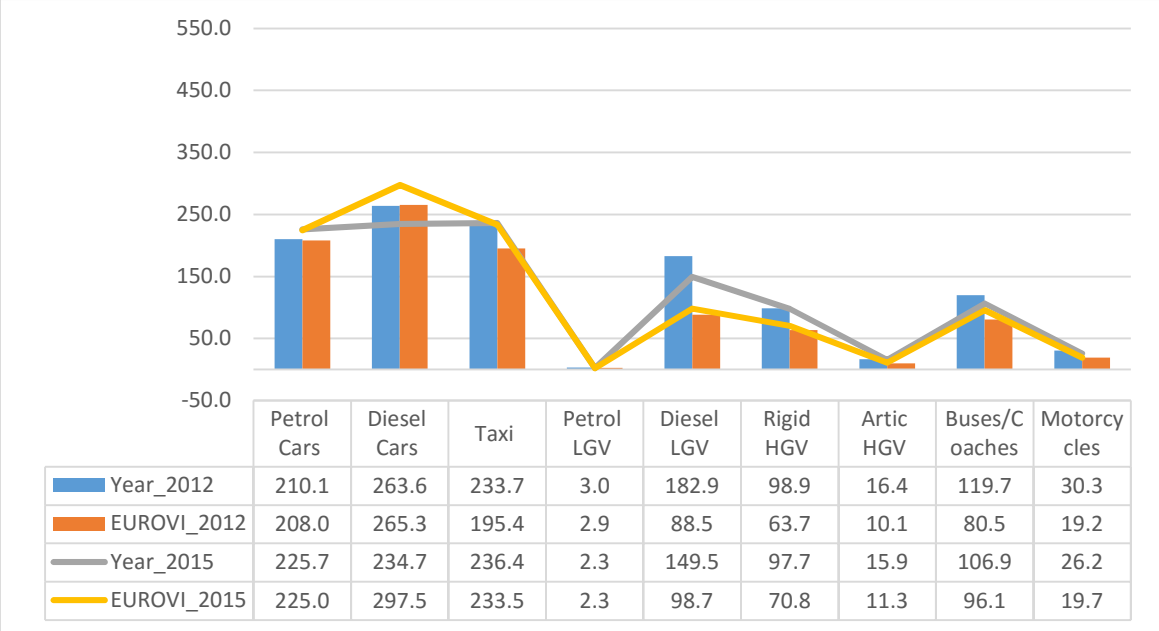
PM_{10} emissions of the diesel LGV and buses/coaches were reduced by 39.7 kg/yr and 53.7 kg/yr respectively.



	Petrol Cars	Diesel Cars	Taxi	Petrol LGV	Diesel LGV	Rigid HGV	Artic HGV	Buses/Coaches	Motorcycles
Year_2011	401.5	381.8	311.9	8.2	267.4	174.5	27.2	201.8	44.9
EUROVI_2011	398.2	382.1	121.1	8.0	159.3	122.0	17.9	165.5	30.8
Year_2015	367.0	427.2	306.8	4.3	233.2	163.5	24.9	172.4	38.6
EUROVI_2015	366.3	324.5	133.6	4.3	179.5	164.2	24.9	132.7	31.7

Figure 8.1 Estimated annual PM_{10} emission rates (kg/yr) with and without Euro4/VI scenario for MY1.

For the $PM_{2.5}$ emissions (see Figure 8.2), the results of the implementation of the scenario in 2012 and 2015 followed the same trend as in the case of PM_{10} with one important difference in the case of diesel car where there was increase of 1.7 kg/yr and 62 kg/yr in $PM_{2.5}$ emissions respectively. Overall, the implementation of the scenario resulted in higher reductions in the emissions of Taxis, Diesel LGV and Buses/Coaches.



	Petrol Cars	Diesel Cars	Taxi	Petrol LGV	Diesel LGV	Rigid HGV	Artic HGV	Buses/Coaches	Motorcycles
Year_2012	210.1	263.6	233.7	3.0	182.9	98.9	16.4	119.7	30.3
EUROVI_2012	208.0	265.3	195.4	2.9	88.5	63.7	10.1	80.5	19.2
Year_2015	225.7	234.7	236.4	2.3	149.5	97.7	15.9	106.9	26.2
EUROVI_2015	225.0	297.5	233.5	2.3	98.7	70.8	11.3	96.1	19.7

Figure 8.2 Estimated annual $PM_{2.5}$ emission rates (kg/yr) with and without Euro 4/VI scenario for MY1

The emission rates determined with the scenario restrictions were then used in the training of machine learning models for the prediction of the PM_{10} and $PM_{2.5}$. Also, the emission rates were used in the ADMS-Roads for the same purpose. The application years were 2011 and 2015 for PM_{10}, and 2012 and 2015 for $PM_{2.5}$. These years were selected based on the availability of the data at the monitoring stations in the study area.

The scenario was first implemented in the City of Westminster in London where there were only two sites (see Figure 8.3) with the sufficient PM_{10} data and one site with sufficient $PM_{2.5}$ data. The emission inventory used in the ADMS modelling included all the major roads within the City of Westminster. For the sites outside Westminster City, only the roads adjacent to the monitoring sites were considered (see Figure 4.1). However, for the machine learning models, traffic data obtained from Marylebone Road was taken as the average traffic data in the area as it does not require traffic data from all the roads as in the case of ADMS-Roads. The details of the ADMS-Roads and machine learning modelling processes can be found in Chapter 3.

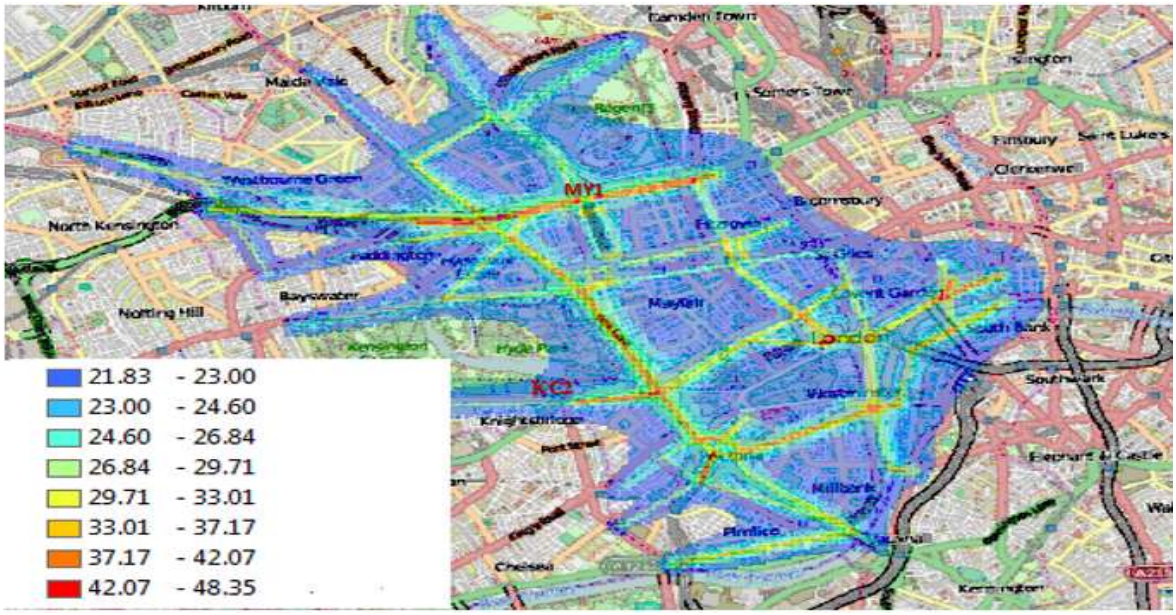


Figure 8.3 Map showing the levels of ADMS-Roads modelled concentrations of PM_{10} ($\mu g/m^3$) in Westminster City. Note: MY1 and KC2 in Figure 8.3 are the air quality monitoring units.

8.4 Comparison Between the Performance of Machine Learning Models and ADMS-Roads in the Predictions of PM_{10} and $PM_{2.5}$ (without scenario)

This section presents the comparison between the performance of the machine learning models and the ADMS-Roads model in the spatial and temporal predictions of PM_{10} and $PM_{2.5}$ concentrations. Also, the performance of the different machine learning models in spatial and temporal prediction of PNC concentrations was compared. ADMS-Roads was not used for PNC predictions because it has no provision for PNC prediction. The performance of the models was evaluated using the Coefficient of Efficiency (CoE), Index of Agreement (IoA), and a FACtor of two (FAC2) methods. Others are the coefficient of correlation (R), Root Mean Squared Error (RMSE), Normalised Mean Bias (NMB), Normalised Mean Gross Error (NMGE). The conditional quantile plots, time variation plots and bivariate polar plots were also used to compare the performance of the models. For the details of the evaluation methods see Section 2.11. It is imperative to use more than one performance criteria because the model to be evaluated might have more than one objective

or the end user might want to use a different performance criterion to evaluate the model accuracy. Also, most of the evaluation methods have their shortcomings, therefore using many evaluation methods would help in determining the strengths and weaknesses of the models.

8.4.1 Statistical Performance

The performance of the ANN and BRT is similar as indicated by most of the performance statistics (see Table 8.2). However, the SVM show lesser performance than the ANN and BRT methods as shown by the average COE values of 0.45, 0.54 and 0.65 for PM_{10}, $PM_{2.5}$ and PNC predictions. All the remaining metrics shows this trend. The predictions of the machine learning models are much better than what could be explained by the mean of the measured particle concentrations as indicated by the average COE values (0.53 – 0.63). The only exception is in the case of SVM for the prediction of PM_{10}. The average values of *R* (more than 0.8) and *IOA* (more than 0.7) for the machine learning models show that their predictions have better agreement with the PM_{10} and $PM_{2.5}$ observations than the predictions of the ADMS-Roads model with average values of *R* between 0.67 and 0.75. The FAC2 values of the machine learning models indicate that more than 95% of their predictions fell within the factor of two of the measure particle concentrations at most of the sites while the predictions of ADMS-Roads fell below 90% at the majority of the sites.

Also, it could be seen that the predictions of the ANN and BRT models have shown smaller prediction bias than the ADMS-Roads as indicated by the NMB (0.01 – 0.08) against NMB (0.1 – 0.22). Also, the ANN and BRT show average RMSE values of 10.05 $\mu g/m^3$- 10.1, 4.67 $\mu g/m^3$ - 4.8 $\mu g/m^3$ and 13532 - 13783 for PM_{10}, $PM_{2.5}$ and PNC predictions respectively. However, the ADMS-Roads show much higher RMSE values of 15.3 $\mu g/m^3$ and 8.72 $\mu g/m^3$ for the PM_{10} and $PM_{2.5}$ predictions respectively. The bias values of the ADMS-Roads and the SVM models for PM_{10} predictions are dominantly negative which signifies under

prediction, except at BT4 and KC5, where the ADMS-Roads model shows over prediction indicated by positive NMB values. The ADMS-Roads model overestimated the $PM_{2.5}$ concentrations at all the sites except at MY1 where it shows underestimation. The R-values measure the correlation between the predictions of the models and the particle observations and the predictions of the machine learning models show higher correlations than the predictions of the ADMS-Roads. The performance of the SVM models for the prediction of PM_{10} is less than the performance of ANN and BRT models while in the case of $PM_{2.5}$ predictions, all the machine learning models, have shown similar performance (see Appendix H for the detail performance statistics).

Table 8.2. Test performance of the machine learning and ADMS-Roads models

Pollutants →		PM_{10}		$PM_{2.5}$		PNC	
Model	**Performance Statistics**	**Lower – Upper**	**Average for all sites**	**Lower – Upper**	**Average for all sites**	**Lower – Upper**	**Average for all sites**
ADMS	FAC2	0.8 - 0.95	0.86	0.70 - 0.92	0.83		
ANN	FAC2	0.84 - 0.99	0.97	0.93 - 0.98	0.95	0.75 - 0.97	0.91
BRT	FAC2	0.82 - 1.00	0.97	0.94 - 0.99	0.97	0.75 - 0.97	0.89
SVM	FAC2	0.84 - 0.99	0.95	0.91 - 0.97	0.95	0.66 - 0.99	0.9
ADMS	NMB	0.22 - 0.14	-0.1	0.12 - 0.56	0.21		
ANN	NMB	0.07 - 0.11	0	0.02 - 0.12	0.03	0.02 - 0.11	0.08
BRT	NMB	0.03 - 0.15	0.02	0.01 - 0.04	0.02	0.01 - 0.21	0.14
SVM	NMB	0.26 - 0.04	-0.13	0.01 - 0.06	0.01	0.01 - 0.10	0.06
ADMS	R	0.4 - 0.84	0.67	0.57 - 0.91	0.75		
ANN	R	0.45 - 0.95	0.81	0.82 - 0.95	0.87	0.89 - 0.93	0.91
BRT	R	0.43 - 0.95	0.81	0.83 - 0.95	0.88	0.91 - 0.93	0.92
SVM	R	0.43 - 0.95	0.79	0.81 - 0.95	0.87	0.90 - 0.94	0.92
ADMS	COE	0.10 - 0.18	0.16	0.35 - 0.57	0.41		
ANN	COE	0.31 - 0.71	0.53	0.37 - 0.70	0.54	0.57 - 0.68	0.63
BRT	COE	0.35 - 0.73	0.56	0.45 - 0.68	0.56	0.50 - 0.66	0.58
SVM	COE	0.33 - 0.70	0.45	0.44 - 0.70	0.54	0.61 - 0.68	0.65
ADMS	RMSE	9.16 - 20.38	15.3	5.66 - 11.49	8.72		
ANN	RMSE	4.69 - 19.17	10.12	4.15 - 6.30	4.80	9900 - 19115	13532
BRT	RMSE	4.48 - 20.98	10.05	3.47 - 6.33	4.67	10570 - 17644	13783
SVM	RMSE	4.91 - 19.17	11.44	3.50 - 6.74	4.84	8827 - 14058	11873
ADMS	NMGE	0.25 - 0.42	0.33	0.32 - 0.63	0.40		
ANN	NMGE	0.13 - 0.38	0.2	0.17 - 0.26	0.20	0.22 - 0.27	0.24
BRT	NMGE	0.14 - 0.44	0.19	0.16 - 0.22	0.19	0.22 - 0.29	0.27
SVM	NMGE	0.13 - 0.37	0.22	0.17 - 0.24	0.20	0.19 - 0.33	0.23
ADMS	IOA	0.51 - 0.71	0.57	0.22 - 0.79	0.53		
ANN	IOA	0.58 - 0.86	0.75	0.69 - 0.85	0.77	0.78 - 0.84	0.81
BRT	IOA	0.52 - 0.86	0.75	0.73 - 0.86	0.78	0.75 - 0.83	0.79
SVM	IOA	0.59 - 0.85	0.71	0.72 - 0.85	0.77	0.80 - 0.84	0.83

Note: The first and the second columns display the names of the models and the performance statistics respectively. The rest of the columns show the upper, lower and average values of the performance statistics for all the sites. The third and the fourth columns represent the statistics for PM_{10} concentrations while the fifth and the sixth columns represent the statistics for $PM_{2.5}$ concentrations

Key: Coefficient of Efficiency (CoE), Index of Agreement (IoA), FACtor of two (FAC2), the coefficient of correlation (R), Root Mean Squared Error (RMSE), Normalised Mean Bias (NMB), Normalised Mean Gross Error (NMGE).

Mishra et al. (2015) applied ANN, Neuro-fuzzy and MLR in forecasting $PM_{2.5}$ concentrations during haze episodes in Delhi, India and found that Neuro-fuzzy outperformed ANN models with R values 0.53 and 0.72, IOA 0.78, and 0.80 and FAC2 0.81 and 0.84. The ANN models in this study performed better than those obtained by Mishra et

al. (2015) considering the results from the majority of the sites (see Table H.2). The average R, IOA and FAC2 values for all the sites for the ANN are 0.87, 0.77 and 0.95 respectively. Also, Wang et al. (2015a) compare the use of a hybrid model combining wavelet neural network and genetic algorithm (GA -WNN) for predicting 5-min series of carbon monoxide (CO) and fine particulate matter ($PM_{2.5}$) concentrations in proximity to an intersection in comparison with ordinary ANN. The study found that the ANN model has R, IOA, RMSE and NMB values 0.88, 0.93, 8.2 and -0.01 respectively. This is also similar to the performance of the ANN model obtained in this study.

8.4.2 Graphical Performance Evaluation

This section presents the evaluation of the models using graphical tools that include: Taylor's diagrams, conditional quantile plots, time variation plots and bivariate polar plots. This type of assessment is very important as it reveals information about the weakness and the strength of the models in capturing the extremely high and low concentrations since they the most important for the policy makers and the public (Fei et al., 2005). This analysis will, in turn, boost the confidence of the air quality modellers in using the machine learning methods for air quality predictions and management.

8.4.1.1 Conditional Quantile Plots

The conditional quantile plots shown in Figure 8.4 were constructed by partitioning the prediction of the models into certain intervals such that for each interval, the median and the spread in the concentrations (25 – 75th and 10 – 90th percentiles) of each interval and its corresponding observations are calculated. Conditional quantile plots can be used to find the behaviour of the models at certain intervals of the observations. The conditional quantile plots revealed that the predictions of the ANN models had shown more agreement with the observed PM_{10} concentrations than the remaining models while the predictions of the ADMS-Roads model have shown more data coverage (lower – upper range), but with wider

spread around the perfect model line (blue). Moreover, in most cases, the ADMS-Roads predictions (red lines) of the higher concentrations intervals deviate from the blue lines which suggest less accurate prediction. For example, at BT4 only PM_{10} concentrations up to 30µg/m^3 were accurate but beyond that, the model tends to overestimate the concentrations. The BRT models also show substantial data coverage, but its prediction of the PM_{10} concentrations in the range of 100 µg/m^3 are less accurate than those of the ANN and SVM models.

In the case of the prediction of $PM_{2.5}$ concentrations (see Table H.2), the performance of the machine learning models was similar at GR9, MY1, and TH4 where they show good agreement with the observation intervals. However, the predictions show slight disagreement with the higher concentrations. Also, the predictions of the machine learning models at BT4, GR8 and HK6 are much better than the predictions of the ADMS-Roads model considering the amount of deviation of the model lines from the perfect model lines. The ADMS-Roads model grossly overestimated almost all the observation intervals, but the machine learning models could only have problems with the $PM_{2.5}$ concentrations above 50µg/m^3. The PNC models have shown similar performance, and their overestimation mostly lies in the range of 100,000 number/cm^3 where they show large disparity with the observations.

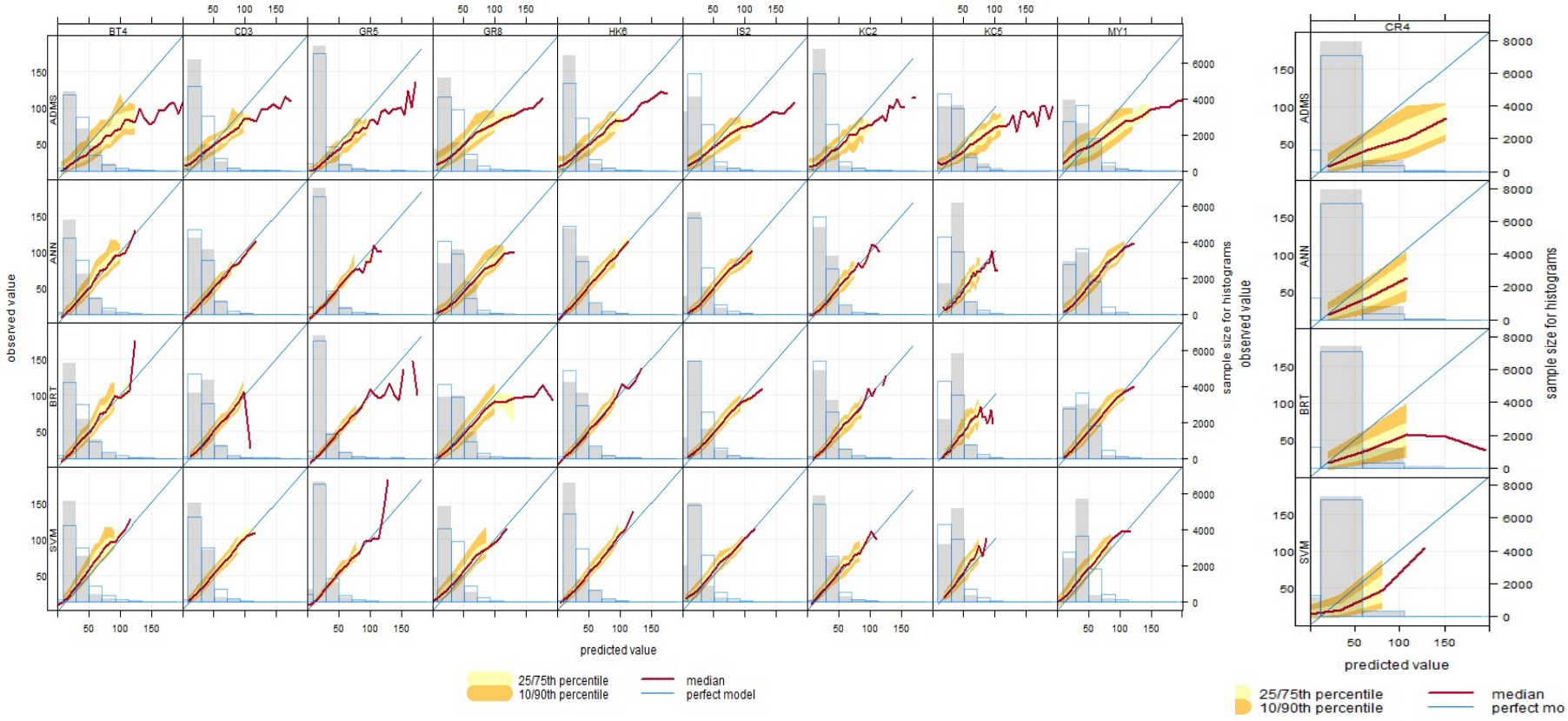


Figure 8.4 Conditional quantile plots showing the prediction performance of the models at 10 London sites (see Table 4.1)

Note: The BT4, CR4, GR5, GR8, HK6, IS2, KC2, KC5 and MY1 represent the air quality monitoring sites.

8.4.1.2 Time Variation Plots

The conditional quantile plots have shown that the models have problems with the prediction of higher concentrations, but it is not known whether these problems have relationships with the time of the day. Therefore, it is vital to establish the time when these variations occur. Time variation plot is one of the tools that can reveal more about the relationships between the time and the prediction of the models.

The time variation plots shown in Figure 8.5 indicates that the ADMS-Roads and SVM models under-predicted the hourly PM_{10} concentrations in almost all the sites except at BT4 and KC5 where the ADMS-Roads model overestimates most of the hourly concentrations. The over or under prediction by the two models did not vary with the time of the day at most of the sites except at BT4 and MY1 where they tend to overestimate night and early morning concentrations.

At site KC5, the data collected were daily averages repeated for each hour, therefore, evaluating the performance of the models on an hourly basis might be meaningless as shown in Figure 8.5 where the models tried to show some hourly variations. The predictions of the ANN and BRT have demonstrated a higher degree of agreement with the observation, and they captured more accurately the hourly pattern of the observations. For example, the models indicate morning and afternoon peaks shown in the observations. Also, ANN and BRT overestimated the early morning concentrations i.e. between midnight and 6:00 am. All the models performed better in predicting PM_{10} at GR5 and performed poorly at BT4 and KC5. ADMS-Roads overestimated the hourly $PM_{2.5}$ concentrations in all the sites except at MY1 where it underestimated the concentrations, and its prediction accuracy did not vary with time. The prediction of the machine learning models has shown a higher degree of agreement with the $PM_{2.5}$ observations, and they captured more accurately the hourly pattern of the observations (see Appendix F - Figure F.12).

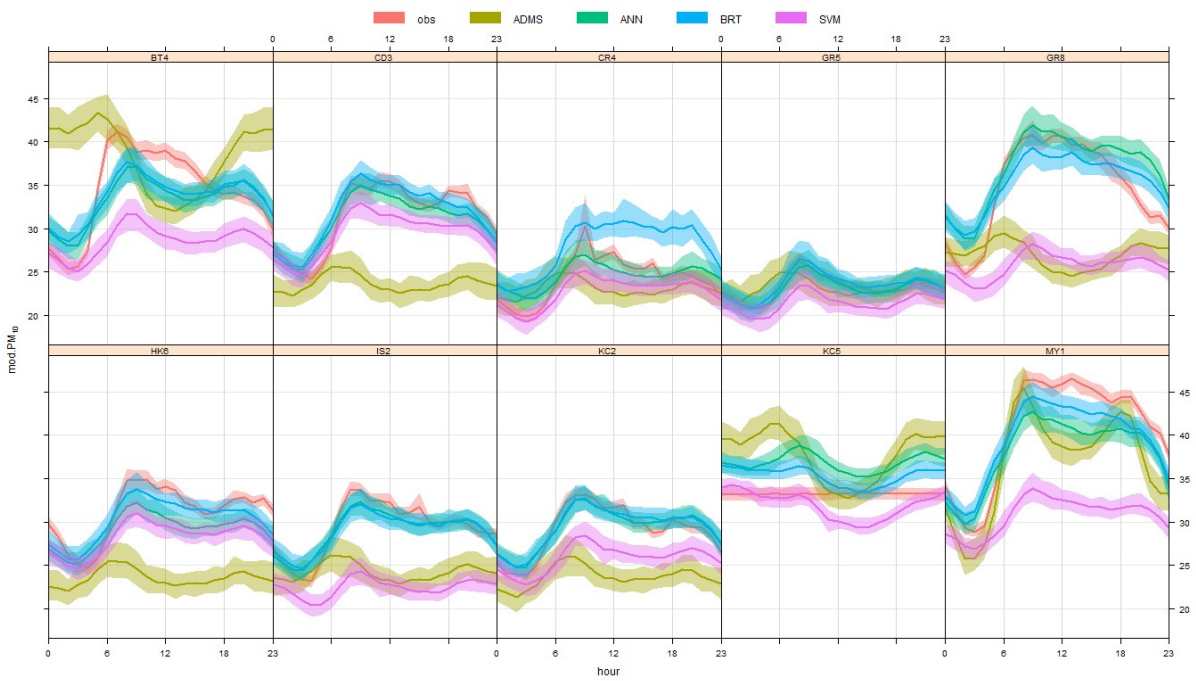


Figure 8.5 Hourly time variation plots of the observed and predicted PM_{10} concentrations

Also, there is a close agreement between the observed PNC concentrations and the prediction of the models, and the predicted PNC concentrations captured the hourly pattern of the observations (see Appendix F - Figure F.13). However, the time variation plots revealed that the models have more problems predicting concentrations during peak periods as the gap between the observations and the predictions mostly widened at those times.

8.4.1.3 Bivariate Polar Plots

The analysis of hourly predictions thus far has shown that the machine learning models performed slightly better, and their predictions have consistently shown more agreement with the observed concentrations than the predictions of the ADMS-Roads model. However, there is no information regarding the relationship between the emission sources and the predictions. Although it can be argued that the predicted concentration profiles are similar to the hourly traffic profiles as shown in Figures 8.5. One way of establishing that fact is to consider the behaviour of the predicted concentrations and the wind speeds and wind directions. The wind plays an essential role in the dispersion of pollutants in the streets and

the urban environment at large. Bivariate polar plots are used to reveal these relationships, and they may provide information about the likely source of the pollutants (Carslaw et al., 2006) and the likely effect of the street canyon geometry on the concentrations where they exist. They also highlight the evidence of whether the pollution source has been modelled or not (Carslaw D et al., 2013).

Considering the bivariate polar plots shown in Figure 8.6, it could be observed that the machine learning models captured the general pattern of the relationships between the observed concentrations and the wind directions and wind speeds at most of the sites. For example, the predictions of the ANN and BRT models captured the higher concentrations associated with low and higher winds in most of the sites.

However, the ADMS-Roads model has performed relatively less accurately. It overestimated the higher concentrations associated with the low easterly winds at BT4, KC5 and MY1 sites and it did not show the higher concentrations associated with the winds from the south-west directions at GR8 and MY1. The MY1 monitoring site is located on the southern side of Marylebone Road that lies along the axes of 75^0 and 225^0. On the bivariate polar plot of MY1, it could be seen that there were higher concentrations along the same axes especially towards the south-west, which is evidence of Canyon recirculation vortices that deliver most of the concentrations on the leeward side of the street canyon. The machine learning models reproduced the same pattern, but the SVM model slightly underestimated this pattern in the prediction of PM_{10} concentrations.

At GR8 where the road lies along a North-west/South-east axis, and the monitoring station is located to the north of the road, there was evidence of high PM_{10} concentrations associated with the winds coming from the west and south-west. This association shows that most of the concentration recorded at this station were from the road, and there was evidence of

recirculation considering the elevated concentrations associated with the winds coming from the north-east and east (see Figure 8. 6). The monitoring unit is located under the Woolwich flyover in Greenwich London and is bounded by trees to the north and the bridge underpass to the south. It is not clear how this complex air flow is formed, and it may require further investigation to ascertain how the bridge and the trees affect the dispersion of the pollutants at this site. However, the flow pattern was only captured by the ANN and BRT models and slightly underestimated by the SVM while ADMS-Roads accounted for only the concentrations along the road axes.

BT4 is located to the northern side of the London North Circular Road (north-east, south-west directions) adjacent to the Brent Park Ikea store. The Bivariate polar plot of the observations at this site did not show any evidence of recirculation as the monitoring site is located in a relatively open area. The higher PM_{10} concentrations at this site are associated with the winds coming from the south-west, signifying the effect of the road, but it is not clear whether the parking area located to the south of the monitoring unit might have affected the concentrations as there was no evidence of higher concentrations associated with the winds coming from that direction. The parking space, though elevated, has no solid walls that can cause recirculation. However, an in-depth analysis needs to be carried out to determine the effect of the parking area on the level of concentrations at the site.

The $PM_{2.5}$ concentrations (see Appendix F - Figure F.14) at this site show only an inclination of the winds coming from the east, showing that most of the $PM_{2.5}$ concentrations recorded at this site came from the road. The machine learning models estimated the flow pattern better than ADMS-Roads at this site because it shows an association between the concentrations and north-west winds which were not shown in the observations.

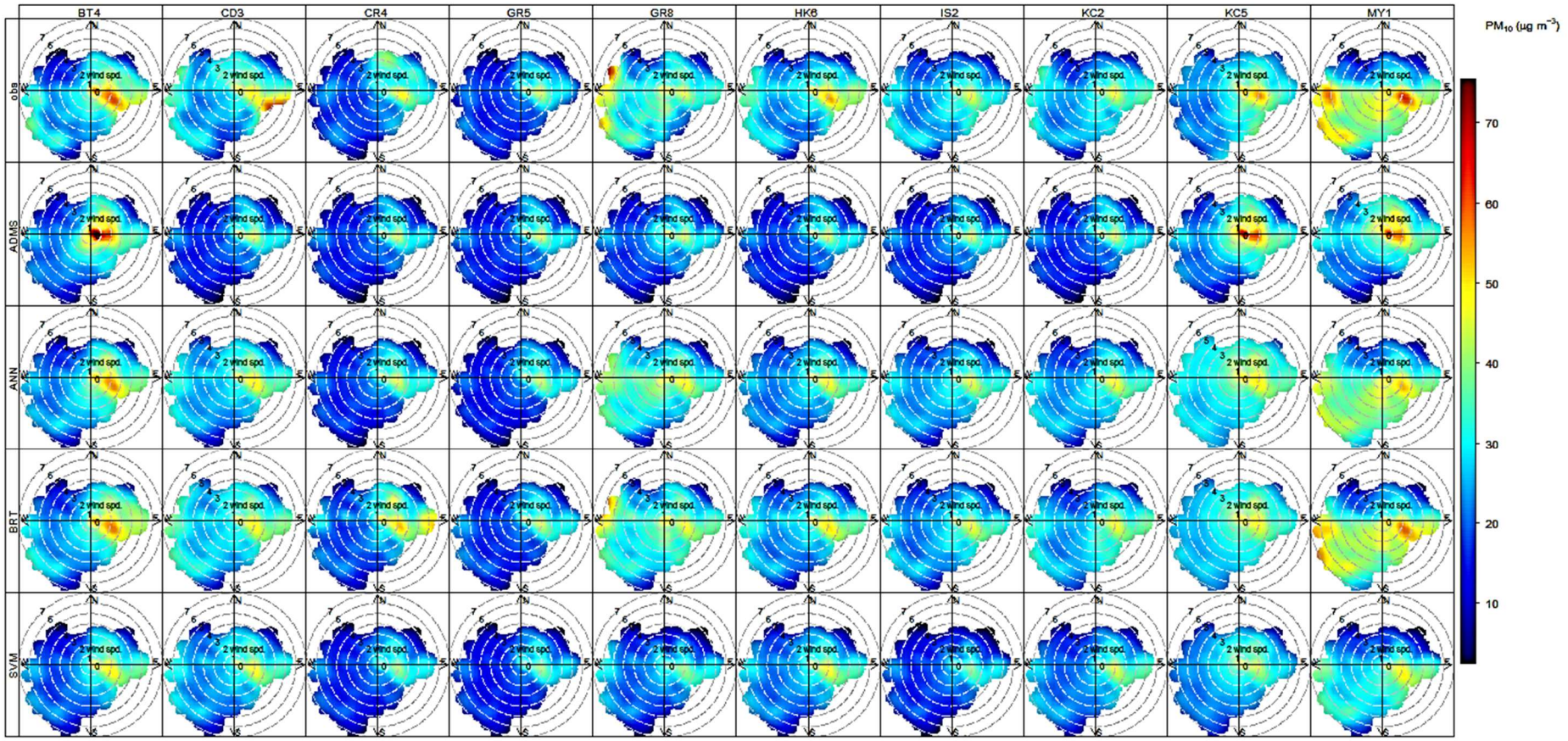


Figure 8.6 Bivariate Polar plots showing the variation of the PM_{10} concentrations with wind speeds and wind directions in the model predictions and the observations.

The IS2 and HK6 sites are located in street canyons and at both sites, the higher concentrations are associated with the winds coming from the direction where the monitoring units are located. This indicates the effects of complex canyon flows and the predictions of the machine learning models have shown good agreement with the observations. GR5 and KC2 are located at intersections formed by crossing street canyons, and the observations have demonstrated that there was an effect of Canyon recirculation at GR5, and a much more complex pattern at KC2 and the predictions of the machine learning models adequately captured these patterns. Whereas ADMS-Roads only predicted the pattern at GR5 and failed at KC2. The CD3 site is located at the heart of intersecting street canyons and the higher concentrations at this site, are associated with the winds coming from the south-west which is perpendicular to the road running along a north-east/south-west axis.

The machine learning models have captured the same pattern at this site but underestimated the concentrations associated with higher winds from the south-west. TH4 monitoring site is located to at an intersection formed by the Blackwall Tunnel northern approach road (North –South direction) and Abbott Road (East-west direction) and towards the east of Blackwall Tunnel northern approach road. The polar plot shows that the higher $PM_{2.5}$ concentrations at this site were associated with the winds coming from the east showing that much of the pollution was coming from Abbott Road. The monitoring unit at this site is not in a street canyon. Therefore, recirculation is not expected. However, from the plot, there was no evidence pointing to how the Blackwall Tunnel northern approach road contributed to the concentrations at this site. All the models did capture the flow pattern at the site but with a slight exaggeration of the dispersion of medium concentrations by the ADMS-Roads.

At the Instrumented Junction in Leeds, the air flow is also complex as it is being influenced by the adjacent buildings enclosing the roads and the monitoring units. The machine learning

models performed very well in capturing these complex air flows at all the three sites, but the ANN model underestimated the elevated concentrations associated with the winds coming from the north-west at ENV1.

ENV1 is located on the eastern side of Otley Road (north-west/south-east axis) and approximately 50m northwards from the junction. Therefore, the elevated concentrations associated with the north-west winds represent the concentrations channelled through the street canyons and those associated with the north-east winds resulted in the canyon recirculation. ENV2 is located at the heart of the T-junction formed by the Otley Road and North Lane on the western side of Otley Road.

The higher PNC concentrations at this site are much more concentrated along the axes of the roads. Therefore, it could be said that the channelling flow is dominant in this case. However, despite the complex nature of the flow at this site, the prediction of the models captures the pattern of the observed concentrations adequately as shown in (see Appendix F – Figure F15).

ENV3 is located at the southern side of North Lane approximately 25m from the junction housed in a street canyon. The higher PNC concentrations at this site were more associated with the south-east winds which are perpendicular to North Lane and with the direction where the monitoring unit is located. This shows evidence of Canyon recirculation and the models performed well in capturing this pattern. At MY1 also the models captured the flow pattern of the PNC observations. The BRT models slightly overestimated the higher PNC concentrations at the Instrumented Junction sites than the ANN and SVM models.

8.4.1.4 Performance of The Models in Predicting Air Quality Statistics

This section presents the evaluation of the performance of the models on prediction of some annual air quality statistics. The statistics predicted were annual mean concentrations and the number of days where PM_{10} is greater than $50\mu g/m^3$.

1. Annual Mean Concentrations

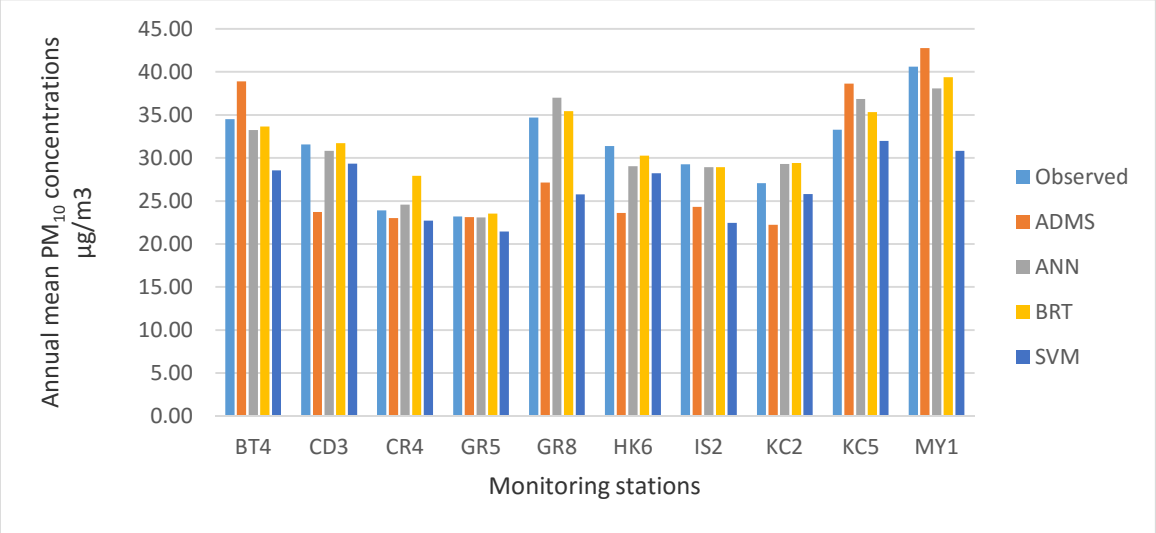


Figure 8.7 Predicted and observed annual mean PM_{10} ($\mu g/m^3$) concentrations

The models performed well in predicting the annual mean concentrations with slight overestimations and underestimations at some sites as shown in Figure 8.7. Here also the prediction of the machine learning models is much closer to the observed annual mean concentrations than those of the ADMS-Roads predictions. The SVM models performed poorly in predicting PM_{10} but show similar performance with the ANN and BRT in predicting $PM_{2.5}$ concentrations and was better in predicting PNC (see Appendix G Figures G.1 and G.2).

2. Number of days where PM_{10} is > 50 µg/m^3

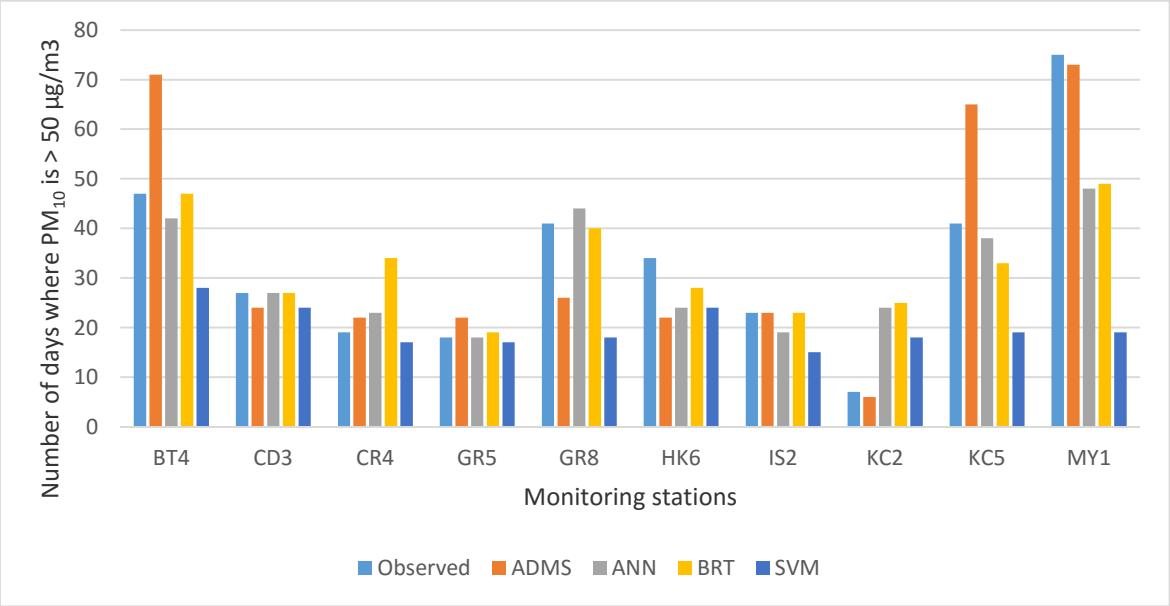


Figure 8.8. Predicted and observed number of days where PM_{10} is > 50 µg/m^3

The number of days where PM_{10} is > 50 µg/m^3 is of particular importance as it relates to health concerns and is very difficult to predict by most models as it can be affected by small uncertainties in the predictions (Carslaw D et al., 2013). However, the ADMS-Roads, ANN and BRT models performed well in predicting this statistic with various degrees of accuracy (see Figure 8.8). The predictions did not compromise the actual observations on the sites except at GR8 where ADMS-Roads predicted that the threshold provided in the EU directives of less than 35 days exceedance a year had been underestimated by 9 days which shows that the site is within the target, but the observation indicates that it has actually exceeded the target by 6 days. The SVM have severely underestimated the statistic in all the sites but most importantly at BT4, GR8, KC5 and MY1 where the threshold has been exceeded, but the model predicted no exceedance. The ADMS-Roads also overestimated the statistic at BT4 and KC5.

8.5 Performance of The Models in Predicting Daily Air Quality Index (without scenario)

Air quality index is a means of communicating information about real time and short term forecast of outdoor air pollution levels. This forecast serves as an advanced warning of the likely occurrence of an air pollution event that can be injurious to health (COMEAP, 2011). The pollutant (i.e. PM_{10}, $PM_{2.5}$, SO_2, O_3 and NO_2) indexes were developed by the Committee on the Medical Effects of Air Pollutants (COMEAP) in the UK, based on health evidence (COMEAP, 2011). The bands for the indexes include low, moderate, high and very high with each band having three different levels (e.g. low1, low2, and low 3) as shown in Table 8.3.

In its report, COMEAP (2011) recommended that the daily Air Quality Index (AQI) to be communicated to the public should be taken as the highest pollutant index among the five pollutant indexes since there is no sufficient understanding of the effects of mixtures of air pollutants. Therefore, COMEAP (2011) discourages the use of composite air quality index in the UK. The machine learning models (ANN, BRT, RF and SVM) were tested for the prediction of the pollutant indexes and the daily AQI. There have been studies on the use of the machine learning methods in the prediction of AQI and Composite Air Quality Index (CAQI) estimated based on EPA (1999) methods and suggested that the machine learning methods are suitable tools for air quality prediction and management (Singh et al., 2013, Kumar and Goyal, 2013). However, in this research, both the pollutant indexes and the daily AQI were estimated based on the methods recommended by the COMEAP. The pollutant indexes for the PM_{10} and $PM_{2.5}$ were estimated from the prediction of the machine learning models and the ADMS-Roads (see Section 8.4), and their performance was compared. Conversely, in the case of the prediction of the daily AQIs, the machine learning models were trained for multilevel classification of the AQIs. The AQIs were estimated from the

air pollution data covering a period between 2007 and 2012 using COMEAP (2011) method. The predictor variables considered were only meteorological variables and the temporal variables (i.e. day, month and year). The pollutant variables were removed from the predictor variables because they were used to estimate the indexes. The traffic variables were found to be less significant in the prediction of the AQIs and therefore discarded. The daily AQI was taken as the maximum pollutant index among the O_3, PM_{10}, $PM_{2.5}$, SO_2 and NO_2 indexes. The machine learning models including ANN, BRT, RF and SVM were then applied in training the models for the prediction of the AQIs.

Table 8.3. Recommended Index Pollutants and their Breakpoints for each Band (COMEAP, 2011).

Band	Index	NO_2 1-hour mean ($\mu g\ m^{-3}$)	O_3 Running 8-hour mean ($\mu g\ m^{-3}$)	SO_2 15-minute mean ($\mu g\ m^{-3}$)	PM_{10} 24-hour mean ($\mu g\ m^{-3}$)	$PM_{2.5}$ 24-hour mean ($\mu g\ m^{-3}$)
1	Low	0–66	0–33	0–88	0–16	0–11
2	Low	67–133	34–65	89–176	17–33	12–23
3	Low	134–199	66–99	177–265	34–49	24–34
4	Moderate	200–267	100–120	266–354	50–58	35–41
5	Moderate	268–334	121–140	355–442	59–66	42–46
6	Moderate	335–399	141–159	443–531	67–74	47–52
7	High	400–467	160–187	532–708	75–83	53–58
8	High	468–534	188–213	709–886	84–91	59–64
9	High	535–599	214–239	887–1063	92–99	65–69
10	Very High	600 or more	240 or more	1064 or more	100 or more	70 or more

8.5.1 Comparison of The Performance of the Machine Learning Models in Predicting Pollutant Index

The performance of the models in predicting the pollutant index is crucial especially if the models are to be used for forecasting. These indexes are used by the regulatory agencies to estimate the daily AQIs to issue warnings to the people against possible health implications of their outdoor activities especially during air quality episodes. The indexes for PM_{10} and $PM_{2.5}$ concentrations are calculated daily using the 24-hour mean (see Table 8.3). This section presents the performance of the models in predicting the pollutant indexes in terms of normalised mean bias and RMSE values.

Figures 8.9 show the graphical comparison of the normalised mean biases of the machine learning and ADMS-Roads models in predicting the daily air quality indexes. For the prediction of PM_{10} indexes, the machine learning models have lower biases than the ADMS-Roads model. The ADMS-Roads model tends to have negative biases for lower indexes and positive biases for the higher indexes except at BT4, KC5 and MY1 sites where both the lower and higher indexes have positive biases. There is little difference between the predictions of the different machine learning models and they show negative bias for the higher indexes at CR4, GR8 and KC5 sites. For $PM_{2.5}$ concentrations (Figure G.8 in Appendix G), the ADMS-Roads model has shown larger positive bias at BT4, HK6 and GR8 while the machine learning models show much less bias in predicting the indexes at all the sites. The ADMS-Roads model showed less bias at GR9, TH4 and MY1, but it has many problems in predicting the lower indexes where it over-predicted at GR9 and TH4. However, the machine learning models have most of their bias close to zero signifying much more accurate prediction. At MY1, all the models under-predicted the lower index.

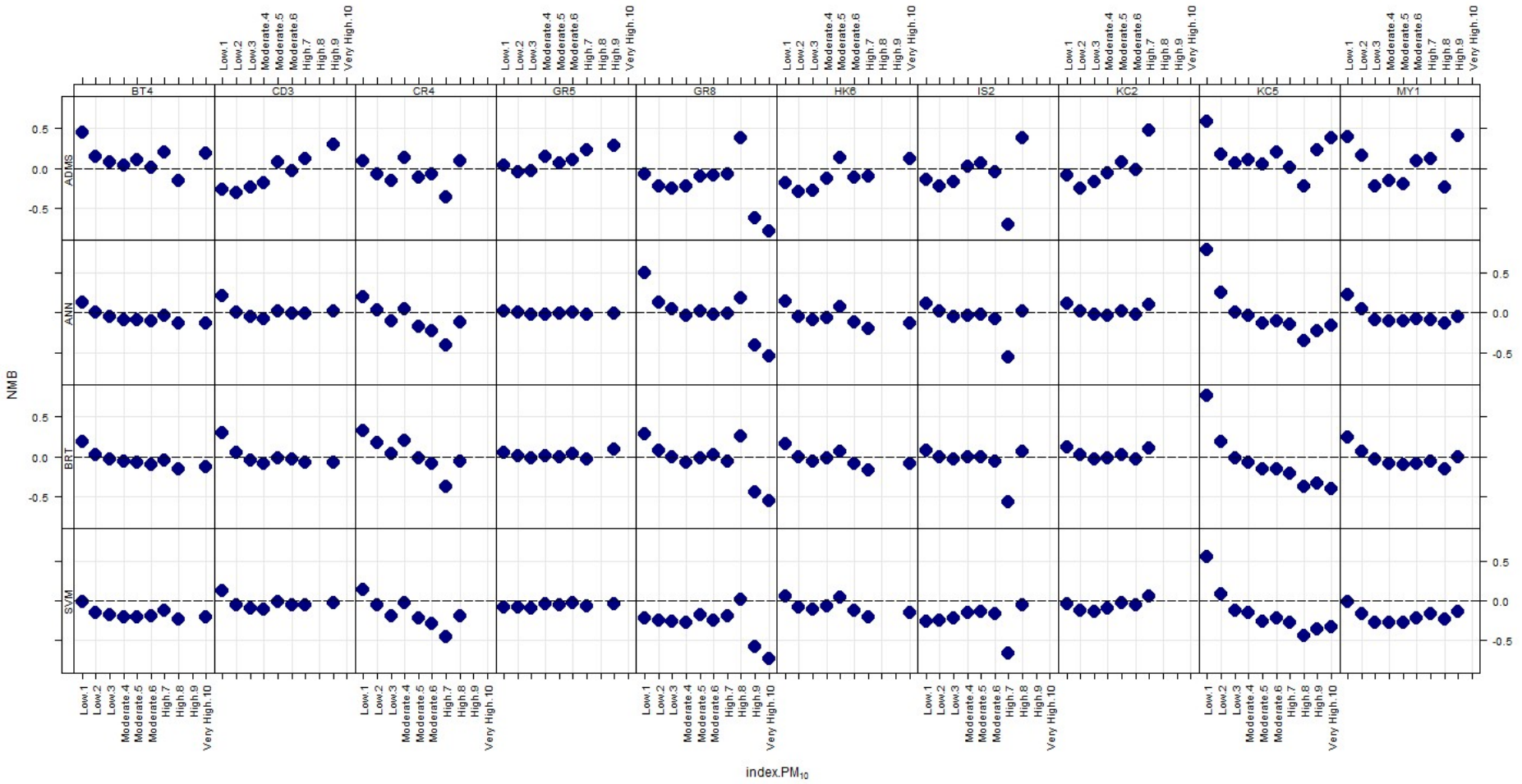


Figure 8.9 Graphical comparison of model performance (normalised mean bias) against daily air quality index for PM_{10}

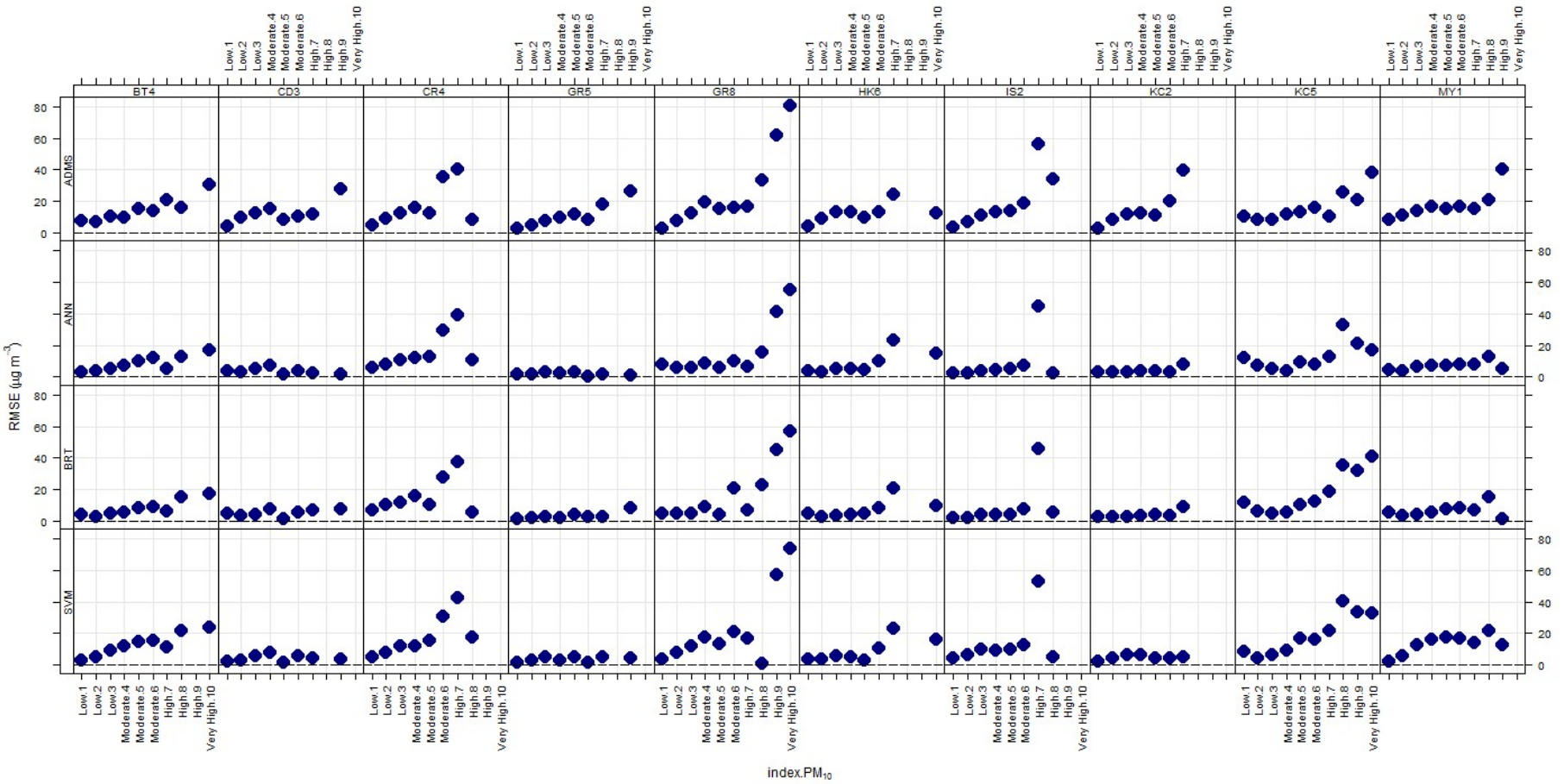


Figure 8.10 Graphical comparison of model performance (RMSE) against daily air quality index for PM_{10}

Figures 8.10 and G.9 show that the machine learning models have a much lower root mean squared errors than the ADMS-Roads model, and they both have higher errors in predicting the higher indexes. All the models show identical performance at CR4 and much higher errors in estimating higher indexes at GR8. The poor performance might be connected with the complexity of air flows at these sites and the rare occurrence of the lower and higher indexes.

8.5.2 Comparison of The Performance of the Machine Learning Models in Predicting AQI

The training performance of the machine learning models in predicting the AQIs is shown in Table 8.4, in terms of R^2 and RMSE. The ANN models have shown higher training performance with R^2 and RMSE values of 0.99 and 0.09 respectively.

Table 8.4. Training performance results of the AQI prediction models

Models	**RMSE**	**R^2**
ANN	0.09	0.99
RF	0.63	0.69
BRT	0.67	0.65
SVM	0.90	0.45

The least performing model during the training was the SVM with R^2 and RMSE values 0.45 and 0.90 respectively. BRT and RF have shown nearly similar results with moderate performance as demonstrated by the R^2 and RMSE values of 0.63 and 0.69 for RF, and 0.67 and 0.65 for the BRT. The performance of the ANN is a source of concern since it has shown performance values close to the values for an ideal model. Therefore, the ANN model might have overfitted the data. The models were then tested using a test data set not used in the training. The test performance results are shown in Table 8.5

Table 8.5 Test performance results of the AQI prediction models

Models	Accuracy	Kappa	RMSE	R^2
ANN	0.53	0.32	0.19	0.77
RF	0.57	0.33	0.14	0.80
BRT	0.55	0.31	0.17	0.78
SVM	0.67	0.50	0.11	0.92

Accuracy is a statistical measure of how well a classification test correctly identifies or excludes a particular class or index level in this case. The accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined. The overall accuracy of the test of the ANN, BRT, RF and SVM models were 53%, 57%, 55%, and 67%. The Kappa statistics values of 0.32, 0.33 and 0.31 suggest a moderate agreement between the observation and the predicted AQIs for the ANN, RF and BRT models. The SVM showed the highest agreement with Kappa values of 0.5. The models have performed moderately well as indicated by a very low RMSE values between 0.11 and 0.19 and higher R^2 values between 0.77 and 0.92. The SVM show the higher performance considering the all the performance metrics used. ANN is the least performing model closely followed by the BRT. Kumar and Goyal (2013) reported similar R^2 values (0.63 – 0.88) for the seasonal performance of ANN models in predicting the AQIs in India. However, Singh et al. (2013) reported the model accuracy of more than 90% and the coefficient of correlation values of more than 0.9 for the ANN, BRT, RF and SVM models for the prediction of seasonal AQIs. The main difference with our study is that Singh et al. (2013) use SO_2 and NO_2 as part of the predictor variables, and the response variables were seasonal AQIs and composite AQIs while here daily AQIs were response variables. The performance of the models in the prediction of pollutant indexes is by far better than the prediction for the daily

AQIs. Therefore, if the AQIs are to be determined according to COMEAP (2011) method, it is recommended that the pollutants concentrations are predicted first and then estimate the daily indexes using the predicted concentrations.

8.6 Determining The Effects of Euro4/VI Scenario On the PM_{10} and $PM_{2.5}$ Air Quality Metrics Using Machine Learning Models and ADMS-Roads

This section presents an evaluation of the capabilities of the models in assessing the effectiveness of air quality control scenarios. Given this, the models were tested for Euro4/VI scenario. The Euro4/VI scenario assumed that in 2011, and in 2015 only petrol vehicles meeting EuroIV/4 design standards and diesel vehicles meeting EuroVI/6 design standards will be allowed to drive on the roads in the study areas. The scenario was estimated by dis-aggregating traffic volume into eight traffic categories (i.e. Petrol car, Diesel car, Taxi, LGV, Rigid, Artic, Bus and coach and Motorcycle) based on the UK traffic composition projections. Their corresponding hourly emission rates were estimated using LAQM emission factor toolkit version 6.0.1.

The emission rates were then used as part of the input data for the training of the three machine learning models. The use of the emission rate is expected to provide a channel through which the response of the models to the changes in the emissions can be investigated. If the response of the models is positive, then they could be used as management tools for measuring traffic-related air quality control scenarios. Otherwise, they could only be used for prediction of the actual concentrations.

To compare the predictions of the machine learning models with the prediction of the ADMS-Roads, the same scenario was implemented in the ADMS-Roads model covering the Westminster City Council area. The procedure for implementing the Euro4/VI scenario in ADMS-Roads involved developing a detailed emissions inventory within the study area, however since the focus of this study is on the traffic-related air pollution, only 109 major

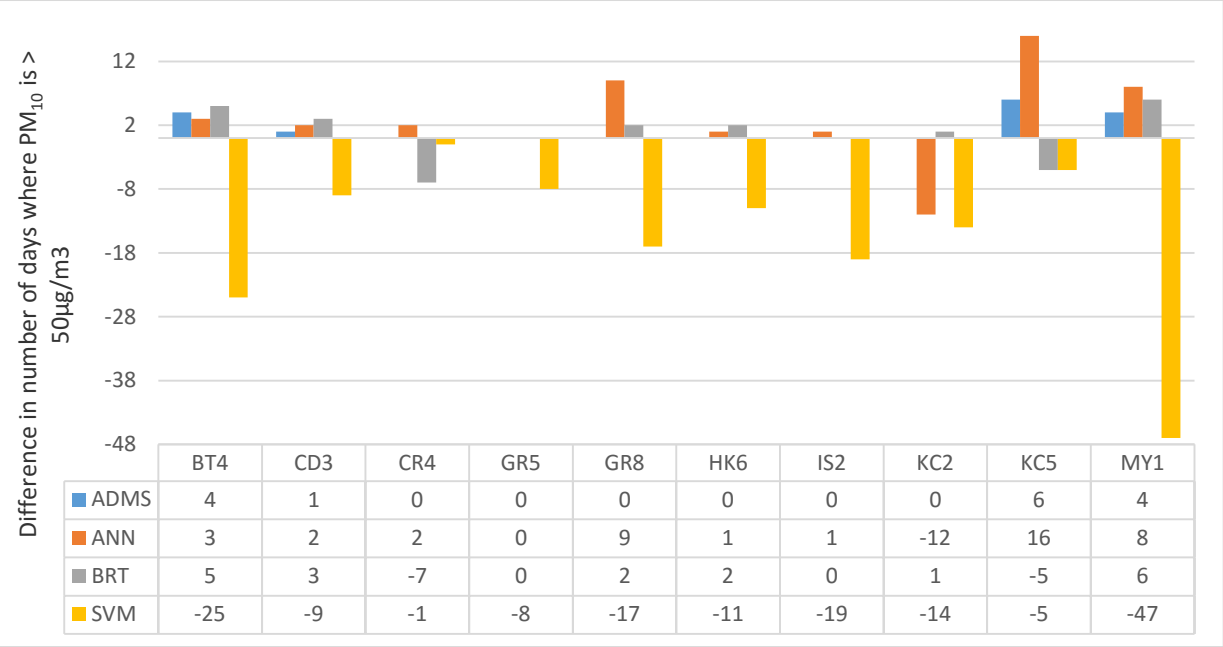
road links within the area were considered. The data collected on each link included traffic volume, average traffic speed, link length, width and elevation of the roads, height and width of the street canyons. The data were then used to prepare emission inventories with and without the scenario imposed using LAQM emission factor toolkit version 6.0.1 and the ADMS-Roads emission inventory utilities. The years 2011 and 2012 were chosen as the base years to match the test years of the machine learning models for PM_{10} and $PM_{2.5}$ predictions respectively.

The scenario was implemented for the 2011 and 2015 in the case of PM_{10}, while for $PM_{2.5}$, it was applied for the years 2012 and 2015. Two monitoring sites, KC2 and MY1, were the only sites within Westminster City Council with sufficient PM_{10} data in 2011, and only MY1 has sufficient $PM_{2.5}$ data in 2012, among the ten PM_{10} and six $PM_{2.5}$ sites used in the machine learning modelling. Hence, the detailed modelling in the ADMS-Roads model involved only these two sites while, for the remaining sites, only the roads adjacent to the monitoring stations were considered.

8.6.1 Comparison of the Estimated Effects of Euro4/VI Scenario on the PM_{10} Concentrations

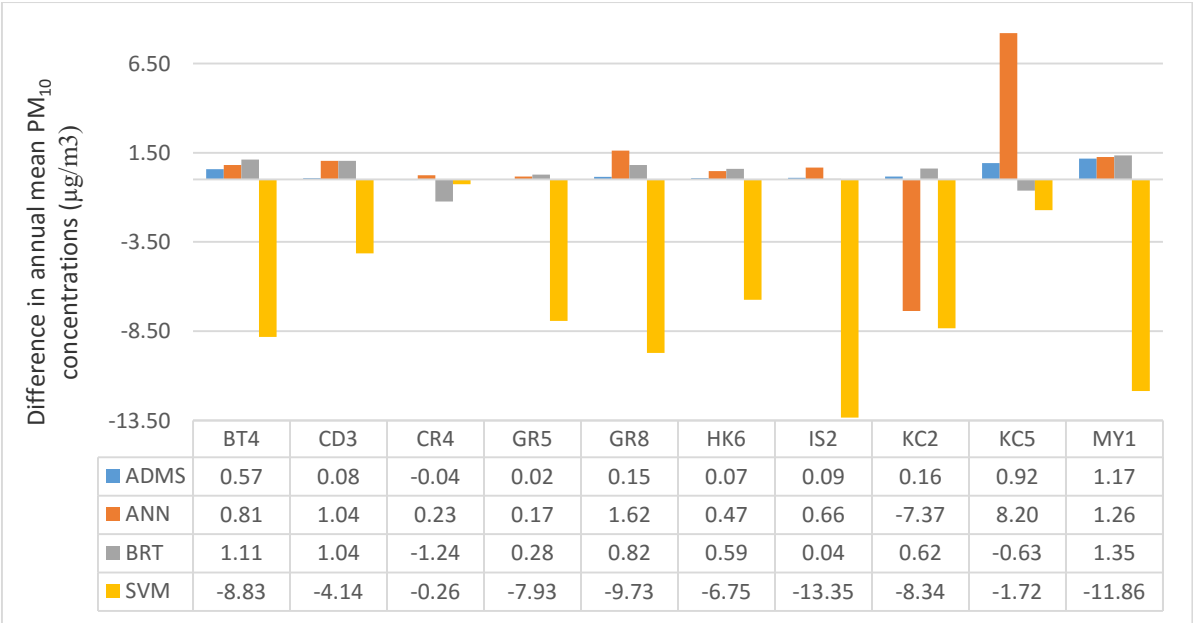
The Euro4/VI scenario implemented is hypothetical. Therefore, there is no actual data to compare the performance of the models. However, we compared the machine learning models against the ADMS-Roads since the latter has been used for the same purpose in many projects. Figures 8.12 and 8.13 show the predicted effects of Euro4/VI scenario on the number of days with $PM_{10} > 50\mu g/m^3$ and the predicted effects of Euro4/VI on the annual mean PM_{10} concentrations in 2011. From the figures, it could be observed that the base year predictions of the ADMS-Roads model compare well with the observed PM_{10} concentrations with slight under or overestimations depending on the site under consideration. When the scenario was applied to the data collected in 2011, the ADMS-Roads model predicted the

slight effect on the PM_{10} concentrations. It predicted that the annual mean PM_{10} concentrations would be reduced by 0.16µg/m^3 and 1.09 µg/m^3, and the number of days where PM_{10} is greater than 50 µg/m^3 was predicted to be reduced by only zero and four days at KC2 and MY1 respectively. The predictions of the ADMS-Roads model showed that the EUROVI scenario would have very little or no effect if it was to be implemented in 2011 at the CR4, GR5, GR8, HK6 and IS2 monitoring sites. However, it showed that the PM_{10} concentration would have been slightly reduced at BT4 and KC5 monitoring sites.



	BT4	CD3	CR4	GR5	GR8	HK6	IS2	KC2	KC5	MY1
ADMS	4	1	0	0	0	0	0	0	6	4
ANN	3	2	2	0	9	1	1	-12	16	8
BRT	5	3	-7	0	2	2	0	1	-5	6
SVM	-25	-9	-1	-8	-17	-11	-19	-14	-5	-47

Figure 8.11 Predicted effects of Euro4/VI scenario on the number of with PM_{10} > 50µg/m^3 in 2011



	BT4	CD3	CR4	GR5	GR8	HK6	IS2	KC2	KC5	MY1
ADMS	0.57	0.08	-0.04	0.02	0.15	0.07	0.09	0.16	0.92	1.17
ANN	0.81	1.04	0.23	0.17	1.62	0.47	0.66	-7.37	8.20	1.26
BRT	1.11	1.04	-1.24	0.28	0.82	0.59	0.04	0.62	-0.63	1.35
SVM	-8.83	-4.14	-0.26	-7.93	-9.73	-6.75	-13.35	-8.34	-1.72	-11.86

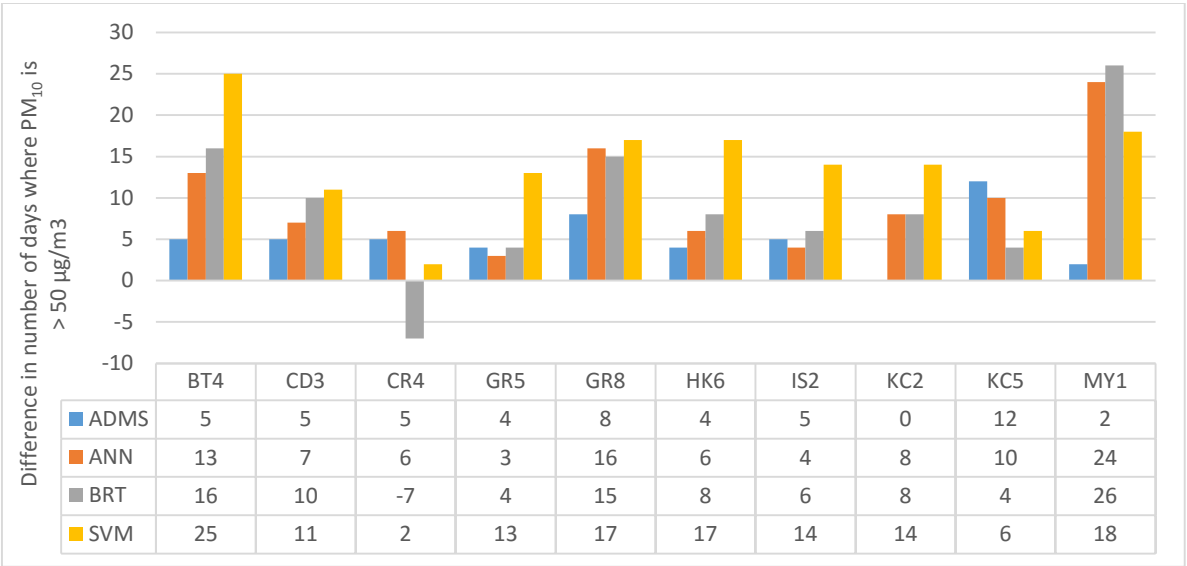
Figure 8.12 Predicted effects of Euro4/VI on the annual mean PM_{10} concentrations in 2011

On the other hand, the machine learning models predicted more reduction in the PM_{10} concentrations when the scenario was implemented for the year 2011. The ANN and BRT predicted that at MY1 monitoring site, the annual mean PM_{10} concentrations would be reduced by 1.26 µg/m^3 and 1.35 µg/m^3, respectively. Also, they predicted eight and six days reduction in the number of days with PM_{10} higher than 50µg/m^3 respectively. However, the SVM model has consistently shown that there will be an increase of annual mean PM_{10} concentrations by 11.86 µg/m^3 and 47 days with PM_{10} greater than 50µg/m^3 which is contrary to the expected results since the emission has been reduced.

At KC2, only the BRT model prediction agrees with the ADMS-Roads predictions where the PM_{10} concentrations will be slightly reduced with the implementation of the scenario. Whereas, the ANN and SVM predicted a large increase in the PM_{10} concentrations in 2011 with the implementation of the scenario. This failure might be attributed to the number of missing values imputed in the data used for the training of the models. The data captured at this site in 2011, was 73%, and after the imputation, there were 28 days when PM_{10} concentrations were higher than 50µg/m^3 as against the 7 days in the original data, and the

95 and 99 percentiles increased by 10µg/m^3 in the imputed data. Therefore, the performance of the models might be affected as the imputation heavily influenced the original data.

When the scenario was implemented for the year 2015, the ADMS-Roads model predicted that without the scenario the number of days where the PM_{10} greater than 50µg/m^3 at MY1 will be reduced by only 2 days, and with the scenario implemented, it will be further reduced by 21 days. However, the ANN and BRT predicted less reduction in the PM_{10} concentrations than what was predicted by the ADMS-Roads as shown in Appendix G - Figure G.3. They predicted the reduction of 2.18 µg/m3 and 1.53 µg/m^3 in annual mean PM_{10} concentrations as against a 3.86 µg /m^3 predicted by the ADMS-Roads as shown in Appendix G - Figure G.4. The disparity could be traced to the performance of the models where the ADMS-Roads overestimated the observations while the ANN and BRT models underestimated the observations. At the KC2 monitoring site, all the models have shown approximately the same performance where they predicted that the scenario would have much less effect and mostly remained the same considering most of the statistics.



	BT4	CD3	CR4	GR5	GR8	HK6	IS2	KC2	KC5	MY1
ADMS	5	5	5	4	8	4	5	0	12	2
ANN	13	7	6	3	16	6	4	8	10	24
BRT	16	10	-7	4	15	8	6	8	4	26
SVM	25	11	2	13	17	17	14	14	6	18

Figure 8.13 Predicted Change in days with PM_{10} > 50 µg/m^3 from 2011 to 2015 at monitoring stations

	BT4	CD3	CR4	GR5	GR8	HK6	IS2	KC2	KC5	MY1
ADMS	1.97	1.60	1.50	1.62	2.21	1.59	1.61	0.16	2.08	1.09
ANN	3.08	3.16	1.83	1.79	4.29	2.34	2.63	2.67	1.49	3.87
BRT	3.52	2.74	-1.48	1.98	4.40	2.79	3.80	2.41	0.86	5.35
SVM	19.47	8.21	0.40	12.13	19.28	13.69	23.77	14.74	1.81	24.40

Figure 8.14 Predicted change in the annual mean PM_{10} concentrations from 2011 to 2015 at the monitoring stations.

The models were also evaluated for their performance in predicting the change in the PM_{10} concentrations from 2011 to 2015 without the Euro4/VI scenario as shown in Figures 8.14 and 8.15. It could be seen that they predicted various degrees of reduction in annual mean PM_{10} concentrations and the number of days where PM_{10} is greater than 50µg/m^3. The ANN and BRT models show higher degrees of reduction than the ADMS-Roads model and the SVM models exaggerated the reductions in most of the sites.

8.6.2 Comparison of the Estimated Effects of Euro4/VI Scenario on the $PM_{2.5}$ Concentrations

The machine learning models have performed well when compared with the $PM_{2.5}$ observations which are a good sign that the models can be applied to real-life problems. The Euro4/VI scenario explained above was also implemented in predicting the 2012 and 2015 $PM_{2.5}$ concentrations. Figures G.5 – G.6 in Appendix G show the predicted effect of the scenario on the annual mean $PM_{2.5}$ concentrations when it was implemented for the years

2012 and 2015. In all the six sites considered, the ADMS-Roads models predicted the reduction of less than 0.05µg/m^3 in the annual mean $PM_{2.5}$ concentrations in 2012 at 4 sites while it predicted no change at GR8 and a slight increase of 0.01 µg/m^3 at GR9 (see Appendix G – Figure G.5).

However, the machine learning models predicted more reduction of the annual mean $PM_{2.5}$ concentrations ranging between 0.3 µg/m^3 at GR9 to 2µg/m^3 at MY1. In most cases, the predictions of the ANN and SVM were similar while BRT showed a slightly different result, but all of them predicted a much higher reduction than the ADMS-Roads. The same trend was also observed when the scenario was implemented in 2015 where the machine learning models predicted more reduction than the ADMS-Roads model and much higher reduction at MY1 and GR8 as shown in Appendix G – Figure G.6. The change in the levels of the concentrations predicted by the models without the Euro4/VI scenario is shown in Appendix G – Figure G.7. It could be seen that all the models predicted a reduction in the concentrations from 2012 to 2015, even though there was an increase in the traffic volume. The reduction might be attributed to improvements in vehicle technology and other air quality control measures being implemented presently in London.

The machine learning models developed for the prediction of PNC at the Instrumented Junction in Leeds were also tested for use in managing air quality. The Euro4/VI scenario could not be estimated for these models because in the LAQM emission factor toolkit version 6.0.1, there was no provision for PNC emission factors. Therefore, it was assumed that a certain scenario was implemented in and around the monitoring sites in Leeds resulting in a 20% emission reduction at the sites and the reduced emission rates were used as part of the test input data. The results show that all the models predicted reductions in the annual mean PNC concentrations at the sites (see Figure 8.16). However, ANN and BRT have

shown a much higher reduction at ENV1and ENV3 respectively while the BRT predicted a slight increment even with the emission reduction at ENV2.

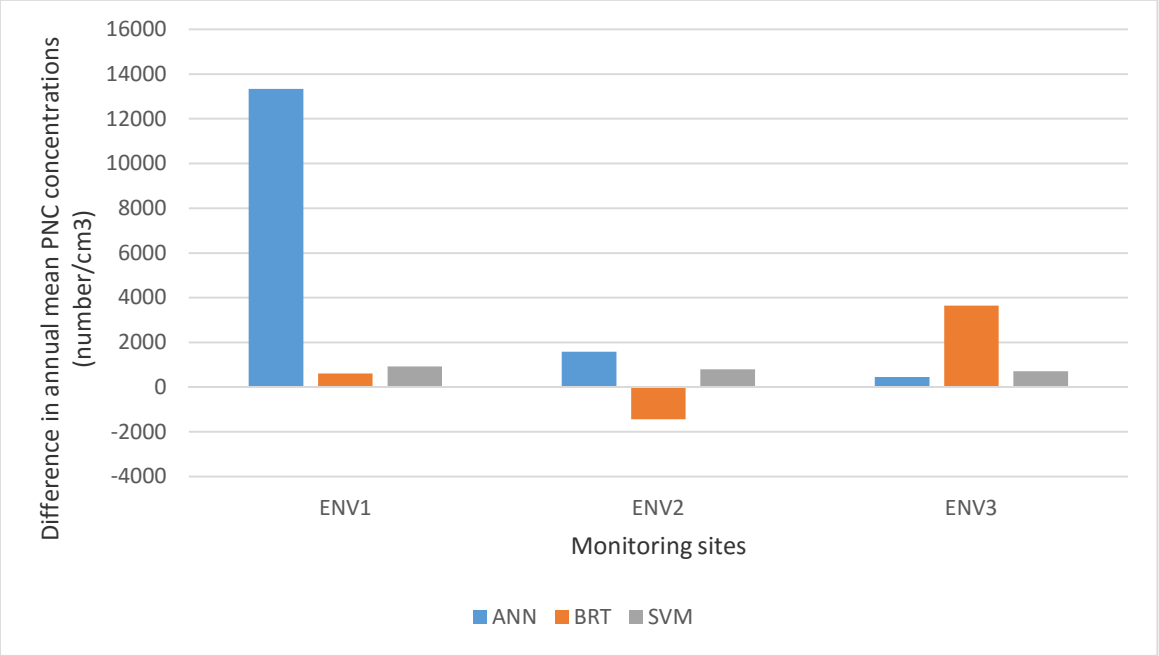


Figure 8.15 Predicted the effect of 20% emission reduction in the annual mean PNC concentrations.

8.7 Summary

This chapter has evaluated the application of the machine learning based air quality models discussed in Chapter 7 in both spatial and temporal predictions and their application to an air quality management scenario involving PM_{10} and $PM_{2.5}$ concentrations.

The performance of the machine learning models is better than the ADMS-Roads model. Moreover, they can perform much better when the geometry of the street canyon and the complex air flow within it affect the PM concentrations. The ANN and BRT performed better than SVM in predicting PM_{10} while there was little difference in their performance when predicting $PM_{2.5}$ and PNC concentrations. However, SVM performed slightly better

in predicting PNC at the Instrumented Junction in Leeds. The models also did not show the difference in performance when predicting PM in mass or number concentrations.

When evaluating the effectiveness of the Euro4/V,I scenario, the ANN and BRT models predicted much higher reductions in the PM_{10} and $PM_{2.5}$ concentrations than ADMS-Roads. While in a few cases, they predicted that the concentrations would remain unchanged. The SVM model consistently predicted higher PM_{10} concentrations when tested with the scenario while predicting much smaller reductions in $PM_{2.5}$ concentrations. However, the SVM model predicted a much higher reduction in the PM_{10} concentrations in 2015 without the scenario. According to all the performance metrics and charts shown above the SVM model was the poorest of the machine learning models in predicting PM_{10} whereas it showed similar performance with ANN and BRT in predicting $PM_{2.5}$ and PNC. The reason for this behaviour needs to be investigated further. Also, the ADMS-Roads model estimates the monthly and annual concentrations better than hourly and daily concentrations while the performance of machine learning models did not show much difference.

Chapter 9

Conclusions and Recommendations

9.1 Introduction

This chapter presents the summary of the seven main chapters of this study and how they have fulfilled the study aim and objectives. Implications of the research findings, recommendations for air quality modelling community and recommendations for further studies are also presented.

9.2 Conclusions

This research developed models for point and spatiotemporal prediction of the traffic-related particulate matter using statistical and machine learning methods. The performance of the models was evaluated and compared with the performance of an operational model (ADMS-Roads) both in the predictions and for assessing the effectiveness of a hypothetical air quality management scenario. Particulate matter including PM_{10}, $PM_{2.5}$ and PNC were considered as the pollutants of choice for the study because of the lack of proper treatment of particle chemistry in the existing models and also most of the models did not include PNC as part of the particle pollutant metric to be estimated and the need for their accurate estimation due to their health implications.

- It was estimated that the PM_{10} observations related to the road constitute about 22% - 44%. Moreover, the percentage mean PM_{10} increment contributed by the roads was between 24% and 62% of the roadside mean PM_{10} concentrations. It was also found that the percentage frequency of the contribution of other sources was only between 2% and 4%. However, the average PM_{10} increments due to the other sources were also found to be nearly the same as the roadside PM_{10} increments.

- The use of Genetic Algorithms (GA) and Simulated Annealing (SA) combined with Random Forests (RF) as feature selection methods for the machine learning and statistical models resulted in a reduction of 18, 11 and 12 out 28, 27 and 25 predictor variables for the prediction of PM_{10}, $PM_{2.5}$ and PNC respectively. Training the models with few predictors consequently, reduces the operational and computational costs of the models. It was also discovered that there was a slight increase in performance when using feature selection before the statistical modelling.

- The most important predictor variables identified by all the models are the roadside oxides of nitrogen and background particle concentrations. The meteorological variables, background pollutants and traffic variables are the second most important variables and have similar contributions in all the models except in MLR. The temporal variables are more important in predicting PNC than $PM_{2.5}$ and PM_{10} concentrations.

- The models were found to perform better in predicting roadside concentrations than roadside increments.

- The BRT, RF, ELM and deep learning algorithms were considered to be most suitable for the prediction of the PM_{10}, $PM_{2.5}$ and PNC concentrations due to their predictive accuracy and faster training speed.

- The feature selection is most appropriate for the traditional MLP neural networks because they show some improvement in performance when trained with the selected variables. The machine learning models performed slightly better in predicting $PM_{2.5}$ and PNC than PM_{10} concentrations.

- Among the machine learning methods considered BRT shows consistent and better performance as shown by the higher COE, IOA, R values for all the particle metrics used

and has the additional capability of producing partial dependence plots which provide more information about the interactions between the predictor variables and response variables during the modelling process.

- Deep learning and the ELM algorithms, though under active development, are purported to be more sophisticated algorithms that can handle a variety of cases including image processing, character and speech recognition and even language translation. However, in this study they did not show much difference in prediction performance than the traditional neural network and the tree-based models. The advantages of using the deep learning and ELM algorithms over traditional ANN observed in this study are the faster training speed and scalability especially when using high-performance computing. The BRT and RF algorithms are also scalable and speedy like Deep learning and ELM, and they produced similar predictions.

- Deep learning, ELM, BRT, and RF are recommended to be the best machine learning algorithms because of their outstanding performance, scalability, and consistency. BRNN and PCA-MLP could equally produce good results especially when computing capacity and training speed is not priorities.

- The machine learning models can perform much better in the spatiotemporal prediction of the particles concentrations than the ADMS-road model when the geometry of a street canyon and the complex air flow within it affect the particle concentrations. The ANN and BRT performed better than SVM in predicting PM_{10} while there was little difference in their performance when predicting $PM_{2.5}$ and PNC concentrations.

- When evaluating the effectiveness of the Euro4/VI scenario, the ANN and BRT models predicted higher reductions in the PM_{10} and $PM_{2.5}$ concentrations than did ADMS-Roads. While in some few cases, the ANN and BRT models predicted that the concentrations

will remain unchanged. The SVM model consistently predicted higher PM_{10} concentrations when tested with the scenario while predicting a much lower reduction in $PM_{2.5}$ concentrations. However, the SVM model predicted a much larger decrease in the PM_{10} concentrations in 2015 without the scenario. According to all the performance metrics, the SVM model was the poorest of the machine learning models in predicting PM_{10} whereas it shows similar performance with ANN and BRT in predicting $PM_{2.5}$ and PNC.

9.3 Fulfilment of the Research Aim and Objectives

The main goal of this study is to examine the application of Machine Learning and Statistical Methods for developing roadside particle (number/mass concentrations) prediction models that can be used for air quality management. Given this aim, the main conclusion drawn from the study is that both machine learning and statistical methods can be used for developing models for the prediction of roadside particle concentrations (PM_{10}, $PM_{2.5,}$ and PNC). However, machine learning methods are more suitable and can be applied in the evaluation of traffic-related air quality control scenarios. Machine learning models can produce more accurate Spatio-temporal predictions of roadside particles more accurately than the ADMS-Roads.

The following questions have been answered in the course of this study

1. What are the most relevant predictor variables to be used for modelling roadside particle concentrations using machine learning and statistical methods? Moreover, how can we select them from the available predictor variables?

In this study, it was observed that using an evolutionary search algorithm combined with Random Forest (GA-RF) can be used to choose the most relevant predictor variables for both the machine learning and statistical models. The variables selected were found to have

produced cheaper, simpler and more interpretable models. The most relevant variables selected were background particle concentrations of particles, roadside NOx, Barometric pressure, relative humidity, Temperature, wind directions and speed, the day of the week and month of the year. Roadside NO was also found to be paramount for PNC.

2. Which among the roadside concentration and roadside increments is a better predictor or response variable as the case may be?

In this study, it was observed that there was no difference in using either roadside concentrations or roadside increments as predictors. However, the methods predict the roadside concentrations more accurately than the roadside increments. This failure might be attributed to the noise caused by negative values in the roadside increments. The predictions might improve if more appropriate background concentrations are determined.

3. Why use machine learning methods if simple statistical methods can be utilised?

It was established that evolutionary search algorithms combined with machine learning algorithms could select more relevant predictor variables for statistical models and make them less complex and more interpretable. Therefore, where the interpretation of the modelling results is more important than the prediction accuracy, the use of GA-RF with any one of the Statistical methods be recommended. However, when both interpretation and prediction accuracy are important, the use of BRT or RF is highly recommended. Conversely, when only prediction accuracy is important, the use of deep learning algorithms is recommended because they produce more stable prediction accuracy across the particle metrics.

4. ***Can the machine learning methods be applied for Spatio-temporal modelling of the particle concentrations?***

In this study, it was shown that using average background and roadside concentrations of NOx, meteorological and traffic variables; machine learning methods can be applied for spatiotemporal predictions of roadside particle concentrations. Moreover, they can produce more accurate predictions than the ADMS-Roads models especially in street canyons and at intersections.

5. ***Can the machine learning models be applied for evaluating the effectiveness of traffic-related air quality control scenarios?***

The machine learning models have been applied for evaluating hypothetical air quality control scenario, and the results obtained were compared with the results obtained from ADMS-Roads. Although there was a disparity in the results obtained, the pattern of the results is the same. The machine learning predicted more reduction in the concentrations as a consequence of the implementation of the scenario in most cases. The study recommends the use of ANN and BRT methods in this case but not SVM as it produced less accurate results.

9.4 Recommendations

9.4.1 Policy Implications of the Findings of this Study

Owing to their impact on human health and mortality rate, traffic-related particle concentrations should be extensively studied and their standards and limits revised and/or improved. Although the EU targets on particulate matter are being met recently in most UK cities, the effects of high particle concentrations are still important, which raises questions about the integrity of the targets in reducing the harmful consequences of the particles. This

paradox calls for using more robust methods of prediction and analysing the particle concentrations not only in mass metrics but also in particle number counts since the latter is more related to the studies on the health implications of the particle concentrations.

This study recommends the use of machine learning based air quality models to predict the effects of existing limits and the various air quality control strategies put in place in a given area. The models are expected to provide more accurate predictions considering the results obtained in this study. Also, new regulations and control strategies related to the PNC metric are highly recommended, and the machine learning methods can provide an effective way of evaluating their effectiveness with greater accuracy as demonstrated in this study. This feature is missing in most of the operational models.

The machine learning models would also provide better estimates of particulate matter concentrations when dealing with street canyons and intersections. Therefore, this study highly recommends the use of machine learning methods in this context.

Due to the strength of the associations between the roadside particles and the NOx concentrations, a more robust monitoring and control of NOx concentrations are recommended as this will help to develop more accurate machine learning models for the predictions of roadside particles. Also, any further reduction in NOx concentrations might likely result in a reduction of roadside particle concentrations.

Separate measurements or estimates of non-exhaust particles could enhance the quality of the model predictions as they might come from sources other than traffic. Therefore, if such variables are considered as part of predictor variables they might add to the accuracy of the predictions.

9.4.2 Transferability of the Machine Learning Models

The machine learning and statistical methods used in this research can be applied to any pollutant metric of choice and at any place where the meteorological, traffic and pollutants variables are available for training the models. The methods are also flexible in terms of the type and the number of variables to be used. The user is free to choose any relevant predictor variable that has some predictive power and later use RF-GA to select the most important predictor variables from the pool of predictor variables initially identified.

9.4.3 Recommendations for Local Authorities and Environmental Agencies

This study shows that both the machine learning and statistical methods are capable of producing models that can be used for point and spatiotemporal predictions of roadside particle concentrations with higher accuracy than the ADMS-Roads model. Also, the machine learning models can be applied in evaluating the effectiveness of traffic-related roadside particle reduction scenarios. The result of the evaluation of a hypothetical roadside particle reduction scenario shows that they predicted higher reduction than the ADMS-Roads slightly.

It is recommended that the Local Authorities and environmental agencies take the advantages offered by these methods by incorporating them into their modelling tools. The methods are freely available on many software platforms and do not require any particular computational training.

Using these methods will allow the environmental agencies a comprehensive utilisation of the air quality data being taken over the years. Also using machine learning methods provides flexibility in the modelling process as the modeller controls which variable to be involved in the modelling and the prediction time unit. The trained models can always be updated as the new data become available.

The training of machine learning methods involving large datasets might require high computing capacity. However, using the trained models require an only personal computer.

From the results obtained in this study, the machine learning methods will provide higher prediction accuracy than the ADMS-Roads at a relatively lower operational and computational costs. If these models are adopted in large scale projects by the Local Authorities and environmental agencies, they will provide savings and hence manage the air quality at a reduced cost.

9.4.4 Recommendations on the Modelling Procedure using Machine Learning and Statistical Methods

This study recommends the use of the hybrid feature selection (GA-RF) method before training either statistical or machine learning methods (ANN and SVM) for the prediction of roadside particle concentrations. In this study, it was shown that the use of GA-RF before using either statistical or machine learning methods for the modelling reduces the complexity of the models, the cost of operations and improves prediction accuracy.

Also, the strong relationships found between the NOx and all the particle metrics suggests that the concentrations of NOx can be used as surrogates of traffic-related particle concentrations where the particle concentration data is not available. Furthermore, when using machine learning and statistical methods to model traffic-related particle concentrations, background particle concentrations, roadside NOx, barometric pressure, relative humidity, temperature, wind direction and speed, the day of the week and month of the year are the most relevant predictor variables. Roadside NO was also found to be vital for PNC modelling.

9.5 Recommendations for Further Research

This study recommends the use of air quality image data as part of the predictors to harness the full potentials of newer machine learning algorithms such as Deep learning and Extreme learning machines; These methods have been proven to perform very well in image processing. Their inclusion in the modelling might allow a more robust spatiotemporal analysis of air quality of a very large area.

Also, we recommend a study on the use of machine learning methods in providing real-time predictions of air quality implications of various traffic control strategies and measures applied in densely congested areas.

Government agencies such as DEFRA should explore the application of the machine learning methods on a very large scale using more sophisticated computing systems and larger datasets. The results from such implementations should be compared with similar results from operational models such OSPM and CMAQ. There is a possibility of new discoveries about their application over such large scales which could enhance the confidence of air quality researchers to accept them as better and more modern modelling tools.

We also recommend the implementation of the machine learning methods in modelling the particle components as they are vital to understanding the health impact of particles.

9.6 Limitation of the Study

The study used meteorological data collected at Heathrow meteorological station and manual traffic counts in some cases instead of data collected at the individual sites which will provide microscopic details about the characteristics of the sites. This might affect the

performance of the models in predicting extreme events since they are infrequent and are influenced by local factors.

Using evolutionary algorithms as feature selection methods to both the machine learning and statistical methods adds to the computational requirements for training the models, although it has resulted in the reduction of a number of predictor variables required (Carslaw et al., 2014).

The roadside particle reduction scenario (Euro4/VI) used in this study is hypothetical. Therefore, there was no observed data with which to compare the performance of the models.

References

ABDEL-ATY, M., EKRAM, A.-A., HUANG, H. & CHOI, K. 2011. A study on crashes related to visibility obstruction due to fog and smoke. *Accident Analysis & Prevention,* 43**,** 1730-1737.

ABDERRAHIM, H., CHELLALI, M. R. & HAMOU, A. 2016. Forecasting PM10 in Algiers: efficacy of multilayer perceptron networks. *Environmental Science and Pollution Research,* 23**,** 1634-1641.

AGARWAL, A. K. 2007. Biofuels (alcohols and biodiesel) applications as fuels for internal combustion engines. *Progress in Energy and Combustion Science,* 33**,** 233-271.

AMIRSASHA BNANANKHAH, F. N. 2012. Artificial Neural Networks: A Non-Linear Tool for Air Quality Modeling and Monitoring. *International Conference on Applied Life Sciences*

ICALS2012.

ANDERSON, J., THUNDIYIL, J. & STOLBACH, A. 2012. Clearing the Air: A Review of the Effects of Particulate Matter Air Pollution on Human Health. *Journal of Medical Toxicology,* 8**,** 166-175.

AQEG 1999. Source Apportionment of Airborne Particulate Matter in the United Kingdom. *Department for Environment, Food & Rural Affairs*.

AQEG 2005. Particulate Matter in the United Kingdom . *Department for the Environment, Food and Rural Affairs*.

AQMRSG 2011. REVIEW OF AIR QUALITY MODELLING IN DEFRA. *Defra Publications*.

ASTM 2010. Standard guide for Statistical Evaluation of Atmospheric Dispersion Model Performance (D6589). *ASTM International.* West Conshohocken,PA: ASTM.

AYE, G. C. & GUPTA, R. 2013. Forecasting Real House Price of the US: An Analysis Covering 1890 to 2012.

BALAGUER, B. & CARPIN, S. 2012. Bimanual Regrasping from Unimanual Machine Learning. *2012 Ieee International Conference on Robotics and Automation (Icra)***,** 3264-3270.

BALSAMÀ, A. P., DE BIASE, L., JANSSENS-MAENHOUT, G. & PAGLIARI, V. 2014. Near-term projection of anthropogenic emission trends using neural networks. *Atmospheric Environment,* 89**,** 581-592.

BANERJEE, T., SINGH, S. B. & SRIVASTAVA, R. K. 2011. Development and performance evaluation of statistical models correlating air pollutants and meteorological variables at Pantnagar, India. *Atmospheric Research,* 99**,** 505-517.

BARMPADIMOS, I., HUEGLIN, C., KELLER, J., HENNE, S. & PRÉVÔT, A. 2011. Influence of meteorology on PM 10 trends and variability in Switzerland from 1991 to 2008. *Atmospheric Chemistry and Physics,* 11**,** 1813-1835.

BARRETT, S. H. L. Y. A. S. R. H. 2012. Public health impacts of combustion emissions in the United Kingdom. *Environ Sci Technol,* 46**,** 4291-6.

BEDDOWS, D. C. S. & HARRISON, R. M. 2008. Comparison of average particle number emission factors for heavy and light duty vehicles derived from rolling chassis dynamometer and field studies. *Atmospheric Environment,* 42**,** 7954-7966.

BENAS, N., BELOCONI, A. & CHRYSOULAKIS, N. 2013. Estimation of urban PM10 concentration, based on MODIS and MERIS/AATSR synergistic observations. *Atmospheric Environment,* 79**,** 448-454.

BENGIO, Y. Deep learning of representations: Looking forward. International Conference on Statistical Language and Speech Processing, 2013. Springer, 1-37.

BENGIO, Y., COURVILLE, A. & VINCENT, P. 2013. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* 35**,** 1798-1828.

BENNETT, N. D., CROKE, B. F. W., GUARISO, G., GUILLAUME, J. H. A., HAMILTON, S. H., JAKEMAN, A. J., MARSILI-LIBELLI, S., NEWHAM, L. T. H., NORTON, J. P., PERRIN, C., PIERCE, S. A.,

ROBSON, B., SEPPELT, R., VOINOV, A. A., FATH, B. D. & ANDREASSIAN, V. 2013. Characterising performance of environmental models. *Environmental Modelling & Software,* 40**,** 1-20.

BISHOP, C. M. 1995. Neural networks for pattern recognition.

BORDES, A., GLOROT, X., WESTON, J. & BENGIO, Y. Joint Learning of Words and Meaning Representations for Open-Text Semantic Parsing. AISTATS, 2012. 423-424.

BRANKE, J. Evolutionary algorithms for neural network design and training. In Proceedings of the First Nordic Workshop on Genetic Algorithms and its Applications, 1995. Citeseer.

BRAUER, M., HOEK, G., VAN VLIET, P., MELIEFSTE, K., FISCHER, P. H., WIJGA, A., KOOPMAN, L. P., NEIJENS, H. J., GERRITSEN, J., KERKHOF, M., HEINRICH, J., BELLANDER, T. & BRUNEKREEF, B. 2002. Air pollution from traffic and the development of respiratory infections and asthmatic and allergic symptoms in children. *Am J Respir Crit Care Med,* 166**,** 1092-8.

BREIMAN, L. 2001. Random forests. *Machine learning,* 45**,** 5-32.

BREIMAN, L. 2006. randomForest: Breiman and Cutler's random forests for classification and regression. URL http://stat-www. berkeley. edu/users/breiman/RandomForests, R package version.

BROWN, T., DASSONVILLE, C., DERBEZ, M., RAMALHO, O., KIRCHNER, S., CRUMP, D. & MANDIN, C. 2015. Relationships between socioeconomic and lifestyle factors and indoor air quality in French dwellings. *Environmental Research,* 140**,** 385-396.

BRUNEKREEF, B., BEELEN, R., HOEK, G., SCHOUTEN, L., BAUSCH-GOLDBOHM, S., FISCHER, P., ARMSTRONG, B., HUGHES, E., JERRETT, M. & VAN DEN BRANDT, P. 2009. Effects of long-term exposure to traffic-related air pollution on respiratory and cardiovascular mortality in the Netherlands: the NLCS-AIR study. *Res Rep Health Eff Inst,* 139**,** 5-71.

BUUREN, S. & GROOTHUIS-OUDSHOORN, K. 2011. mice: Multivariate imputation by chained equations in R. *Journal of statistical software,* 45.

BYUN, D. W., CHING, J. K. S., NOVAK, J. & YOUNG, J. 1998. Development and implementation of the EPA's models-3 initial operating version: Community multi-scale air quality (CMAQ) model. *Air Pollution Modeling and Its Application Xii,* 22**,** 357-368.

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.

CARRUTHERS, D. J., DICKSON, P., MCHUGH, C. A., NIXON, S. G. & OATES, W. 2001. Determination of compliance with UK and EU air quality objectives from high-resolution pollutant concentration maps calculated using ADMS-Urban. *International Journal of Environment and Pollution,* 16**,** 460-471.

CARRUTHERS, D. J., EDMUNDS, H. A., MCHUGH, C. A., RICHES, P. J. & SINGLES, R. J. 1997. ADMS Urban - an integrated air quality modelling system for local government. *Air Pollution V***,** 45-58.

CARSLAW D, APSIMON H, BEEVERS S, BROOKES D, CARRUTHERS D, COOKE S, KITWIROON N, OXLEY T, STEDMAN J, A. & STOCKER J 2013. Defra Phase 2 urban model evaluation. Defra Air Quality Library.

CARSLAW, D., APSIMON, H., BEEVERS, S., BROOKES, D., CARRUTHERS, D., COOKE, S., KITWIROON, N., OXLEY, T., STEDMAN, J. & STOCKER, J. 2014. Defra Phase 2 urban model evaluation.

CARSLAW, D. C. & BEEVERS, S. D. 2013. Characterising and understanding emission sources using bivariate polar plots and k-means clustering. *Environmental Modelling & Software,* 40**,** 325-329.

CARSLAW, D. C., BEEVERS, S. D., ROPKINS, K. & BELL, M. C. 2006. Detecting and quantifying aircraft and other on-airport contributions to ambient nitrogen oxides in the vicinity of a large international airport. *Atmospheric Environment,* 40**,** 5424-5434.

CARSLAW, D. C., BEEVERS, S. D. & TATE, J. E. 2007. Modelling and assessing trends in traffic-related emissions using a generalised additive modelling approach. *Atmospheric Environment,* 41**,** 5289-5299.

CARSLAW, D. C. & ROPKINS, K. 2012. openair — An R package for air quality data analysis. *Environmental Modelling & Software,* 27-28**,** 52-61.

CARSLAW, D. C. & TAYLOR, P. J. 2009. Analysis of air pollution data at a mixed source location using boosted regression trees. *Atmospheric Environment,* 43**,** 3563-3570.

CARSLAW, D. C. R., K. 2012. openair - Data Analysis Tools for the Air Quality Community. *R Journal,* 4**,** 20-29.

CATALANO, M., GALATIOTO, F., BELL, M., NAMDEO, A. & BERGANTINO, A. S. 2016. Improving the prediction of air pollution peak episodes generated by urban transport networks. *Environmental Science & Policy,* 60**,** 69-83.

CERC 2013. *ADMS-Roads Air Quality Management Systems User Guide* Cambridge Environmental Research Consultants Ltd

CHANDRASHEKAR, G. & SAHIN, F. 2014. A survey on feature selection methods. *Computers & Electrical Engineering,* 40**,** 16-28.

CHANG, J. C. & HANNA, S. R. 2004. Air quality model performance evaluation. *Meteorology and Atmospheric Physics,* 87**,** 167-196.

CHARRON, A. & HARRISON, R. M. 2005. Fine (PM2.5) and coarse (PM2.5-10) particulate matter on a heavily trafficked London highway: sources and processes. *Environmental Science & Technology,* 39**,** 7768.

CHAVE, J. & LEVIN, S. 2003. Scale and scaling in ecological and economic systems. *Environmental and Resource Economics,* 26**,** 527-557.

CHEN, Y., SHI, R., SHU, S. & GAO, W. 2013a. Ensemble and enhanced PM10 concentration forecast model based on stepwise regression and wavelet analysis. *Atmospheric Environment,* 74**,** 346-359.

CHEN, Y. Y., SHI, R. H., SHU, S. J. & GAO, W. 2013b. Ensemble and enhanced PM10 concentration forecast model based on stepwise regression and wavelet analysis. *Atmospheric Environment,* 74**,** 346-359.

CHERKASSKY, V. & MA, Y. 2004. Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Networks,* 17**,** 113-126.

CIMORELLI, A. J., PERRY, S. G., VENKATRAM, A., WEIL, J. C., PAINE, R. J. & PETERS, W. D. 1998. AERMOD–Description of model formulation.

CIREGAN, D., MEIER, U. & SCHMIDHUBER, J. Multi-column deep neural networks for image classification. Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, 2012. IEEE, 3642-3649.

CIREŞAN, D. C., MEIER, U. & SCHMIDHUBER, J. Transfer learning for Latin and Chinese characters with deep neural networks. The 2012 International Joint Conference on Neural Networks (IJCNN), 2012. IEEE, 1-6.

COLLS, J. 2002a. Gaseous air pollutants. *Air Pollution.* Routledge.

COLLS, J. 2002b. Meteorology and modelling. *Air Pollution.* First published 2002 by Spon Press 11 New Fetter Lane, London EC4P 4EE: Routledge.

COLVILE, R. N., HUTCHINSON, E. J. & WARREN, R. F. 2002. Chapter 6 The transport sector as a source of air pollution. *In:* JILL AUSTIN, P. B. & WILLIAM, S. (eds.) *Developments in Environmental Science.* Elsevier.

COMEAP 2010. The mortality effects of long-term exposure to particulate air pollution in the UK. *The Committee on the Medical Effects of Air Pollutants (COMEAP)*

COMEAP 2011. Review of the UK Air Quality Index

DA MOTA, B., TUDORAN, R., COSTAN, A., VAROQUAUX, G., BRASCHE, G., CONROD, P., LEMAITRE, H., PAUS, T., RIETSCHEL, M., FROUIN, V., POLINE, J. B., ANTONIU, G., THIRION, B. & CONSORTIUM, I. 2014. Machine learning patterns for neuroimaging-genetic studies in the cloud. *Frontiers in Neuroinformatics,* 8.

DABBERDT, W. F., CARROLL, M. A., APPLEBY, W., BAUMGARDNER, D., CARMICHAEL, G., DAVIDSON, P., DORAN, J. C., DYE, T. S., GRIMMOND, S., MIDDLETON, P., NEFF, W. & ZHANG, Y. 2006. USWRP workshop on air quality forecasting. *Bulletin of the American Meteorological Society,* 87**,** 215-221.

DAN FORESEE, F. & HAGAN, M. T. Gauss-Newton approximation to Bayesian learning. Neural Networks,1997., International Conference on, 9-12 Jun 1997 1997. 1930-1935 vol.3.

DE GENNARO, G., TRIZIO, L., DI GILIO, A., PEY, J., PEREZ, N., CUSACK, M., ALASTUEY, A. & QUEROL, X. 2013. Neural network model for the prediction of PM10 daily concentrations in two sites in the Western Mediterranean. *Sci Total Environ,* 463-464**,** 875-83.

DE PAULA, P. H. M., MATEUS, V. L., ARARIPE, D. R., DUYCK, C. B., SAINT'PIERRE, T. D. & GIODA, A. 2015. Biomonitoring of metals for air pollution assessment using a hemiepiphyte herb (Struthanthus flexicaulis). *Chemosphere,* 138**,** 429-437.

DEAN, J., CORRADO, G., MONGA, R., CHEN, K., DEVIN, M., MAO, M., SENIOR, A., TUCKER, P., YANG, K. & LE, Q. V. Large scale distributed deep networks. Advances in neural information processing systems, 2012. 1223-1231.

DEBRY, E. & MALLET, V. 2014. Ensemble forecasting with machine learning algorithms for ozone, nitrogen dioxide and PM10 on the Prev'Air platform. *Atmospheric Environment,* 91**,** 71-84.

DEFRA 2007. The Air Quality Strategy for England, Scotland, Wales and Northern Ireland. *Department for Environment, Food and Rural Affairs in partnership with the Scottish Executive, Welsh Assembly Government and Department of the Environment Northern Ireland*.

DEFRA 2010. UK air quality Forecasting: Annual Report. *AEA Report For Defra and the Devolved Administrations,* AEAT/ENV/R/3260 ED48946 1.

DEFRA 2013. Emissions of Air Quality Pollutants 1970 - 2011 *In:* DEPARTMENT FOR ENVIRONMENT, F. A. R. A. U. (ed.). AQPI Summary Report.

DEFRA. 2015a. *Air Pollution in the UK 2014.*

DEFRA. 2015b. *DEFRA, Emissions Factors Toolkit* [Online]. Available: http://laqm.defra.gov.uk/review-and-assessment/tools/emissions-factors-toolkit.html [Accessed 06/04/2015 2016].

DEKA, P., BHUYAN, P., DAIMARI, R., SARMA, K. P. & HOQUE, R. R. 2016. Metallic species in PM10 and source apportionment using PCA-MLR modeling over mid-Brahmaputra Valley. *Arabian Journal of Geosciences,* 9.

DENG, L. 2014. A tutorial survey of architectures, algorithms, and applications for deep learning. *APSIPA Transactions on Signal and Information Processing,* 3**,** e2 (29 pages).

DERKSEN, S. & KESELMAN, H. J. 1992. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology,* 45**,** 265-282.

DERVILIS, N., CHOI, M., TAYLOR, S., BARTHORPE, R., PARK, G., FARRAR, C. & WORDEN, K. 2014. On damage diagnosis for a wind turbine blade using pattern recognition. *Journal of sound and vibration,* 333**,** 1833-1850.

DERWENT, D., FRASER, A., ABBOTT, J., JENKIN, M., WILLIS, P. & MURRELLS, T. 2010. Evaluating the Performance of Air Quality Models *In:* (DEFRA), D. F. E. F. A. R. A. (ed.). Department for Environment Food and Rural Affairs (DEFRA): Department for Environment Food and Rural Affairs (DEFRA).

DIAZ-DE-QUIJANO, M., JOLY, D., GILBERT, D. & BERNARD, N. 2014. A more cost-effective geomatic approach to modelling PM10 dispersion across Europe. *Applied Geography,* 55**,** 108-116.

DIMITRIADOU, E., HORNIK, K., LEISCH, F., MEYER, D. & WEINGESSEL, A. 2008. Misc functions of the Department of Statistics (e1071), TU Wien. *R package*, 1.5-24.
DING, S., JIA, W., SU, C., ZHANG, L. & LIU, L. 2011a. Research of neural network algorithm based on factor analysis and cluster analysis. *Neural Computing and Applications,* 20, 297-302.
DING, S., LI, H., SU, C., YU, J. & JIN, F. 2011b. Evolutionary artificial neural networks: a review. *Artificial Intelligence Review,* 39, 251-260.
DING, S., SU, C. & YU, J. 2011c. An optimizing BP neural network algorithm based on genetic algorithm. *Artificial Intelligence Review,* 36, 153-162.
DING, S., ZHAO, H., ZHANG, Y., XU, X. & NIE, R. 2015. Extreme learning machine: algorithm, theory and applications. *Artificial Intelligence Review,* 44, 103-115.
DOMINGOS, P. 2012. A Few Useful Things to Know About Machine Learning. *Communications of the Acm,* 55, 78-87.
DOOVE, L., VAN BUUREN, S. & DUSSELDORP, E. 2014. Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis,* 72, 92-104.
DORA, C. & PHILLIPS, M. 2000. *Transport, environment and health*, WHO Regional Office Europe.
DORE, A. J., CARSLAW, D. C., BRABAN, C., CAIN, M., CHEMEL, C., CONOLLY, C., DERWENT, R. G., GRIFFITHS, S. J., HALL, J., HAYMAN, G., LAWRENCE, S., METCALFE, S. E., REDINGTON, A., SIMPSON, D., SUTTON, M. A., SUTTON, P., TANG, Y. S., VIENO, M., WERNER, M. & WHYATT, J. D. 2015. Evaluation of the performance of different atmospheric chemical transport models and inter-comparison of nitrogen and sulphur deposition estimates for the UK. *Atmospheric Environment,* 119, 131-143.
DOS SANTOS-JUUSELA, V., PETÄJÄ, T., KOUSA, A. & HÄMERI, K. 2013. Spatial-temporal variations of particle number concentrations between a busy street and the urban background. *Atmospheric Environment,* 79, 324-333.
ELANGASINGHE, M. A., SINGHAL, N., DIRKS, K. N. & SALMOND, J. A. 2014a. Development of an ANN-based air pollution forecasting system with explicit knowledge through sensitivity analysis. *Atmospheric Pollution Research,* 5.
ELANGASINGHE, M. A., SINGHAL, N., DIRKS, K. N., SALMOND, J. A. & SAMARASINGHE, S. 2014b. Complex time series analysis of PM10 and PM2.5 for a coastal site using artificial neural network modelling and k-means clustering. *Atmospheric Environment,* 94, 106-116.
ELITH, J., LEATHWICK, J. R. & HASTIE, T. 2008. A working guide to boosted regression trees. *Journal of Animal Ecology,* 77, 802-813.
EPA 1999. Air quality index reporting: final rule. *In:* FEDERAL REGISTER. PART III, C. P. A. S., RESEARCH TRIANGLE PARK. (ed.).
ESPLIN, G. J. 1995. Approximate explicit solution to the general line source problem. *Atmospheric Environment,* 29, 1459-1463.
FEI, L., PING, M. & MING, Y. A validation methodology for AI simulation models. Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on, 18-21 Aug. 2005 2005. 4083-4088 Vol. 7.
FOUSKAKIS, D. & DRAPER, D. 2002. Stochastic optimization: a review. *International Statistical Review,* 70, 315-349.
FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. 2009. glmnet: Lasso and elastic-net regularized generalized linear models. *R package version,* 1.
FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software,* 33, 1-22.
FRIEDMAN, J. H. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics,* 29, 1189-1232.
FRIEDMAN, J. H. & MEULMAN, J. J. 2003. Multiple additive regression trees with application in epidemiology. *Stat Med,* 22, 1365-81.

FU, A., AIELLO, S., RAO, A., KRALJEVIC, T., MAJ, P., KRALJEVIC, M. T. & JAVA, S. 2014. Package 'h2o'.

GARDNER, M. W. & DORLING, S. R. 1998. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment,* 32**,** 2627-2636.

GARDNER, M. W. & DORLING, S. R. 2000. Statistical surface ozone models: an improved methodology to account for non-linear behaviour. *Atmospheric Environment,* 34**,** 21-34.

GOKHALE, S. & KHARE, M. 2004. A review of deterministic, stochastic and hybrid vehicular exhaust emission models. *International Journal of Transport Management,* 2**,** 59-74.

GOSSO, A. & MARTINEZ-DE-PISON, F. 2012. elmNN: Implementation of ELM (Extreme Learning Machine) algorithm for SLFN (Single Hidden Layer Feedforward Neural Networks). *R package version,* 1.

GÖTSCHI, T., HAZENKAMP-VON ARX, M. E., HEINRICH, J., BONO, R., BURNEY, P., FORSBERG, B., JARVIS, D., MALDONADO, J., NORBÄCK, D., STERN, W. B., SUNYER, J., TORÉN, K., VERLATO, G., VILLANI, S. & KÜNZLI, N. 2005. Elemental composition and reflectance of ambient fine particles at 21 European locations. *Atmospheric Environment,* 39**,** 5947-5958.

GOWERS, A., MILLER, B. & STEDMAN, J. 2014. *Estimating local mortality burdens associated with particulate air pollution*.

GRASKOW, B. R., KITTELSON, D., ABDUL-KHALEK, I., AHMADI, M. & MORRIS, J. 1998. *Characterization of exhaust particulate emissions from a spark ignition engine.* University of Minnesota.

GUERREIRO, C. B. B., FOLTESCU, V. & DE LEEUW, F. 2014. Air quality status and trends in Europe. *Atmospheric Environment,* 98**,** 376-384.

GUO, X. Y., LI, C., GAO, Y., TANG, L., BRIKI, M., DING, H. J. & JI, H. B. 2016a. Sources of organic matter (PAHs and n-alkanes) in PM2.5 of Beijing in haze weather analyzed by combining the C-N isotopic and PCA-MLR analyses. *Environmental Science-Processes & Impacts,* 18**,** 314-322.

GUO, Y., LIU, Y., OERLEMANS, A., LAO, S., WU, S. & LEW, M. S. 2016b. Deep learning for visual understanding: A review. *Neurocomputing,* 187**,** 27-48.

HAGAN, M. T. & MENHAJ, M. B. 1994. Training feedforward networks with the Marquardt algorithm. *Neural Networks, IEEE Transactions on,* 5**,** 989-993.

HARRIS, S. J. & MARICQ, M. M. 2001. Signature size distributions for diesel and gasoline engine exhaust particulate matter. *Journal of Aerosol Science,* 32**,** 749-764.

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. 2008a. *Elements of Statistical Learning, The: Data Mining, Inference, and Prediction*, Springer.

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. 2008b. The elemnts of statistical learning Data Mining, Inference, and Prediction. *Springer,* Second Edition.

HAYKIN, S. 2005. Neural Networks - A Comprehensive Foundation Second Edition. *Pearson (Education), Pearson Printice Hall Publication*.

HE, H.-D., LU, W.-Z. & XUE, Y. 2014. Prediction of particulate matter at street level using artificial neural networks coupling with chaotic particle swarm optimization algorithm. *Building and Environment,* 78**,** 111-117.

HE, H. D., LU, W. Z. & XUE, Y. 2015. Prediction of particulate matters at urban intersection by using multilayer perceptron model based on principal components. *Stochastic Environmental Research and Risk Assessment,* 29**,** 2107-2114.

HEINRICH, J., TOPP, R., GEHRING, U. & THEFELD, W. 2005. Traffic at residential address, respiratory health, and atopy in adults: the National German Health Survey 1998. *Environmental Research,* 98**,** 240-249.

HERNÁNDEZ-LOBATO, J. M. & ADAMS, R. P. 2015. Probabilistic backpropagation for scalable learning of bayesian neural networks. *arXiv preprint arXiv:1502.05336*.

HERTEL, O. & BERKOWICZ, R. 1989. Operational Street Pollution Model (OSPM). Evaluation of the model on data from St. Olavs Street in Oslo. National Environmental Research Institute, Roskilde. *NERI Technical report No A-135*.

HINTON, G. E., OSINDERO, S. & TEH, Y.-W. 2006. A fast learning algorithm for deep belief nets. *Neural computation,* 18**,** 1527-1554.

HIRTL, M. & BAUMANN-STANZER, K. 2007. Evaluation of two dispersion models (ADMS-Roads and LASAT) applied to street canyons in Stockholm, London and Berlin. *Atmospheric Environment,* 41**,** 5959-5971.

HOI, K. I., YUEN, K. V. & MOK, K. M. 2010. Is a Complex Neural Network Based Air Quality Prediction Model Better Than a Simple One? A Bayesian Point of View.

HOLGATE, S., GRIGG, J., AGIUS, R., ASHTON, J. R., CULLINAN, P., EXLEY, K., FISHWICK, D., FULLER, G., GOKANI, N. & GRIFFITHS, C. Every breath we take: The lifelong impact of air pollution, Report of a working party. 2016. Royal College of Physicians.

HOLMES, N. S. & MORAWSKA, L. 2006. A review of dispersion modelling and its application to the dispersion of particles: an overview of different dispersion models available. *Atmospheric Environment,* 40**,** 5902-5928.

HONG, X. & CHEN, S. 2015. Elastic net orthogonal forward regression. *Neurocomputing,* 148**,** 551-560.

HUANG, G., HUANG, G.-B., SONG, S. & YOU, K. 2015. Trends in extreme learning machines: A review. *Neural Networks,* 61**,** 32-48.

HUANG, G., SONG, S., GUPTA, J. N. & WU, C. 2014. Semi-supervised and unsupervised extreme learning machines. *Cybernetics, IEEE Transactions on,* 44**,** 2405-2417.

HUANG, G. B., ZHOU, H. M., DING, X. J. & ZHANG, R. 2012. Extreme Learning Machine for Regression and Multiclass Classification. *Ieee Transactions on Systems Man and Cybernetics Part B-Cybernetics,* 42**,** 513-529.

HUANG, G. B., ZHU, Q. Y. & SIEW, C. K. 2006. Extreme learning machine: Theory and applications. *Neurocomputing,* 70**,** 489-501.

HUSSEIN, T., JOHANSSON, C., KARLSSON, H. & HANSSON, H.-C. 2008. Factors affecting non-tailpipe aerosol particle emissions from paved roads: On-road measurements in Stockholm, Sweden. *Atmospheric Environment,* 42**,** 688-702.

JAKEMAN, A. J., LETCHER, R. A. & NORTON, J. P. 2006. Ten iterative steps in development and evaluation of environmental models. *Environmental Modelling & Software,* 21**,** 602-614.

JAMES, G., WITTEN, D. & HASTIE, T. 2014. An Introduction to Statistical Learning: With Applications in R.

JEAN, S., CHO, K., MEMISEVIC, R. & BENGIO, Y. 2014. On using very large target vocabulary for neural machine translation. *arXiv preprint arXiv:1412.2007*.

JOHANSSON, C., NORMAN, M. & GIDHAGEN, L. 2007. Spatial & temporal variations of PM10 and particle number concentrations in urban air. *Environmental Monitoring and assessment,* 127**,** 477-487.

JONES, A. & HARRISON, R. 2006. Estimation of the emission factors of particle number and mass fractions from traffic at a site where mean vehicle speeds vary over short distances. *Atmospheric Environment,* 40**,** 7125-7137.

JORDAN, M. I. & MITCHELL, T. M. 2015. Machine learning: Trends, perspectives, and prospects. *Science,* 349**,** 255-260.

KANG, L., JIAN-XUN, P. & IRWIN, G. W. 2005. A Fast Nonlinear Model Identification Method. *Automatic Control, IEEE Transactions on,* 50**,** 1211-1216.

KARAMIZADEH, S., ABDULLAH, S. M., MANAF, A. A., ZAMANI, M. & HOOMAN, A. 2013. An Overview of Principal Component Analysis. *Journal of Signal and Information Processing,* 4**,** 173.

KEOGH, D. U., FERREIRA, L. & MORAWSKA, L. 2009. Development of a particle number and particle mass vehicle emissions inventory for an urban fleet. *Environmental Modelling and Software,* 24**,** 1323-1331.
KIM, J. J., SMORODINSKY, S., LIPSETT, M., SINGER, B. C., HODGSON, A. T. & OSTRO, B. 2004. Traffic-related air pollution near busy roads: the East Bay Children's Respiratory Health Study. *Am J Respir Crit Care Med,* 170**,** 520-6.
KITTELSON, D. B. 1998. ENGINES AND NANOPARTICLES: A REVIEW. *J. Aerosol Sci. ,* Vol. 29**,** pp. 575 - 588.
KOROBILIS, D. 2013. Hierarchical shrinkage priors for dynamic regressions with many predictors. *International Journal of Forecasting,* 29**,** 43-59.
KRIVTSOV, V., HOWARTH, M. J. & JONES, S. E. 2009. Characterising observed patterns of suspended particulate matter and relationships with oceanographic and meteorological variables: Studies in Liverpool Bay. *Environmental Modelling & Software,* 24**,** 677-685.
KRIZHEVSKY, A., SUTSKEVER, I. & HINTON, G. E. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 2012. 1097-1105.
KRZYZANOWSKI, M., DORA, C. & BRUCE, N. 2011. Improvement of air quality in low-income countries. *Lancet,* 377**,** 1920-1920.
KUHN, M. 2008. Caret package. *Journal of Statistical Software,* 28.
KUHN, M. 2012. The caret package.
KUHN, M. & JOHNSON, K. 2013. *Applied Predictive Modeling*, Springer.
KUMAR, A. & GOYAL, P. 2013. Forecasting of Air Quality Index in Delhi Using Neural Network Based on Principal Component Analysis. *Pure and Applied Geophysics,* 170**,** 711-722.
KUMAR, P., ROBINS, A., VARDOULAKIS, S. & BRITTER, R. 2010. A review of the characteristics of nanoparticles in the urban atmosphere and the prospects for developing regulatory controls. *Atmospheric Environment,* 44**,** 5035-5052.
LAEI 2014. London Emissions (LAEI) Cleaner air for london.
LAGZI, I., MÉSZÁROS, R., GELYBÓ, G. & LEELŐSSY, Á. 2013. *Atmospheric Chemistry,* Hungary, Eötvös Loránd University.
LAWAL, A. O., ZHANG, M., DITTMAR, M., LULLA, A. & ARAUJO, J. A. 2015. Heme oxygenase-1 protects endothelial cells from the toxicity of air pollutant chemicals. *Toxicol Appl Pharmacol,* 284**,** 281-91.
LECUN, Y., BENGIO, Y. & HINTON, G. 2015. Deep learning. *Nature,* 521**,** 436-444.
LEGATES, D. R. & MCCABE, G. J. 2012. A refined index of model performance: a rejoinder. *International Journal of Climatology*.
LELIEVELD, J., EVANS, J. S., FNAIS, M., GIANNADAKI, D. & POZZER, A. 2015. The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature,* 525**,** 367-371.
LEUNG, M. K., XIONG, H. Y., LEE, L. J. & FREY, B. J. 2014. Deep learning of the tissue-regulated splicing code. *Bioinformatics,* 30**,** i121-i129.
LEWIS, A. C., CARSLAW, D. C. & KELLY, F. J. 2015. Vehicle emissions: Diesel pollution long under-reported. *Nature,* 526**,** 195-195.
LIMA, A. R., CANNON, A. J. & HSIEH, W. W. 2013. Nonlinear regression in environmental sciences by support vector machines combined with evolutionary strategy. *Computers & Geosciences,* 50**,** 136-144.
LIMA, A. R., CANNON, A. J. & HSIEH, W. W. 2015. Nonlinear regression in environmental sciences using extreme learning machines: A comparative evaluation. *Environmental Modelling & Software,* 73**,** 175-188.
LIMBACH, L. K., WICK, P., MANSER, P., GRASS, R. N., BRUININK, A. & STARK, W. J. 2007. Exposure of engineered nanoparticles to human lung epithelial cells: influence of chemical composition and catalytic activity on oxidative stress. *Environmental Science & Technology,* 41**,** 4158.

LIN, S.-W., TSENG, T.-Y., CHOU, S.-Y. & CHEN, S.-C. 2008. A simulated-annealing-based approach for simultaneous parameter optimization and feature selection of back-propagation networks. *Expert Systems with Applications,* 34**,** 1491-1499.

LINDGREN, A., STROH, E., NIHLEN, U., MONTNEMERY, P., AXMON, A. & JAKOBSSON, K. 2009. Traffic exposure associated with allergic asthma and allergic rhinitis in adults. A cross-sectional study in southern Sweden. *International Journal of Health Geographics,* 8.

LIU, J. Y. 2002. Evaluation practice on air-borne GPS data quality. *Survey Review,* 36**,** 463-469.

LONDON ASSEMBLY. 2014. *London Assembly ULEZ response* [Online]. Available: https://www.london.gov.uk/sites/default/files/gla_migrate_files_destination/London%20Assembly%20ULEZ%20response_0.pdf [Accessed 02/08/2016 2016].

MALHOTRA, R. 2015. A systematic review of machine learning techniques for software fault prediction. *Applied Soft Computing,* 27**,** 504-518.

MALOHLAVA, C. C. J. L. M. & HANK, V. P. 2015. Gradient Boosted Machines with H2O.

MASIOL, M. & HARRISON, R. M. 2015. Quantification of air quality impacts of London Heathrow Airport (UK) from 2005 to 2012. *Atmospheric Environment,* 116**,** 308-319.

MAY, R., DANDY, G. & MAIER, H. 2011. *Review of input variable selection methods for artificial neural networks*, INTECH Open Access Publisher.

MCCONNELL, R., BERHANE, K., YAO, L., JERRETT, M., LURMANN, F., GILLILAND, F., KUNZLI, N., GAUDERMAN, J., AVOL, E., THOMAS, D. & PETERS, J. 2006. Traffic, susceptibility, and childhood asthma. *Environ Health Perspect,* 114**,** 766-72.

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**,** 3-4.

MICHAL KRZYZANOWSKI & SCHNEIDER, B. K.-D. A. J. 2005. Health effects of transport related air pollution who europe.

MIDAS LAND SURFACE, M. O. 2013. *Met Office Integrated Data Archive System (MIDAS) Land and Marine Surface Stations Data (1853-current)* [Online]. Available: http://badc.nerc.ac.uk/view/badc.nerc.ac.uk__ATOM__dataent_ukmo-midas [Accessed 24/06/2013 2013].

MIKOLOV, T. & DEAN, J. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*.

MISHRA, D., GOYAL, P. & UPADHYAY, A. 2015. Artificial intelligence based approach to forecast PM2.5 during haze episodes: A case study of Delhi, India. *Atmospheric Environment,* 102**,** 239-248.

MOGHADDAM, A. H., MOGHADDAM, M. H. & ESFANDYARI, M. 2016. Stock market index prediction using artificial neural network. *Journal of Economics, Finance and Administrative Science*.

MOISEN, G. G., FREEMAN, E. A., BLACKARD, J. A., FRESCINO, T. S., ZIMMERMANN, N. E. & EDWARDS JR, T. C. 2006. Predicting tree species presence and basal area in Utah: A comparison of stochastic gradient boosting, generalized additive models, and tree-based methods. *Ecological Modelling,* 199**,** 176-187.

MOORCROFT, S. & MARNER, B. 2011. Review of the air quality monitoring network in London Ref: GLA 80090. *In:* LAXEN, P. D. (ed.). Air Quality Consultants Ltd: Greater London Authority

MORAWSKA, L., THOMAS, S., BOFINGER, N., WAINWRIGHT, D. & NEALE, D. 1998. Comprehensive characterization of aerosols in a subtropical urban atmosphere: particle size distribution and correlation with gaseous pollutants. *Atmospheric Environment,* 32**,** 2467-2478.

MOSCHANDREAS, D. J., CHOI, S. W. & MECKLER, M. M. 1996. Indoor air quality and the variable-air-volume bypass filtration system: Chamber experiment. *Environment International,* 22**,** 149-158.

NAEI 2012. Air Quality Pollutant Inventories for England, Scotland, Wales and Northern Ireland: 1990 – 2010. *Department for Environment, Food and Rural Affairs,The Scottish Government, The Welsh Government, The Northern Ireland Department of Environment*.

NAEI. 2014. *UK NAEI - National Atmospheric Emissions Inventory* [Online]. [Accessed 02/08/2016 2016].

NAGENDRA, S. M. S. & KHARE, M. 2005. Modelling urban air quality using artificial neural network. *Clean Technologies and Environmental Policy.*

NAGENDRA, S. M. S. & KHARE, M. 2006. Artificial neural network approach for modelling nitrogen dioxide dispersion from vehicular exhaust emissions. *Ecological Modelling,* 190**,** 99-115.

NAMDEO, A. & BELL, M. C. 2005. Characteristics and health implications of fine and coarse particulates at roadside, urban background and rural sites in UK. *Environment International,* 31**,** 565-573.

NAO 2009. Air quality; Final Report; Briefing for the house of commons Environtal Audit Committee. *National Audit Office United kingdom*

NATIONAL RESEARCH COUNCIL COMMITTEE ON MODELS IN THE REGULATORY DECISION PROCESS 2007. *Models in Environmental Regulatory Decision Making*, National Academies Press: Washhington, DC, USA.

NATIONAL RESEARCH COUNCIL, N. 2007. Models in Environmental Regulatory Decision Making

The National Academies Press: Washington, DC, USA.

NEMMAR, A., HOET, P. H. M., VANQUICKENBORNE, B., DINSDALE, D., THOMEER, M., HOYLAERTS, M., VANBILLOEN, H., MORTELMANS, L. & NEMERY, B. 2002. Passage of inhaled particles into the blood circulation in humans. *Circulation,* 105**,** 411-414.

NGUYEN, D. L. 2014. A Brief Review of Air Quality Models and Their Applications.

NOWACK, B. & BUCHELI, T. D. 2007. Occurrence, behavior and effects of nanoparticles in the environment. *Environmental Pollution,* 150**,** 5-22.

ONG, B. T., SUGIURA, K. & ZETTSU, K. 2016. Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting PM2.5. *Neural Computing and Applications,* 27**,** 1553-1566.

P. THUNIS, E. G., S. GALMARINI 2011. A procedure for air quality models benchmarking. *Joint Research Centre, Ispra*.

PANDEY, G., ZHANG, B. & JIAN, L. 2013. Predicting submicron air pollution indicators: a machine learning approach. *Environmental Science: Processes & Impacts,* 15**,** 996-1005.

PANT, P. & HARRISON, R. M. 2013. Estimation of the contribution of road traffic emissions to particulate matter concentrations from field measurements: A review. *Atmospheric Environment,* 77**,** 78-97.

PEKKANEN, J., TIMONEN, K. L., RUUSKANEN, J., REPONEN, A. & MIRME, A. 1997. Effects of Ultrafine and Fine Particles in Urban Air on Peak Expiratory Flow among Children with Asthmatic Symptoms. *Environmental Research,* 74**,** 24-33.

PELLICCIONI, A. & TIRABASSI, T. 2006. Air dispersion model and neural network: A new perspective for integrated models in the simulation of complex situations. *Environmental Modelling & Software,* 21**,** 539-546.

PERRONE, M., COOPER, L. & MAMMONE, R. 1993. Neural networks for speech and image processing. *When Networks Disagree: Ensemble Methods for Hybrid Neural Networks, Chapman Hall*.

PEY, J., QUEROL, X., ALASTUEY, A., RODRÍGUEZ, S., PUTAUD, J. P. & VAN DINGENEN, R. 2009. Source apportionment of urban fine and ultra-fine particle number concentration in a Western Mediterranean city. *Atmospheric Environment,* 43**,** 4407-4415.

PIRES, J., MARTINS, F., SOUSA, S., FERRAZ, M. & PEREIRA, M. 2008. Prediction of the daily mean PM10 concentrations using linear models. *American Journal of Environmental Sciences,* 4**,** 445.

PUGALENTHI, G., TANG, K., SUGANTHAN, P. N., ARCHUNAN, G. & SOWDHAMINI, R. 2007. A machine learning approach for the identification of odorant binding proteins from sequence-derived properties. *Bmc Bioinformatics,* 8.

QUTEISHAT, A. & LIM, C. P. 2008. A modified fuzzy min–max neural network with rule extraction and its application to fault detection and classification. *Applied Soft Computing,* 8**,** 985-995.

R DEVELOPMENT CORE TEAM 2015. R 3.2. 1. R Project for Statistical Computing Vienna, Austria.

RAGOSTA, M., D'EMILIO, M. & GIORGIO, G. A. 2015. Input strategy analysis for an air quality data modelling procedure at a local scale based on neural network. *Environ Monit Assess,* 187**,** 307.

REN, J. S. & XU, L. 2015. On vectorization of deep convolutional neural networks for vision tasks. *arXiv preprint arXiv:1501.07338*.

RIDGEWAY, G., SOUTHWORTH, M. H. & RUNIT, S. 2013. Package 'gbm'. *Viitattu,* 10**,** 2013.

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**,** 3850-3857.

ROSE, D., WEHNER, B., KETZEL, M., ENGLER, C., VOIGTLÄNDER, J., TUCH, T. & WIEDENSOHLER, A. 2006. Atmospheric number size distributions of soot particles and estimation of emission factors. *Atmospheric Chemistry and Physics,* 6**,** 1021.

RUSSO, A., RAISCHEL, F. & LIND, P. G. 2013. Air quality prediction using optimal neural networks with stochastic variables. *Atmospheric Environment,* 79**,** 822-830.

SALISBURY, E., THISTLETHWAITE, G., PANG, Y. & BAILEY, R. 2014. Air Quality Pollutant Inventories, for England, Scotland, Wales and Northern. *Atmospheric Environment,* 38**,** 2163-2176.

SAMPSON, P. D., RICHARDS, M., SZPIRO, A. A., BERGEN, S., SHEPPARD, L., LARSON, T. V. & KAUFMAN, J. D. 2013. A regionalized national universal kriging model using Partial Least Squares regression for estimating annual PM2.5 concentrations in epidemiology. *Atmospheric Environment,* 75**,** 383-392.

SANCHEZ, A. S., NIETO, P. J. G., FERNANDEZ, P. R., DIAZ, J. J. D. & IGLESIAS-RODRIGUEZ, F. J. 2011. Application of an SVM-based regression model to the air quality study at local scale in the Aviles urban area (Spain). *Mathematical and Computer Modelling,* 54**,** 1453-1466.

SANCHEZ, A. S., NIETO, P. J. G., IGLESIAS-RODRIGUEZ, F. J. & VILAN, J. A. V. 2013. Nonlinear Air Quality Modeling Using Support Vector Machines in Gijon Urban Area (Northern Spain) at Local Scale. *International Journal of Nonlinear Sciences and Numerical Simulation,* 14**,** 291-305.

SÁNCHEZ JIMÉNEZ, A., HEAL, M. R. & BEVERLAND, I. J. 2012. Correlations of particle number concentrations and metals with nitrogen oxides and other traffic-related air pollutants in Glasgow and London. *Atmospheric Environment,* 54**,** 667-678.

SANTOS, G., FERNÁNDEZ-OLMO, I. & IRABIEN, Á. 2015. Estimation of PM10-Bound As, Cd, Ni and Pb Levels by Means of Statistical Modelling: PLSR and ANN Approaches. *Water, Air, & Soil Pollution,* 226.

SANTOS, G., FERNANDEZ-OLMO, I., IRABIEN, A., LEDOUX, F. & COURCOT, D. 2016. Estimating airborne heavy metal concentrations in Dunkerque (northern France). *Arabian Journal of Geosciences,* 9.

SAVIN, I. & WINKER, P. 2013. Lasso–type and Heuristic Strategies in Model Selection and Forecasting. *Towards Advanced Data Analysis by Combining Soft Computing and Statistics.* Springer.

SAYEGH, A., TATE, J. E. & ROPKINS, K. 2016. Understanding how roadside concentrations of NOx are influenced by the background levels, traffic density, and meteorological conditions using Boosted Regression Trees. *Atmospheric Environment,* 127**,** 163-175.

SCHAUER, J. J., LOUGH, G. C., SHAFER, M. M., CHRISTENSEN, W. F., ARNDT, M. F., DEMINTER, J. T. & PARK, J. S. 2006. Characterization of metals emitted from motor vehicles. *Research report (Health Effects Institute)*, 1.

SCHMIDHUBER, J. 2015. Deep learning in neural networks: An overview. *Neural Networks,* 61**,** 85-117.

SEATON, A., GODDEN, D., MACNEE, W. & DONALDSON, K. 1995. Particulate air pollution and acute health effects. *The Lancet,* 345**,** 176-178.

SEN, P. K. 1968. Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association,* 63**,** 1379-1389.

SHARIFI, M. & GHAFOURIAN, T. 2014. Estimation of Biliary Excretion of Foreign Compounds Using Properties of Molecular Structure. *The AAPS journal,* 16**,** 65-78.

SHARMA, M., O'CONNELL, S., GARELLI, B., SATTAYATEWA, C., MOSCHANDREAS, D. & PAGILLA, K. 2012. Case study of odor and indoor air quality assessment in the dewatering building at the Stickney Water Reclamation Plant. *Water Science and Technology,* 65**,** 773-779.

SHARMA, N. C., K. K; CHALAPATI RAO, C. V 2005. vehecular pollution modelling using artificial neural network technique: A review. *journal of scientific and industrial reasearch,* 64**,** 637-647.

SHI, J. P., EVANS, D. E., KHAN, A. A. & HARRISON, R. M. 2001. Sources and concentration of nanoparticles (<10 nm diameter) in the urban atmosphere. *Atmospheric Environment,* 35**,** 1193-1202.

SIEGL, W. O., AKIN, A. C., ZINBO, M., COLEMAN, P. B., MARANO, R. S. & LEE, S. E. 1997. Vehicle interior air quality: Evaluation of odor-removal filters by single-gas testing. *Advances in Filtration and Separation Technology, Vol 11 1997*, 222-233.

SILVA, L. T. & MENDES, J. F. G. 2009. Atmospheric emissions of one pulp and paper mill. Contribution to the air quality of Viana do Castelo. *Proceedings of the 8th Wseas International Conference on System Science and Simulation in Engineering (Icossse '09)*, 21-26.

SIMONS, K., DE SMEDT, T., VAN NIEUWENHUYSE, A., BUYL, R. & COOMANS, D. 2016. Ensemble post-processing is a promising method to obtain flexible distributed lag models. *Air Quality, Atmosphere & Health*, 1-12.

SINGH, K. P., GUPTA, S., KUMAR, A. & SHUKLA, S. P. 2012. Linear and nonlinear modeling approaches for urban air quality prediction. *Science of the Total Environment,* 426**,** 244-255.

SINGH, K. P., GUPTA, S. & RAI, P. 2013. Identifying pollution sources and predicting urban air quality using ensemble learning methods. *Atmospheric Environment,* 80**,** 426-437.

SIWEK, K. & OSOWSKI, S. 2012. Improving the accuracy of prediction of PM10 pollution by the wavelet transformation and an ensemble of neural predictors. *Engineering Applications of Artificial Intelligence,* 25**,** 1246-1258.

SUÁREZ SÁNCHEZ, A., GARCÍA NIETO, P. J., RIESGO FERNÁNDEZ, P., DEL COZ DÍAZ, J. J. & IGLESIAS-RODRÍGUEZ, F. J. 2011. Application of an SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain). *Mathematical and Computer Modelling,* 54**,** 1453-1466.

SULEIMAN, A., TIGHT, M. R. & QUINN, A. D. 2016. Hybrid Neural Networks and Boosted Regression Tree Models for Predicting Roadside Particulate Matter. *Environmental Modeling & Assessment*, 1-20.

SUN, Z., TAO, Y., LI, S., FERGUSON, K. K., MEEKER, J. D., PARK, S. K., BATTERMAN, S. A. & MUKHERJEE, B. 2013. Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons. *Environmental Health,* 12**,** 85.

SZEGEDY, C., LIU, W., JIA, Y., SERMANET, P., REED, S., ANGUELOV, D., ERHAN, D., VANHOUCKE, V. & RABINOVICH, A. 2014. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*.

TAI, A. P. K., MICKLEY, L. J. & JACOB, D. J. 2010. Correlations between fine particulate matter (PM2.5) and meteorological variables in the United States: Implications for the sensitivity of PM2.5 to climate change. *Atmospheric Environment,* 44**,** 3976-3984.

TAN, B., ZHANG, J. & WANG, L. 2011. Semi-supervised Elastic net for pedestrian counting. *Pattern Recognition,* 44**,** 2297-2304.

TASPINAR, F. 2015. Improving artificial neural network model predictions of daily average PM10 concentrations by applying principle component analysis and implementing seasonal models. *J Air Waste Manag Assoc,* 65**,** 800-9.

TASPINAR, F. & BOZKURT, Z. 2014. APPLICATION OF ARTIFICIAL NEURAL NETWORKS AND REGRESSION MODELS IN THE PREDICTION OF DAILY MAXIMUM PM10 CONCENTRATION IN DUZCE, TUKEY. *Fresenius Environmental Bulletin,* 23**,** 2450-2459.

TATE, J., ROPKINS, K., GOODMAN, P., OATES, C., CHEN, H., BELL, M., BALOGUN, A., SMALLEY, R. & AS, T. The influence of traffic congestion, synoptic and in-street winds on NO2 concentrations around a congested intersection: a measurement study. The 7th International Conference on Air Quality - Science and Application (Air Quality 2009), 2009.

TAYLOR, K. E. 2001. Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres (1984–2012),* 106**,** 7183-7192.

TETLEY, T. D. 2007. Health effects of nanomaterials. *Biochem. Soc. Trans.,* 35**,** 527-531.

TFL 2014. Transport Emissions Roadmap (TERM) *In:* LONDON, M. O. (ed.). Transport for London.

TFL. 2016. *Ultra Low Emission Zone* [Online]. TRANSPORT FOR LONDON. Available: https://tfl.gov.uk/modes/driving/ultra-low-emission-zone?cid=ultra-low-emission-zone [Accessed 02/08/2016 2016].

THEIL, H. 1992. A rank-invariant method of linear and polynomial regression analysis. *Henri Theil's Contributions to Economics and Econometrics.* Springer.

THORPE, A. & HARRISON, R. M. 2008. Sources and properties of non-exhaust particulate matter from road traffic: A review. *Science of the Total Environment,* 400**,** 270-282.

THUNIS, P., GEORGIEVA, E. & PEDERZOLI, A. 2012a. A tool to evaluate air quality model performances in regulatory applications. *Environmental Modelling & Software,* 38**,** 220-230.

THUNIS, P., GEORGIEVA, E. & PEDERZOLI, A. 2012b. A tool to evaluate air quality model performances in regulatory applications. *Environmental Modelling & Software,* 38**,** 220-230.

THUNIS, P., PEDERZOLI, A. & PERNIGOTTI, D. 2012c. Performance criteria to evaluate air quality modeling applications. *Atmospheric Environment,* 59**,** 476-482.

TIJANI, K., PLOIX, S., HAAS, B., DUGDALE, J. & NGO, Q. D. 2016. Dynamic Bayesian Networks to simulate occupant behaviours in office buildings related to indoor air quality. *arXiv preprint arXiv:1605.05966*.

TOMLIN, A. S., SMALLEY, R. J., TATE, J. E., BARLOW, J. F., BELCHER, S. E., ARNOLD, S. J., DOBRE, A. & ROBINS, A. 2009. A field study of factors influencing the concentrations of a traffic-related pollutant in the vicinity of a complex urban junction. *Atmospheric Environment,* 43**,** 5027-5037.

TORIJA, A. J. & RUIZ, D. P. 2015. A general procedure to generate models for urban environmental noise pollution using feature selection and machine learning methods. *Science of the Total Environment,* 505**,** 680-693.

TUMER, K. & GHOSH, J. 1996. Analysis of decision boundaries in linearly combined neural classifiers. *Pattern Recognition,* 29**,** 341-348.

UK-AIR. 2013. *Department fro Environment Food and Rural Affairs Data Archive* [Online]. Available: http://uk-air.defra.gov.uk/data/maryleboneroad [Accessed 03/04 2013].

UL-SAUFIE, A. Z., YAHAYA, A. S., RAMLI, N. A., ROSAIDA, N. & HAMID, H. A. 2013. Future daily PM 10 concentrations prediction by combining regression models and feedforward

backpropagation models with principle component analysis (PCA). *Atmospheric Environment,* 77**,** 621-630.
USEPA 2011. The Benefits and Costs of the clean air act fullreport. *Final Report*

U.S. Environmental Protection Agency

Office of Air and Radiation.
VALLERO, D. 2014. Modeling Applications. 683-753.
VAN BUUREN, S. 2007. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research,* 16**,** 219-242.
VAPNIK, V. 2000. *The nature of statistical learning theory*, springer.
VERGUN, S., DESHPANDE, A. S., MEIER, T. B., SONG, J., TUDORASCU, D. L., NAIR, V. A., SINGH, V., BISWAL, B. B., MEYERAND, M. E., BIRN, R. M. & PRABHAKARAN, V. 2013. Characterizing functional connectivity differences in aging adults using machine learning on resting state fMRI data. *Frontiers in Computational Neuroscience,* 7.
VIDYASAGAR, M. 2015. Identifying Predictive Features in Drug Response Using Machine Learning: Opportunities and Challenges. *Annual Review of Pharmacology and Toxicology, Vol 55,* 55**,** 15-34.
VLACHOGIANNI, A., KASSOMENOS, P., KARPPINEN, A., KARAKITSIOS, S. & KUKKONEN, J. 2011. Evaluation of a multiple regression model for the forecasting of the concentrations of NOx and PM10 in Athens and Helsinki. *Science of the Total Environment,* 409**,** 1559-1571.
VONG, C. M., IP, W. F., WONG, P. K. & CHIU, C. C. 2014. Predicting minority class for suspended particulate matters level by extreme learning machine. *Neurocomputing,* 128**,** 136-144.
WALDMANN, P., MÉSZÁROS, G., GREDLER, B., FÜRST, C. & SÖLKNER, J. 2013. Evaluation of the lasso and the elastic net in genome-wide association studies. *Frontiers in Genetics,* 4.
WANG, H. F., FANG, J. & GAO, C. 2009. Research on the Assessment for Air Environment Quality Based on Support Vector Machine. *Ccdc 2009: 21st Chinese Control and Decision Conference, Vols 1-6, Proceedings***,** 4753-4757.
WANG, Z., LU, F., HE, H.-D., LU, Q.-C., WANG, D. & PENG, Z.-R. 2015a. Fine-scale estimation of carbon monoxide and fine particulate matter concentrations in proximity to a road intersection by using wavelet neural network with genetic algorithm. *Atmospheric Environment,* 104**,** 264-272.
WANG, Z., LU, F., LU, Q.-C., WANG, D. & PENG, Z.-R. 2015b. Fine-scale estimation of carbon monoxide and fine particulate matter concentrations in proximity to a road intersection by using wavelet neural network with genetic algorithm. *Atmospheric Environment,* 104**,** 264-272.
WHITTINGHAM, M. J., STEPHENS, P. A., BRADBURY, R. B. & FRECKLETON, R. P. 2006. Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology,* 75**,** 1182-1189.
WHO 2005. Particulate matter air pollution: how it harms health. *http://www.euro.who.int/document/mediacentre/fs0405e.pdf*.
WHO, W. H. O. 2014. Burden of disease from household air pollution for 2012. *World Health Organization: Geneva, Switzerland*.
WHYATT, J. D., METCALFE, S. E., NICHOLSON, J., DERWENT, R. G., PAGE, T. & STEDMAN, J. R. 2007. Regional scale modelling of particulate matter in the UK, source attribution and an assessment of uncertainties. *Atmospheric Environment,* 41**,** 3315-3327.
WICKHAM, H. & WICKHAM, M. H. 2013. Package 'plyr'.
WILAMOWSKI, B. M. & HAO, Y. 2010. Neural Network Learning Without Backpropagation. *Neural Networks, IEEE Transactions on,* 21**,** 1793-1803.
WILKS, D. S. 2011. Chapter 7 - Statistical Forecasting. *In:* DANIEL, S. W. (ed.) *International Geophysics.* Academic Press.

WILLIAMS, D. R. G. H. R. & HINTON, G. 1986. Learning representations by back-propagating errors. *Nature*, 323,533-538.
WILLIAMS, M., BARROWCLIFFE, R., LAXEN, D. & MONKS, P. 2011. Review of Air Quality modelling in Defra. *A report by the Air Quality Modeling Review Steering Group*.
WILLMOTT, C. J. 1981. On the validation of models *Physical geography,* 2, 184 - 194.
WILLMOTT, C. J. 1982. Some comments on the evaluation of models. *Bulletin American Meteorological Society*.
WILLMOTT, C. J. & WICKS, D. E. 1980. An empirical method for the spatial interpolation of monthly precipitation within California. *Physical Geography,* 1, 59-73.
WORLD HEALTH ORGANIZATION, W. 2014a. Air quality deteriorating in many of the world's cities. *WHO, Geneva, Switzerland*.
WORLD HEALTH ORGANIZATION, W. 2014b. *Mortality from ambient air pollution for 2012* [Online]. Geneva. Available: http://www.who.int/phe/health_topics/outdoorair/databases/AAP_BoD_results_March2014.pdf [Accessed 14/03/2016 2016].
XIA, T., NITSCHKE, M., ZHANG, Y., SHAH, P., CRABB, S. & HANSEN, A. 2015a. Traffic-related air pollution and health co-benefits of alternative transport in Adelaide, South Australia. *Environment International,* 74, 281-290.
XIA, X., ZHAO, W., RUI, X., WANG, Y., BAI, X., YIN, W. & DON, J. A comprehensive evaluation of air pollution prediction improvement by a machine learning method. Service Operations And Logistics, And Informatics (SOLI), 2015 IEEE International Conference on, 15-17 Nov. 2015 2015b. 176-181.
YANG, L., ZHOU, X., WANG, Z., ZHOU, Y., CHENG, S., XU, P., GAO, X., NIE, W., WANG, X. & WANG, W. 2012. Airborne fine particulate pollution in Jinan, China: Concentrations, chemical compositions and influence on visibility impairment. *Atmospheric Environment,* 55, 506-514.
YANG, X. J., ZHANG, Y. P. & ZHAO, S. 2008. Multi-granular Ensemble Learning Method for predictions of air quality. *2008 Proceedings of Information Technology and Environmental System Sciences: Itess 2008, Vol 4*, 316-319.
YI, J. & PRYBUTOK, V. R. 1996. A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area. *Environmental Pollution,* 92, 349-357.
YIM, S. H. & BARRETT, S. R. 2012. Public health impacts of combustion emissions in the United Kingdom. *Environ Sci Technol,* 46, 4291-6.
YIM, S. H. L., STETTLER, M. E. J. & BARRETT, S. R. H. 2013. Air quality and public health impacts of UK airports. Part II: Impacts and policy assessment. *Atmospheric Environment,* 67, 184-192.
ZHANG, D. & WANG, Y. 2009. Rough neural network based on bottom-up fuzzy rough data analysis. *Neural processing letters,* 30, 187-211.
ZISSIS SAMARASA, L. N., NEVILLE THOMPSONB, DIANE & HALLB, R. W., PAUL BOULTERD 2005. Characterisation of Exhaust Particulate Emissions from Road Vehicles. *http://vergina.eng.auth.gr/mech/lat/particulates/private/index.htm*.
ZOU, H. & HASTIE, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 67, 301-320.

Appendix A PM Sampling Methods

Table A.1 **Summary of advantages and disadvantages of principal PM sampling methods (AQEG, 2005)**

Technique	Advantages	Disadvantages	Estimated Precision[a]
Filter-based gravimetric samplers	The reference method for PM_{10} specified in the EU First Daughter Directive.	High operating costs. Time resolution of the measurement is limited to 24-h. Reporting requirements of the EU First Daughter Directive cannot be met and results can only be provided some days after the sample was collected.	$\pm 2 \mu g\ m^{-3}$
TEOM analysers	Provide real-time data with short time resolution (<1 h) that can be used for public Information. Improved precision compared to the reference method	Preheated air stream causes a greater loss of semi-volatiles compared to the reference method. High capital cost.	$\pm$ 0.5 μgm^{-3}
ß-attenuation analysers	Provide real-time data with short time resolution (<1 h) that can be used for public information.	If a heated inlet is used some semi-volatile material may be lost. Unheated samplers may suffer from interference due to the presence of water Analyser contains a radioactive source	$\pm 3\ \mu gm^{-3}$, but depends on analyser type
Optical analysers	Portable and often battery operated. Can measure several size fractions simultaneously.	Depends on assumptions about particle characteristics, which may vary from place to place and time to time.	Depends on analyser type

[a] The precision is for PM_{10} mass determined using 24-hours averaging period.

Table A.1 *continues*

Technique	Advantages	Disadvantages	Estimated Precision[a]
Black smoke	Simple, robust, inexpensive and easy to maintain. Long time series of existing data.	Measures an index rather than a gravimetric concentration. Calibration factor not appropriate to the current mix of pollution sources. Time resolution of the measurement is limited to 24-h.	±2 μgm^{-3}, may be higher at current typical concentrations
Personal samplers	Portable samplers that can easily be deployed in the field and used to determine personal exposure to particulate concentrations.	As above, depending on measurement method used All personal exposure analysis is very labour-intensive.	According to technique employed, as above

[a] The precision is for PM_{10} mass determined using 24-hours averaging period.

Appendix B National Air Quality Objectives and European Directive Limit and Target Value

Table B.1 National air quality objectives and European Directive limit and target values for the protection of human health (DEFRA, 2007)

Pollutant	Applies	Objective	Concentration measured as	Date to be achieved by and maintained thereafter	European obligations	Date to be achieved by and maintained thereafter	New or existing
Particles (PM_{10})	UK	50µg.m^{-3} not to be exceeded more than 35 times a year	24-hour mean	31 December 2004	50µg.m^{-3} not to be exceeded more than 35 times a year	1 January 2005	Retain existing
	UK	40µg.m-3	annual mean	31 December 2004	40µg.m^{-3}	1 January 2005	
	Indicative 2010 objectives for PM_{10} (from the 2000 Strategy and 2003 Addendum) have been replaced by an exposure reduction approach for $PM_{2.5}$ (except in Scotland – see below)						
	Scotland	50µg.m-3 not to be exceeded more than seven times a year	24-hour mean	31 December 2010			Retain existing
	Scotland	18µg.m^{-3}	annual mean	31 December 2010			
Particles ($PM_{2.5}$) Exposure Reduction	UK (except Scotland)	25µg.m^{-3}	annual mean	2020	Target value 25µg.m^{-3}	2010	New (European obligations still under negotiation)
	Scotland	12µg.m-3		2020	Limit value 25µg.m-3	2015	
	UK urban areas	Target of 15% reduction in concentrations at urban background		Between 2010 and 2020	Target of 20% reduction in concentrations at urban background3	Between 2010 and 2020	

Table B.1 *continued*

Pollutant	Applies	Objective	Concentration measured as	Date to be achieved by and maintained thereafter	European obligations	Date to be achieved by and maintained thereafter	New or existing
Nitrogen dioxide	UK	200µg.m^{-3} not to be exceeded more than 18 times a year	1-hour mean	31 December 2005	200µg.m^{-3} not to be exceeded more than 18 times a year	1 January 2010	Retain existing
	UK	40µg.m^{-3}	annual mean	31 December 2005	40µg.m^{-3}	1 January 2010	
Ozone	UK	100µg.m^{-3} not to be exceeded more than 10 times a year	8 hour mean	31 December 2005	Target of 120µg.m^{-3} not to be exceeded more than 25 times a year averaged over 3 years	31 December 2010	Retain existing
Sulphur dioxide	UK	266µg.m^{-3} not to be exceeded more than 35 times a year	15 minute mean	31 December 2005			Retain existing
	UK	350µg.m^{-3} not to be exceeded more than 24 times a year	1-hour mean	31 December 2004	350µg.m^{-3} not to be exceeded more than 24 times a year	1 January 2005	
	UK	125µg.m^{-3} not to be exceeded more than 3 times a year	24-hour mean	31 December 2004	125µg.m-3 not to be exceeded more than 3 times a year	1 January 2005	
Polycyclic aromatic hydrocarbons	UK	0.25ng.m^{-3} B[a]P as	annual average	31 December 2010	Target of 1ng.m^{-3}	December 2012	Retain existing

Table B.1 *continued*

Pollutant	Applies	Objective	Concentration measured as	Date to be achieved by and maintained thereafter	European obligations	Date to be achieved by and maintained thereafter	New or existing
Benzene	UK	16.25µg.m^{-3}	running annual mean	31 December 2003			Retain existing
	England and Wales	5µg.m^{-3}	annual average	31 December 2010	5µg.m^{-3}	1 January 2010	
	Scotland, Northern Ireland	3.25µg.m^{-3}	running annual mean	31 December 2010			
1,3- butadiene	UK	2.25µg.m^{-3}	running annual mean	31 December 2003			Retain existing
Carbon monoxide	UK	10mg.m^{-3}	maximum daily running 8 hour mean/in Scotland as running 8 hour mean	31 December 2003	10mg.m^{-3}	1 January 2005	Retain existing
Lead	UK	0.5µg.m^{-3}	annual mean	31 December 2004	0.5µg.m^{-3}	1 January 2005	Retain existing
		0.25µg.m^{-3}	annual mean	31 December 2008			

Appendix C Missing Data Imputation

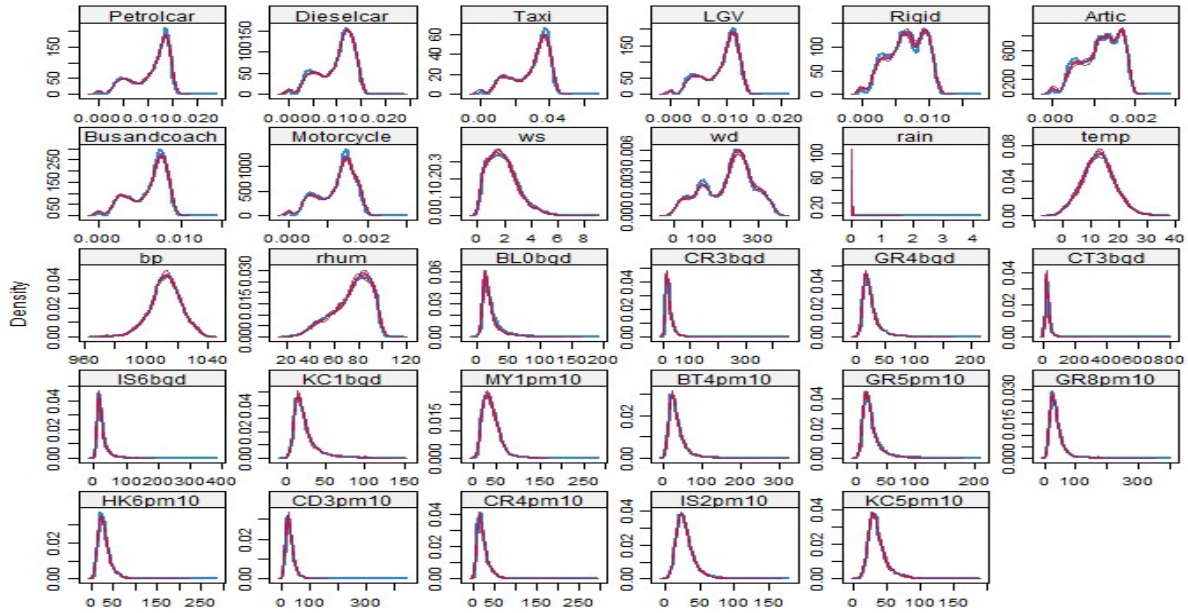


Figure C.1 Density plots of the imputed data with 10% missing

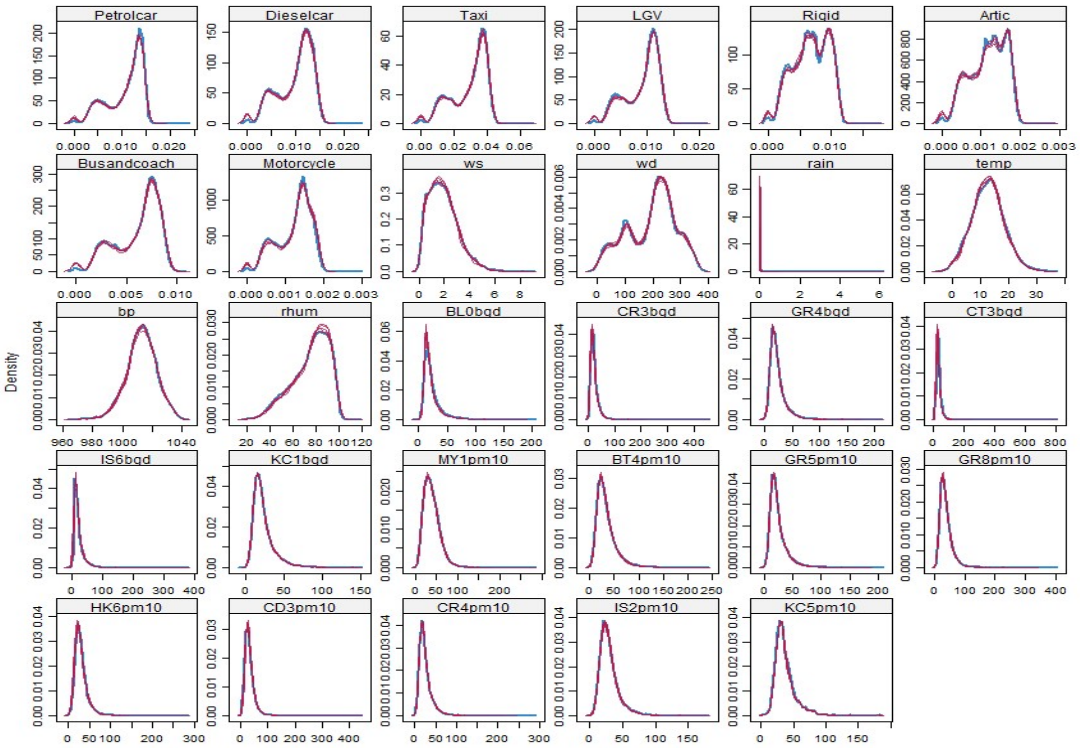


Figure C.2 Density plots of the imputed data with 20% missing

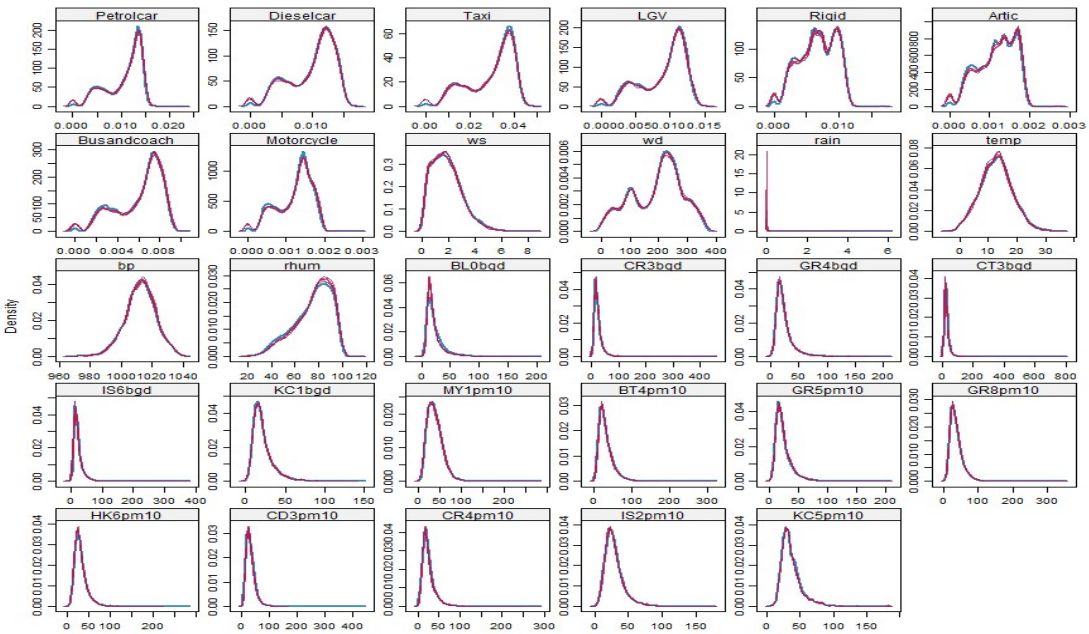


Figure C.3 Density plots of the imputed data with 30% missing

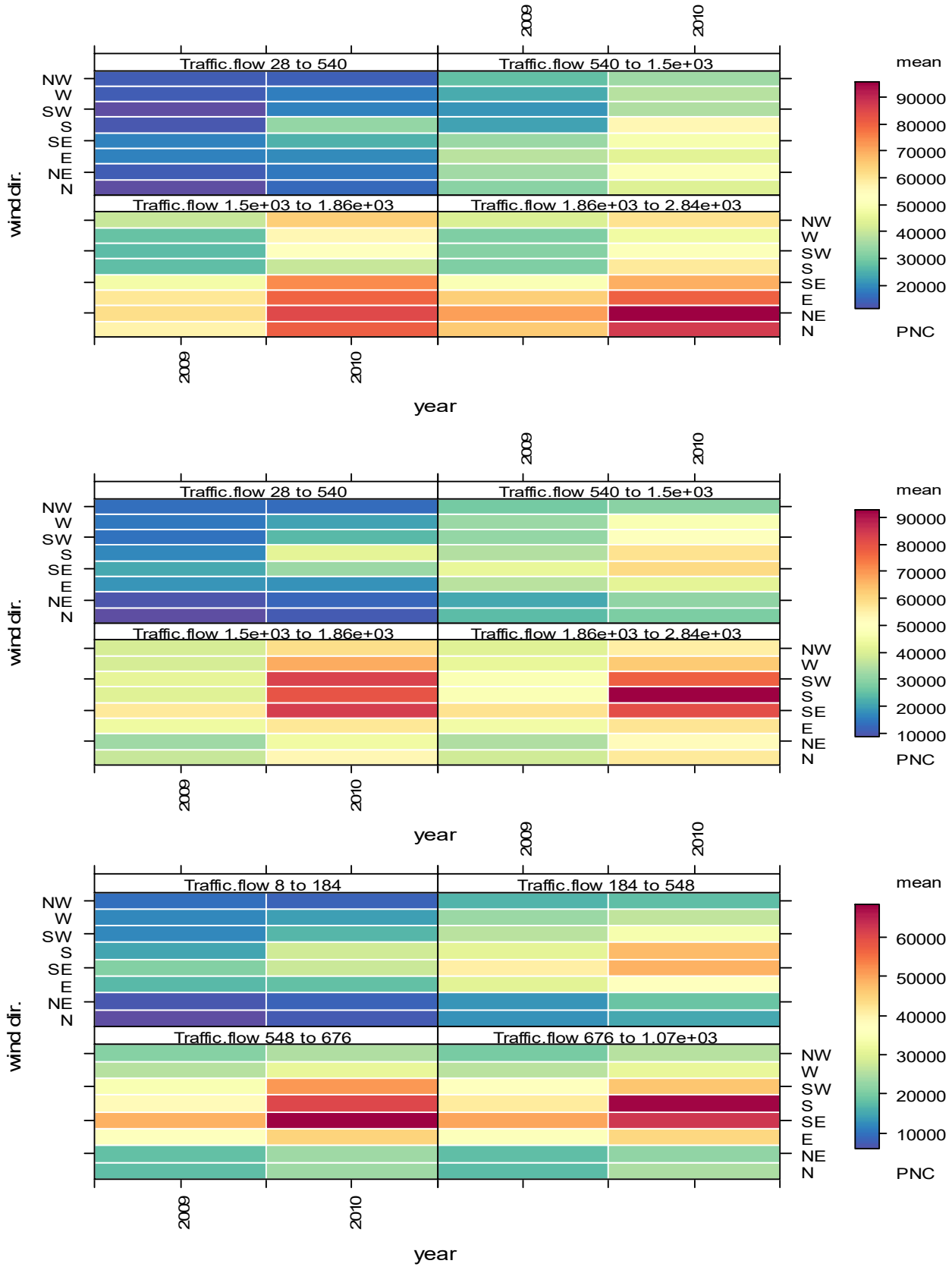


Figure C.4 Trend level plot showing the relationship between traffic volume, particle concentrations and wind directions at the Instrumented Junction

Appendix D Performance of Statistical Methods

Table D.1 Training performance of PCR Models

Pollutant	Model	Data type	Prediction	Number of components	RMSE	R-squared
PM_{10}	PCR	Roadside Data	Roadside prediction	20	10.26	0.77
PM_{10}	PCR	Roadside Data	Increment prediction	19	10.37	0.75
PM_{10}	PCR	Increment data	Roadside prediction	20	10.46	0.77
PM_{10}	PCR	Increment data	Increment prediction	20	10.57	0.74
$PM_{2.5}$	PCR	Roadside data	Roadside prediction	19	5.04	0.87
$PM_{2.5}$	PCR	Roadside data	Increment prediction	20	5	0.73
$PM_{2.5}$	PCR	Increment data	Roadside prediction	20	4.98	0.87
$PM_{2.5}$	PCR	Increment data	Increment prediction	20	5.11	0.73
PNC	PCR	Roadside data	Roadside prediction	19	11283	0.83
PNC	PCR	Roadside data	Increment prediction	16	10857	0.80
PNC	PCR	Increment data	Roadside prediction	16	11096	0.84
PNC	PCR	Increment data	Increment prediction	16	11008	0.80

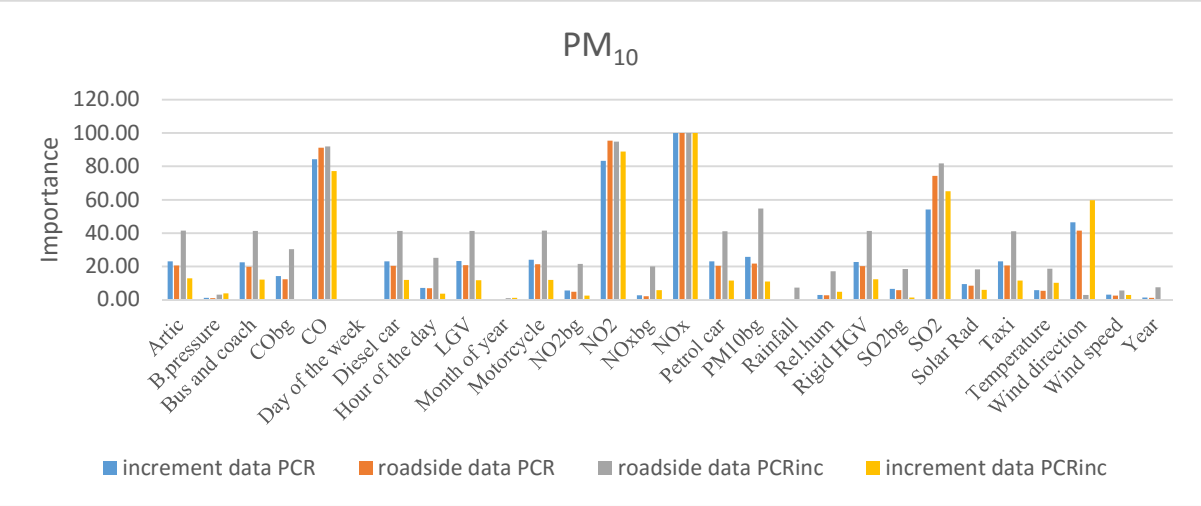


Figure D.1 Predictor variable importance for PM_{10}

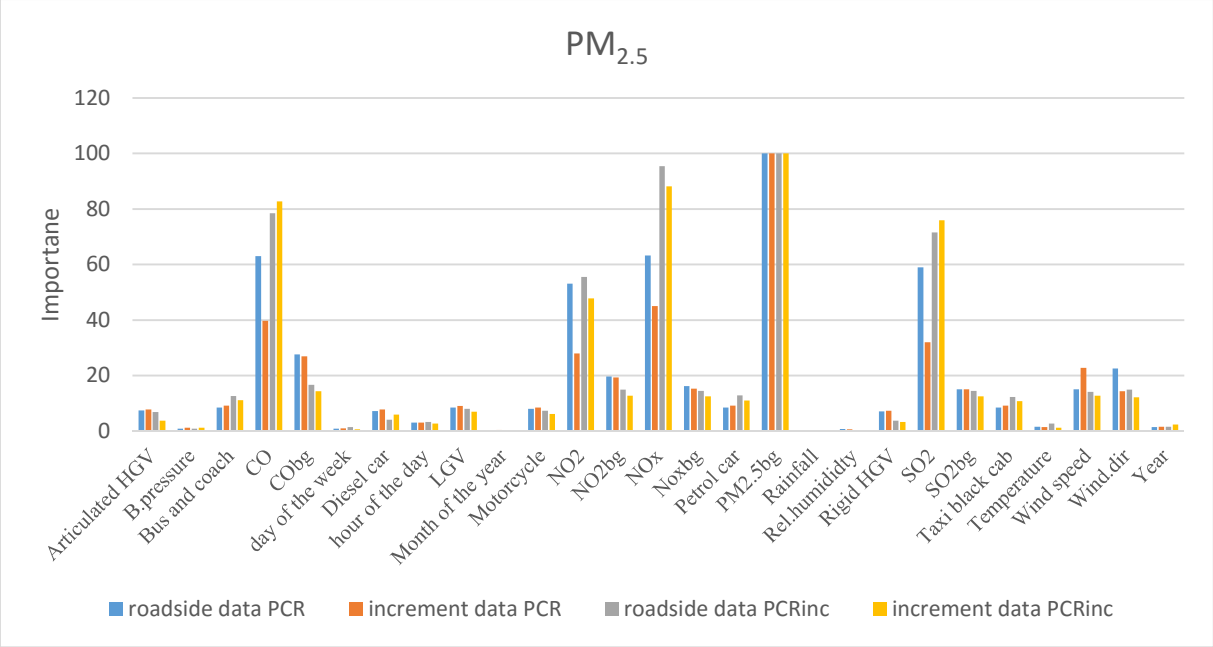


Figure D.2 Predictor variable importance for $PM_{2.5}$ models

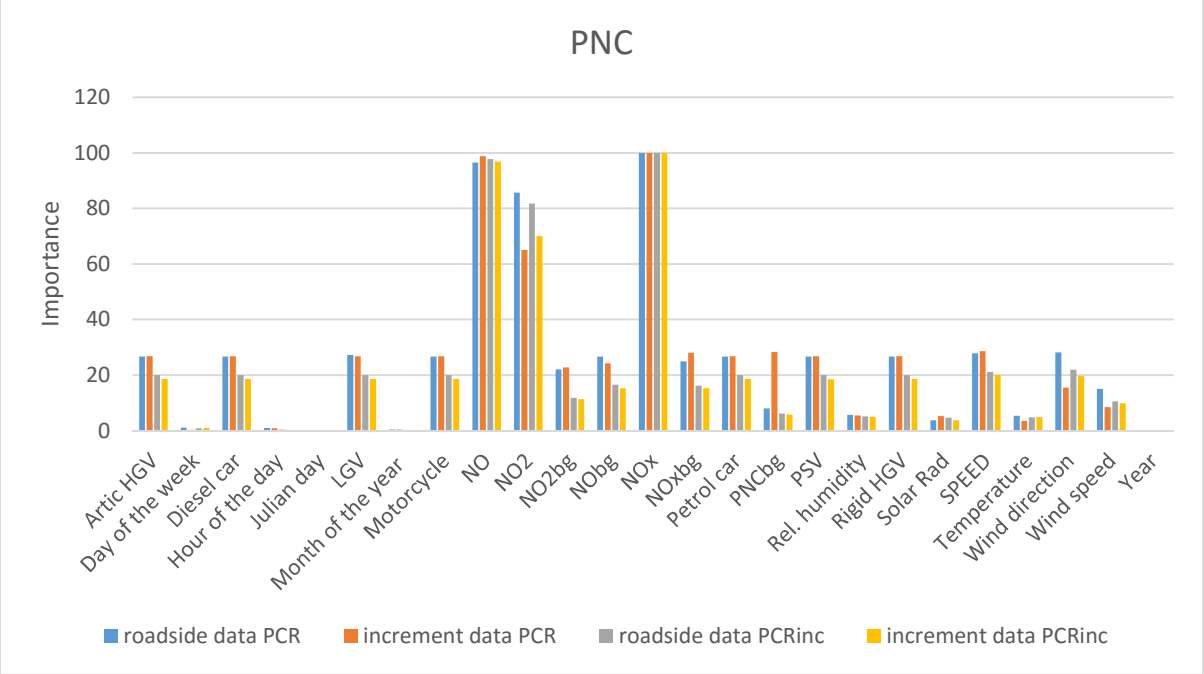


Figure D.3 Predictor variable importance for PNC models

Table D.2 The test performance for PCR models

Pollutant	model	Data type	Prediction	FAC2	MB	MGE	NMB	NMGE	RMSE	r	COE	IOA
PM_{10}	PCR	roadside data	Roadside prediction	0.99	-0.23	6.89	-0.01	0.16	10.98	0.87	0.60	0.80
PM_{10}	PCR	roadside data	Increment prediction	0.75	-0.16	6.88	-0.01	0.33	10.95	0.86	0.59	0.79
PM_{10}	PCR	increment data	Roadside prediction	0.99	-0.06	6.84	0.00	0.16	10.05	0.88	0.60	0.80
PM_{10}	PCR	increment data	Increment prediction	0.75	0.01	6.83	0.00	0.32	10.03	0.88	0.59	0.79
$PM_{2.5}$	PCR	roadside data	Roadside prediction	0.97	0.05	3.59	0.00	0.15	5.13	0.93	0.65	0.83
$PM_{2.5}$	PCR	roadside data	Increment prediction	0.62	0.03	3.58	0.01	0.59	5.11	0.85	0.54	0.77
$PM_{2.5}$	PCR	increment data	Roadside prediction	0.97	-0.05	3.57	0.00	0.15	5.31	0.93	0.66	0.83
$PM_{2.5}$	PCR	increment data	Increment prediction	0.61	-0.04	3.56	-0.01	0.58	5.29	0.85	0.55	0.77
PNC	PCR	roadside data	Roadside prediction	0.95	181	5760	0.01	0.18	10557	0.92	0.70	0.85
PNC	PCR	roadside data	Increment prediction	0.77	169	5694	0.01	0.28	10496	0.90	0.67	0.83
PNC	PCR	increment data	Roadside prediction	0.95	-216	5916	-0.01	0.18	11088	0.92	0.70	0.85
PNC	PCR	increment data	Increment prediction	0.77	-178	5913	-0.01	0.28	11075	0.90	0.66	0.83

Table D.3 Training performance of PLSR Models

Pollutant	Model	Data type	Prediction	Number of components	RMSE	R-squared
PM_{10}	PLSR	roadside data	Roadside prediction	20	9.93	0.79
PM_{10}	PLSR	roadside data	Increment prediction	20	10.1	0.77
PM_{10}	PLSR	increment data	Roadside prediction	20	10.2	0.78
PM_{10}	PLSR	increment data	Increment prediction	20	10.28	0.76
$PM_{2.5}$	PLSR	roadside data	Roadside prediction	18	4.96	0.87
$PM_{2.5}$	PLSR	roadside data	Increment prediction	18	4.93	0.74
$PM_{2.5}$	PLSR	increment data	Roadside prediction	20	4.91	0.87
$PM_{2.5}$	PLSR	increment data	Increment prediction	20	5.05	0.73
PNC	PLSR	roadside data	Roadside prediction	16	11284	0.83
PNC	PLSR	roadside data	Increment prediction	16	10802	0.81
PNC	PLSR	increment data	Roadside prediction	16	11192	0.83
PNC	PLSR	increment data	Increment prediction	16	11051	0.8

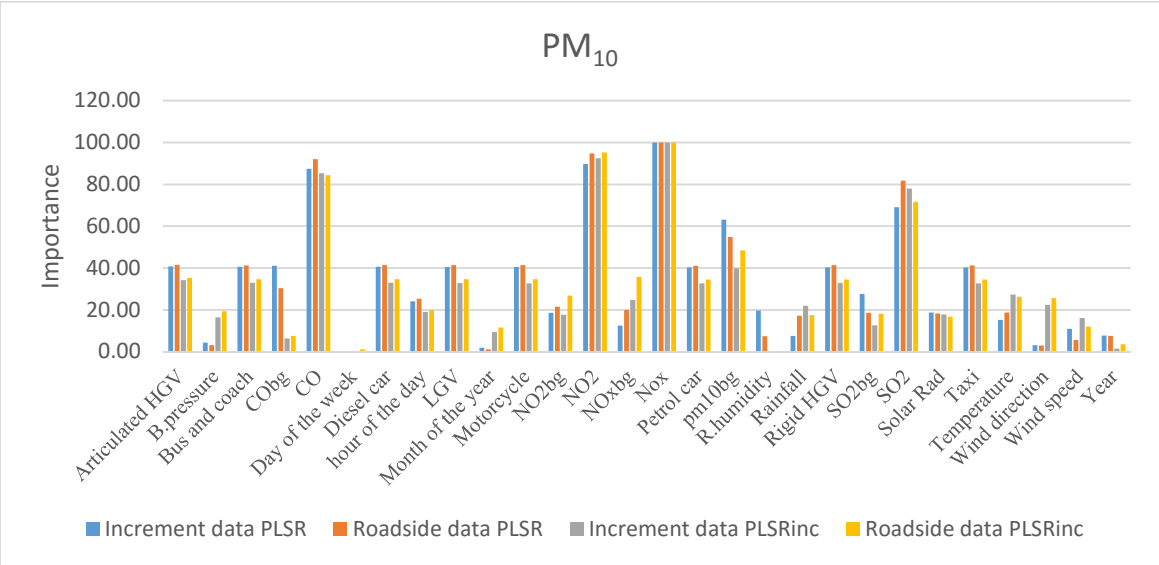


Figure D.4 Predictor variable importance for PM_{10}

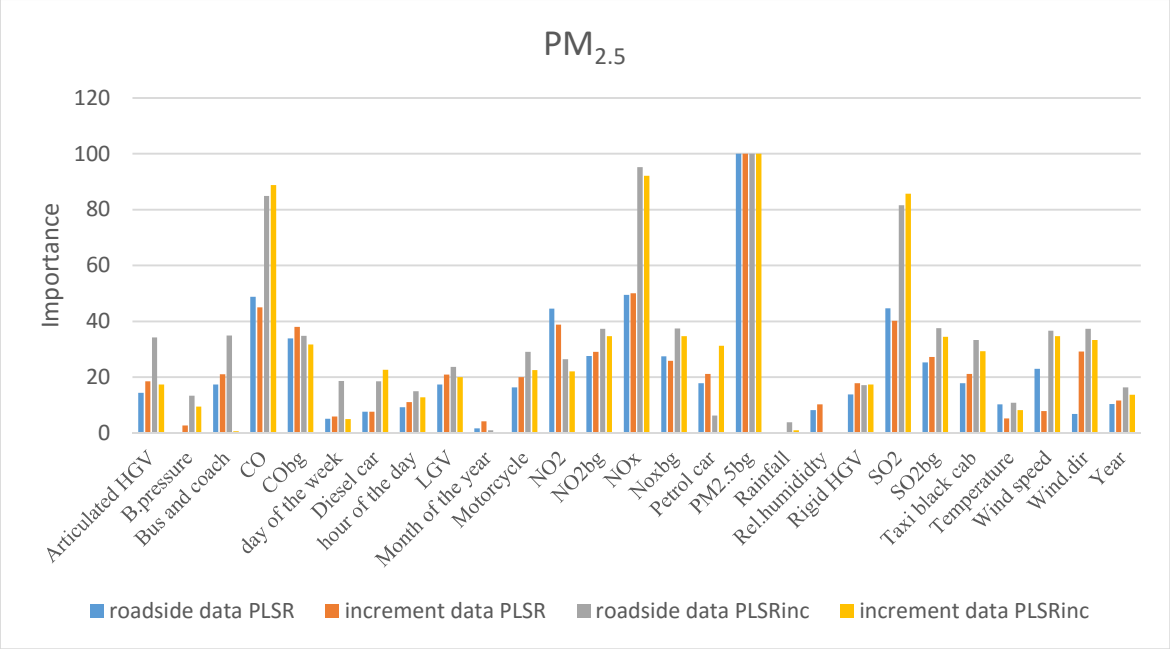


Figure D.5 Predictor variable importance for $PM_{2.5}$

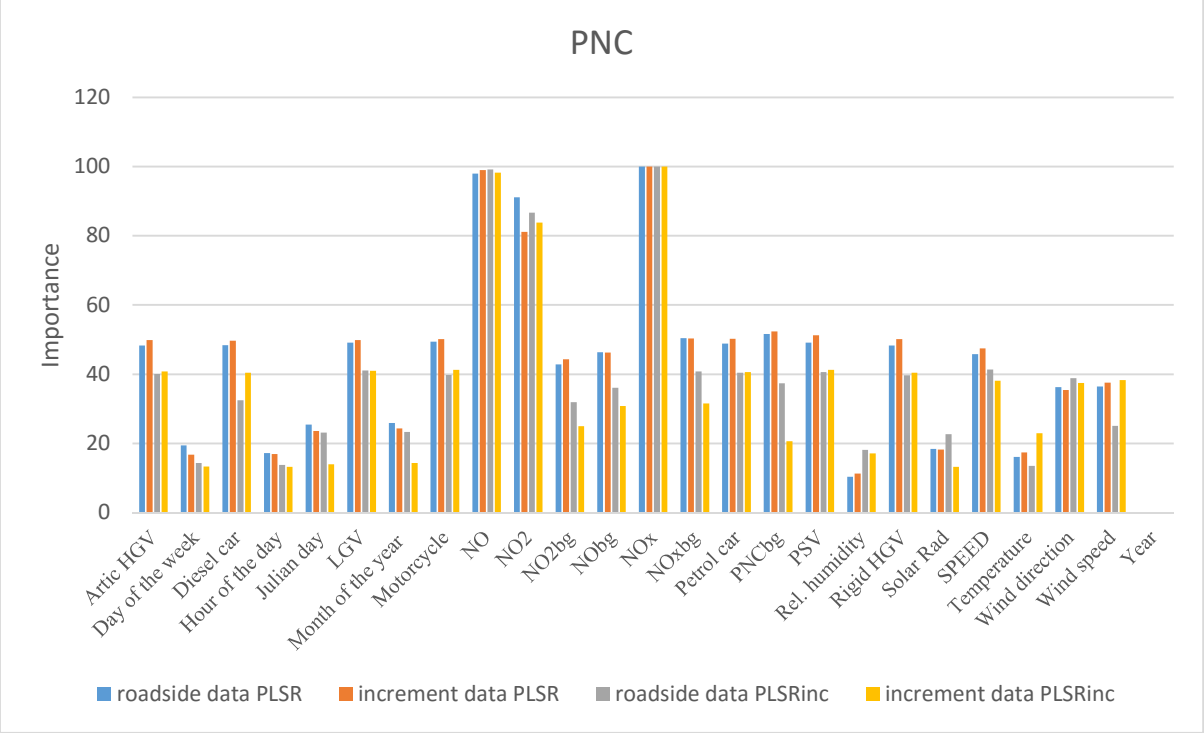


Figure D.6 Predictor variable importance for PNC

Table D.4 Testing performance for PLSR models

Pollutanta	model	Data type	Prediction	FAC2	MB	MGE	NMB	NMGE	RMSE	r	COE	IOA
PM_{10}	PLSR	roadside data	Roadside prediction	0.99	-0.21	6.69	0.00	0.15	10.76	0.87	0.61	0.81
PM_{10}	PLSR	Roadside data	Increment prediction	0.76	-0.15	6.68	-0.01	0.32	10.75	0.86	0.60	0.80
PM_{10}	PLSR	increment data	Roadside prediction	0.99	-0.07	6.62	0.00	0.15	9.77	0.89	0.61	0.81
PM_{10}	PLSR	increment data	Increment prediction	0.76	0.00	6.62	0.00	0.31	9.75	0.88	0.60	0.80
$PM_{2.5}$	PLSR	roadside data	Roadside prediction	0.97	0.04	3.50	0.00	0.15	5.05	0.93	0.66	0.83
$PM_{2.5}$	PLSR	roadside data	Increment prediction	0.63	0.03	3.49	0.00	0.57	5.04	0.86	0.55	0.78
$PM_{2.5}$	PLSR	increment data	Roadside prediction	0.97	-0.03	3.49	0.00	0.15	5.24	0.93	0.67	0.83
$PM_{2.5}$	PLSR	increment data	Increment prediction	0.62	-0.03	3.49	0.00	0.57	5.23	0.85	0.56	0.78
PNC	PLSR	Roadside data	Roadside prediction	0.95	183	5761	0.01	0.18	10556	0.92	0.70	0.85
PNC	PLSR	roadside data	Increment prediction	0.77	172	5692	0.01	0.28	10491	0.90	0.67	0.83
PNC	PLSR	increment data	Roadside prediction	0.95	-218	5915	-0.01	0.18	11085	0.92	0.70	0.85
PNC	PLSR	increment data	Increment prediction	0.78	-179	5912	-0.01	0.28	11071	0.90	0.66	0.83

Table D.5 The training performance of the elastic-net models

Pollutants	Model	Data type	Prediction	alpha	lambda	RMSE	Rsquared
PM_{10}	Elasticnet regression	roadside data	Roadside prediction	0.25	0	9.93	0.79
PM_{10}	Elasticnet regression	roadside data	Increment prediction	0.95	0	10.09	0.77
PM_{10}	Elasticnet regression	increment data	Roadside prediction	0.2	0.01	10.15	0.78
PM_{10}	Elasticnet regression	increment data	Increment prediction	0.15	0.01	10.27	0.76
$PM_{2.5}$	Elasticnet regression	roadside data	Roadside prediction	0.55	0	4.96	0.87
$PM_{2.5}$	Elasticnet regression	roadside data	Increment prediction	0.15	0.01	4.94	0.74
$PM_{2.5}$	Elasticnet regression	increment data	Roadside prediction	0.45	0	4.91	0.87
$PM_{2.5}$	Elasticnet regression	increment data	Increment prediction	0.85	0	5.06	0.73
PNC	Elasticnet regression	roadside data	Roadside prediction	0.15	0.5	11343	0.83
PNC	Elasticnet regression	roadside data	Increment prediction	0.3	0.5	10827	0.80
PNC	Elasticnet regression	increment data	Roadside prediction	0.4	0.5	11214	0.83
PNC	Elasticnet regression	increment data	Increment prediction	0.1	0.5	11048	0.80

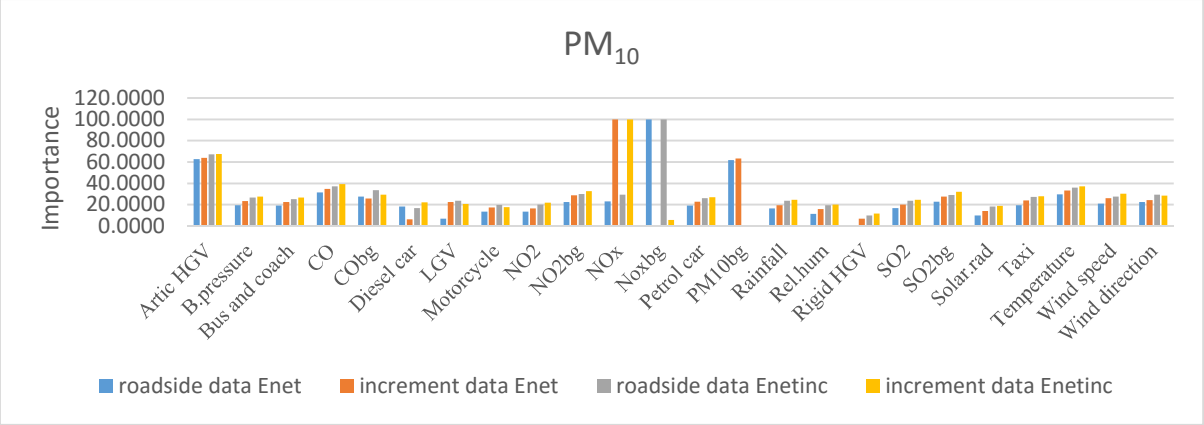


Figure D.7 Predictor variable importance for PM_{10}

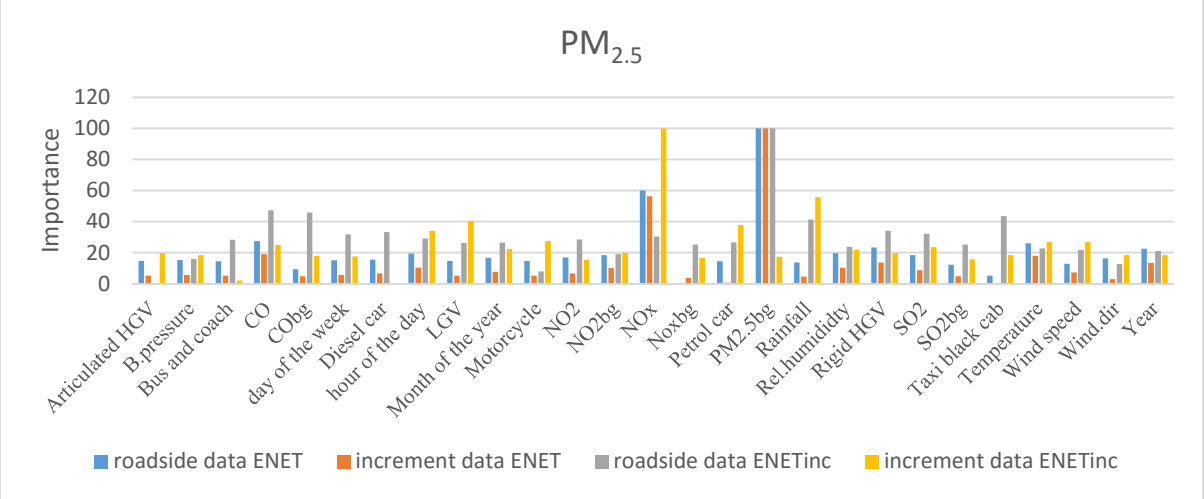


Figure D.8 Predictor variable importance for PM_{10}

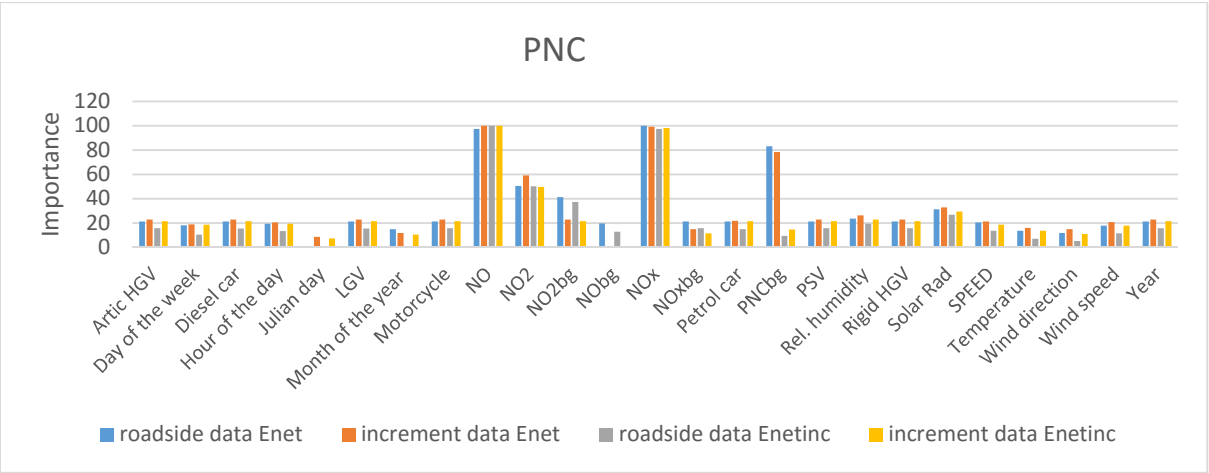


Figure D.9 Predictor variable importance for PNC

Table D.6 Training performance of Stepwise regression models

	Model	Data type	Prediction	RMSE	R-squared
PM_{10}	Stepwise regression	roadside data	Roadside prediction	9.94	0.78
PM_{10}	Stepwise regression	roadside data	Increment prediction	10.11	0.77
PM_{10}	Stepwise regression	increment data	Roadside prediction	10.17	0.78
PM_{10}	Stepwise regression	increment data	Increment prediction	10.29	0.76
$PM_{2.5}$	Stepwise regression	roadside data	Roadside prediction	4.97	0.87
$PM_{2.5}$	Stepwise regression	roadside data	Increment prediction	4.94	0.74
$PM_{2.5}$	Stepwise regression	increment data	Roadside prediction	4.91	0.87
$PM_{2.5}$	Stepwise regression	increment data	Increment prediction	5.05	0.73
PNC	Stepwise regression	roadside data	Roadside prediction	11222	0.84
PNC	Stepwise regression	roadside data	Increment prediction	10822	0.80
PNC	Stepwise regression	increment data	Roadside prediction	11208	0.83
PNC	Stepwise regression	increment data	Increment prediction	11068	0.80

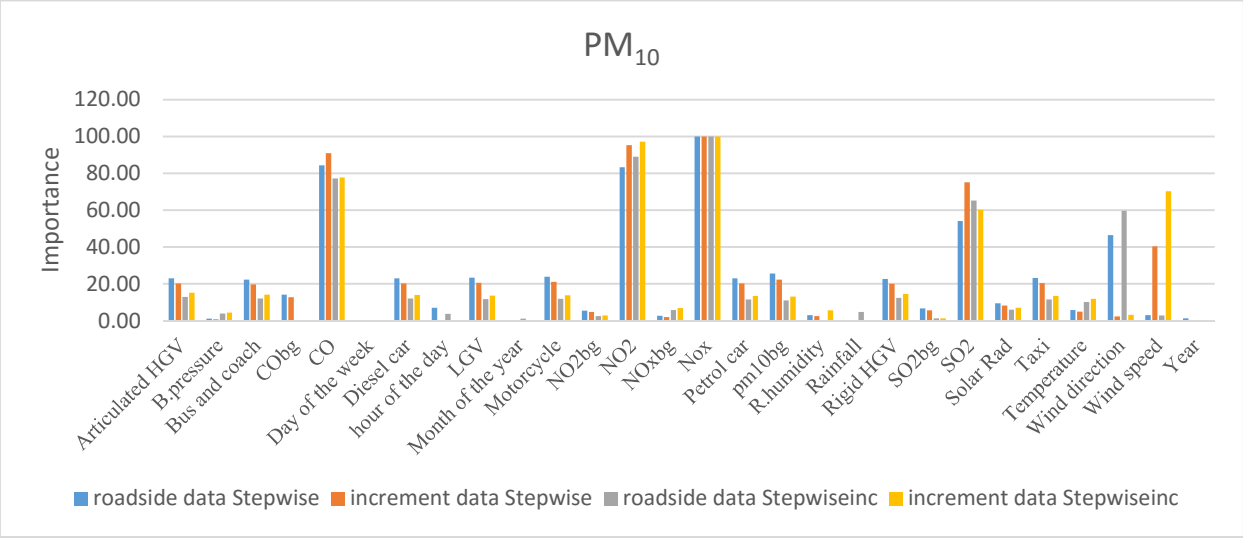


Figure D.10 Predictor variable importance for PM_{10}

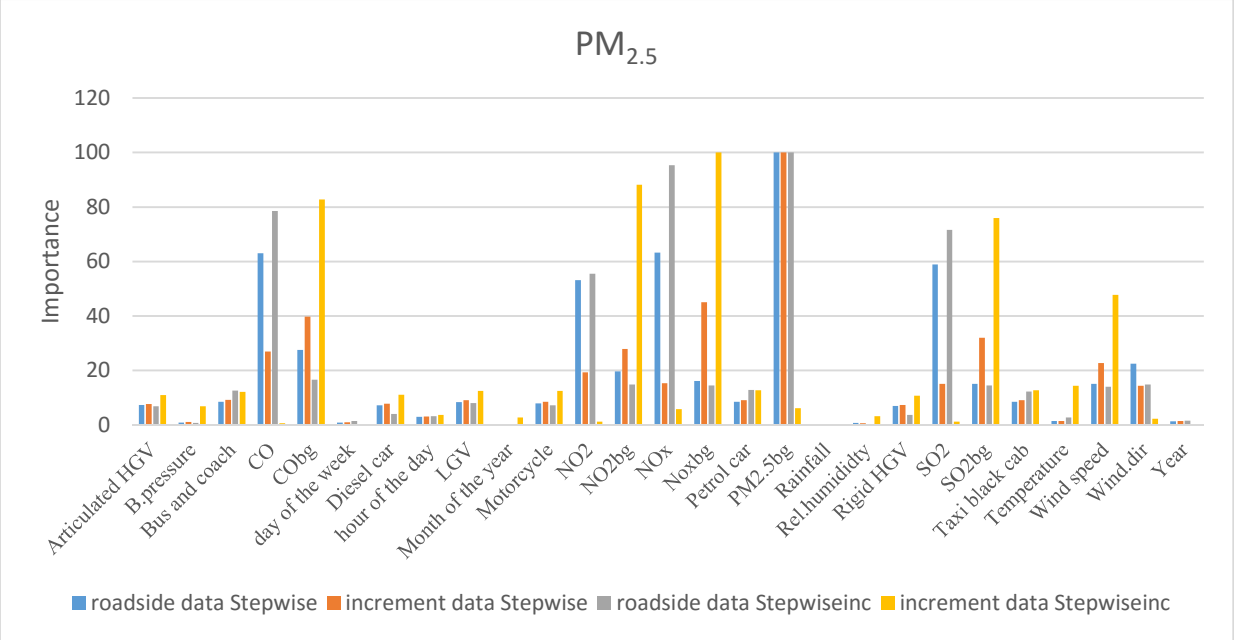


Figure D.11 Predictor variable importance for $PM_{2.5}$

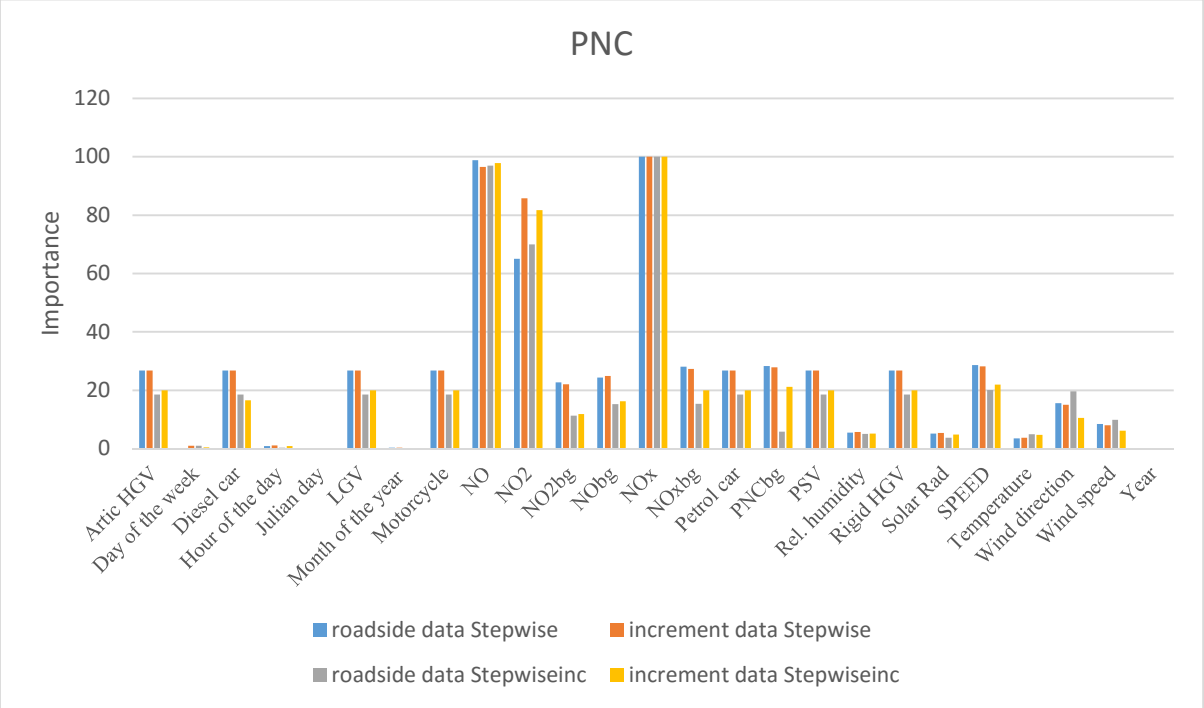


Figure D.12 Predictor variable importance for PNC

Table D.7 The test performance of the Stepwise regression models

Pollutant	model	Data type	Prediction	n	FAC2	MB	MGE	NMB	NMGE	RMSE	r	COE	IOA
PM_{10}	Stepwise.reg	roadside data	Roadside prediction	5537	0.99	-0.20	6.68	0.00	0.15	10.76	0.87	0.61	0.81
PM_{10}	Stepwise.reg	roadside data	Increment prediction	5537	0.75	-0.15	6.68	-0.01	0.32	10.74	0.86	0.60	0.80
PM_{10}	Stepwise.reg	increment data	Roadside prediction	5537	0.99	-0.06	6.61	0.00	0.15	9.76	0.89	0.61	0.81
PM_{10}	Stepwise.reg	increment data	Increment prediction	5537	0.76	0.01	6.61	0.00	0.31	9.75	0.88	0.60	0.80
$PM_{2.5}$	Stepwise.reg	roadside data	Roadside prediction	4180	0.97	0.04	3.49	0.00	0.15	5.05	0.93	0.66	0.83
$PM_{2.5}$	Stepwise.reg	roadside data	Increment prediction	4176	0.63	0.03	3.49	0.00	0.57	5.04	0.86	0.55	0.78
$PM_{2.5}$	Stepwise.reg	increment data	Roadside prediction	4176	0.97	-0.04	3.49	0.00	0.15	5.25	0.93	0.67	0.83
$PM_{2.5}$	Stepwise.reg	increment data	Increment prediction	4176	0.62	-0.03	3.49	0.00	0.57	5.23	0.85	0.56	0.78
PNC	Stepwise.reg	roadside data	Roadside prediction	2467	0.95	185	5752	0.01	0.17	10551	0.92	0.70	0.85
PNC	Stepwise.reg	roadside data	Increment prediction	2467	0.77	176	5691	0.01	0.28	10495	0.90	0.67	0.83
PNC	Stepwise.reg	increment data	Roadside prediction	2467	0.95	-223	5914	-0.01	0.18	11084	0.92	0.70	0.85
PNC	Stepwise.reg	increment data	Increment prediction	2467	0.78	-183	5912	-0.01	0.28	11071	0.90	0.66	0.83

Table D.8 Test performance of the Elastic- net regression models

Pollutant	model	Data type	Prediction	FAC2	MB	MGE	NMB	NMGE	RMSE	r	COE	IOA
PM_{10}	Elasticnet	roadside data	Roadside prediction	0.99	-0.21	6.68	0.00	0.15	10.76	0.87	0.61	0.81
PM_{10}	Elasticnet	roadside data	Increment prediction	0.76	-0.15	6.68	-0.01	0.32	10.74	0.86	0.60	0.80
PM_{10}	Elasticnet	increment data	Roadside prediction	0.99	-0.07	6.61	0.00	0.15	9.76	0.89	0.61	0.81
PM_{10}	Elasticnet	increment data	Increment prediction	0.76	0.00	6.62	0.00	0.31	9.75	0.88	0.60	0.80
$PM_{2.5}$	Elasticnet	roadside data	Roadside prediction	0.97	0.04	3.50	0.00	0.15	5.05	0.93	0.66	0.83
$PM_{2.5}$	Elasticnet	roadside data	Increment prediction	0.62	0.03	3.49	0.00	0.57	5.04	0.86	0.55	0.78
$PM_{2.5}$	Elasticnet	increment data	Roadside prediction	0.97	-0.04	3.49	0.00	0.15	5.25	0.93	0.67	0.83
$PM_{2.5}$	Elasticnet	increment data	Increment prediction	0.62	-0.03	3.49	0.00	0.57	5.23	0.85	0.56	0.78
PNC	Elasticnet	roadside data	Roadside prediction	0.96	158	5713	0.00	0.17	10563	0.92	0.71	0.85
PNC	Elasticnet	roadside data	Increment prediction	0.78	144	5648	0.01	0.27	10512	0.90	0.67	0.84
PNC	Elasticnet	increment data	Roadside prediction	0.96	-252	5857	-0.01	0.18	11095	0.92	0.70	0.85
PNC	Elasticnet	increment data	Increment prediction	0.79	-201	5876	-0.01	0.28	11092	0.90	0.66	0.83

Appendix E Performance of Hybrid Statistical Methods

Table E.1 Comparison of the performance of feature selection methods for $PM_{2.5}$ models

Row Labels	FAC2.	MB.	MGE.	NMB.	NMGE.	RMSE.	R.	COE.	IOA.
ENET									
GA-RF	0.97	-0.04	3.49	0.00	0.15	5.25	0.93	0.67	0.83
Linear	0.97	0.04	3.50	0.00	0.15	5.05	0.93	0.66	0.83
SA-RF	0.98	0.05	3.56	0.00	0.15	4.97	0.93	0.66	0.83
MLR									
GA-RF	0.97	-0.04	3.49	0.00	0.15	5.24	0.93	0.67	0.83
Linear	0.97	0.04	3.49	0.00	0.15	5.05	0.93	0.66	0.83
SA-RF	0.97	0.05	3.56	0.00	0.15	4.97	0.93	0.66	0.83
PCR									
GA-RF	0.97	-0.05	3.57	0.00	0.15	5.31	0.93	0.66	0.83
Linear	0.97	0.05	3.59	0.00	0.15	5.13	0.93	0.65	0.83
SA-RF	0.97	0.04	3.72	0.00	0.16	5.11	0.93	0.64	0.82
PLSR									
GA-RF	0.97	-0.03	3.49	0.00	0.15	5.24	0.93	0.67	0.83
Linear	0.97	0.04	3.50	0.00	0.15	5.05	0.93	0.66	0.83
SA-RF	0.97	0.05	3.56	0.00	0.15	4.97	0.93	0.66	0.83
STEPWISE.REG									
GA-RF	0.97	-0.04	3.49	0.00	0.15	5.25	0.93	0.67	0.83
Linear	0.97	0.04	3.49	0.00	0.15	5.05	0.93	0.66	0.83
SA-RF	0.97	0.05	3.56	0.00	0.15	4.97	0.93	0.66	0.83

Table E.2 Comparison of the performance of feature selection methods for PNC models

Row Labels	FAC2.	MB.	MGE.	NMB.	NMGE.	RMSE.	R.	COE.	IOA.
ENET									
GA-RF	0.96	129.87	5606.89	0.00	0.17	9943.52	0.93	0.71	0.85
Linear	0.96	157.57	5712.75	0.00	0.17	10562.68	0.92	0.71	0.85
SA-RF	0.96	129.87	5606.89	0.00	0.17	9943.52	0.93	0.71	0.85
MLR									
GA-RF	0.96	130.64	5617.99	0.00	0.17	9962.21	0.93	0.71	0.85
Linear	0.95	185.89	5761.36	0.01	0.18	10556.95	0.92	0.70	0.85
SA-RF	0.96	130.64	5617.99	0.00	0.17	9962.21	0.93	0.71	0.85
PCR									
GA-RF	0.96	123.89	5710.54	0.00	0.17	10005.91	0.93	0.70	0.85
Linear	0.95	180.46	5759.80	0.01	0.18	10556.77	0.92	0.70	0.85
SA-RF	0.96	123.89	5710.54	0.00	0.17	10005.91	0.93	0.70	0.85
PLSR									
GA-RF	0.96	130.72	5618.40	0.00	0.17	9963.12	0.93	0.71	0.85
Linear	0.95	183.24	5761.02	0.01	0.18	10555.51	0.92	0.70	0.85
SA-RF	0.96	130.72	5618.40	0.00	0.17	9963.12	0.93	0.71	0.85
STEPWISE.REG									
GA-RF	0.96	130.64	5617.99	0.00	0.17	9962.21	0.93	0.71	0.85
Linear	0.95	185.06	5751.76	0.01	0.17	10550.48	0.92	0.70	0.85
SA-RF	0.96	130.64	5617.99	0.00	0.17	9962.21	0.93	0.71	0.85

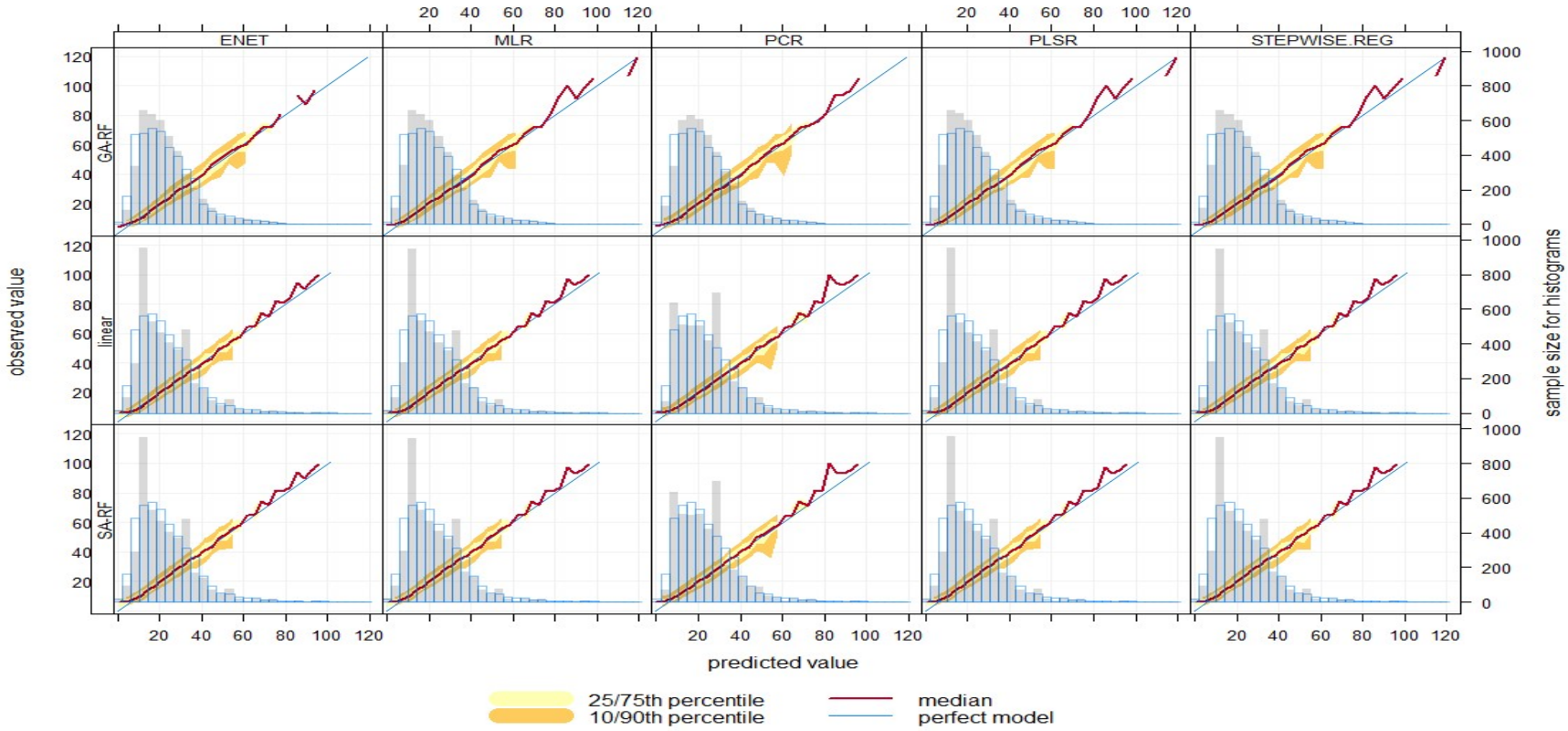


Figure E.1 Conditional Quantile plots comparing the performance of $PM_{2.5}$ ($\mu g/m^3$) models

Note: predicted value and observed value are modelled and observed $PM_{2.5}$ concentrations respectively

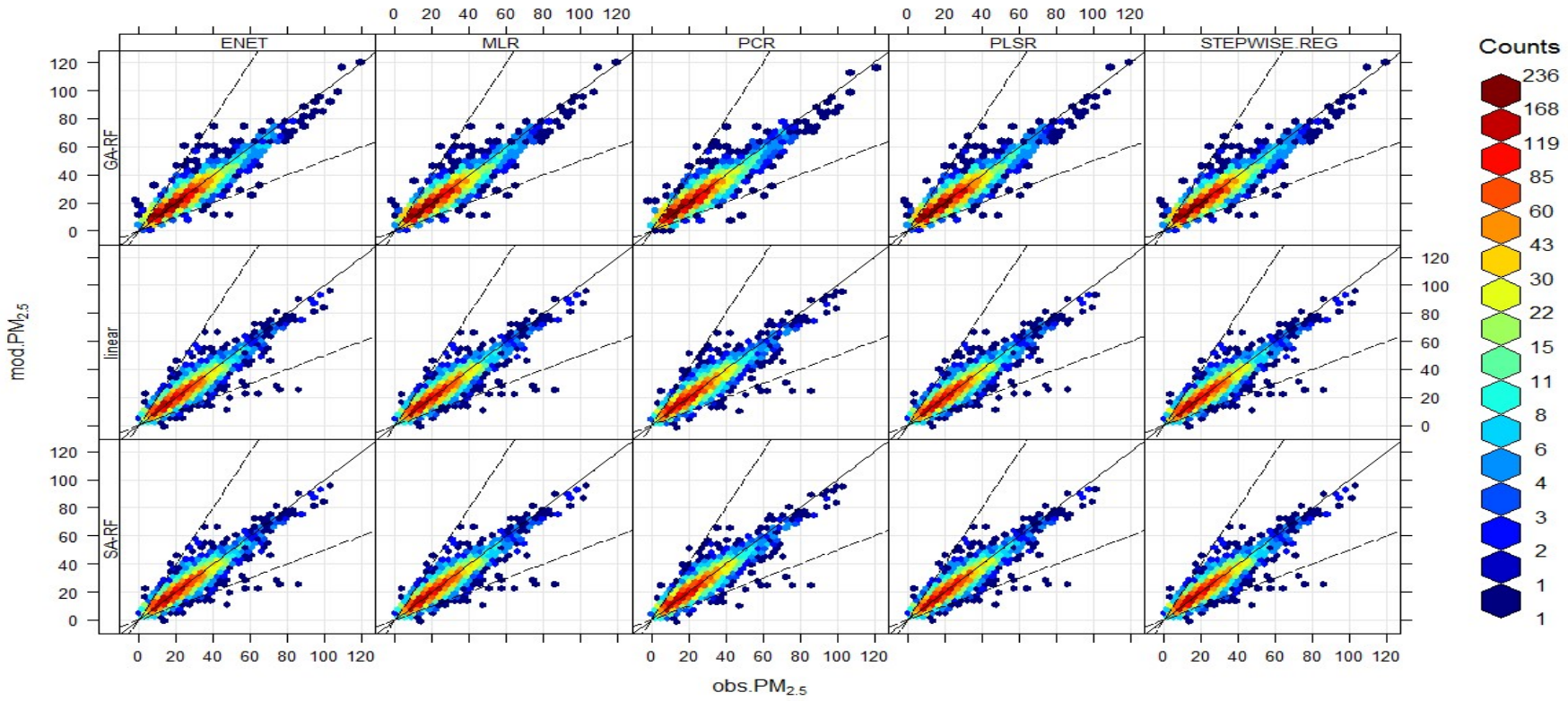


Figure E.2 Scatter plots comparing the performance of $PM_{2.5}$ ($\mu g/m^3$) models

Note: mod$PM_{2.5}$ and obs$PM_{2.5}$ are modelled and observed $PM_{2.5}$ concentrations respectively

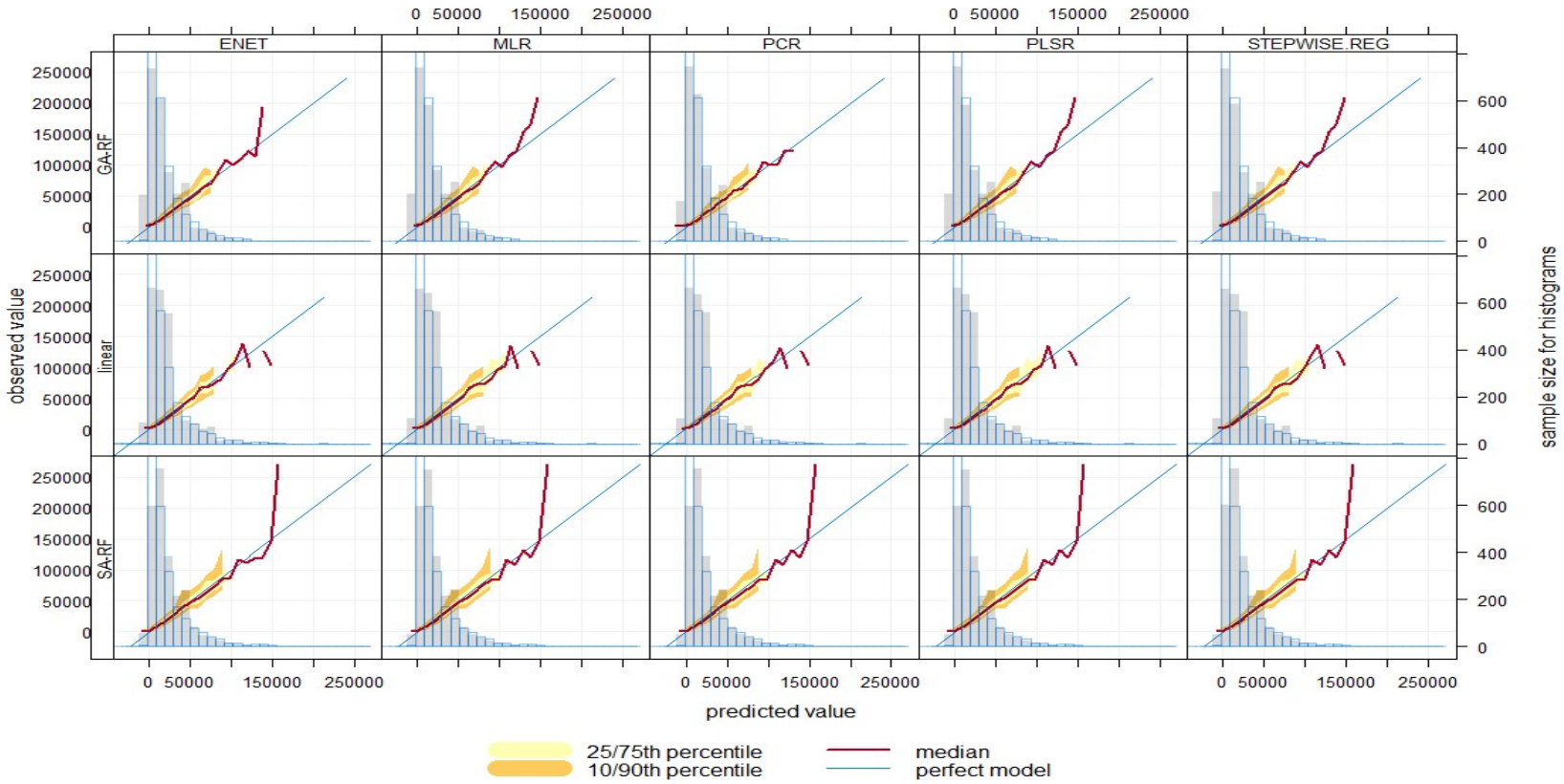


Figure E.3 conditional quantile plots comparing the performance of PNC (number/cm^3) models

Note: predicted value and observed value are modelled and observed PNC concentrations respectively

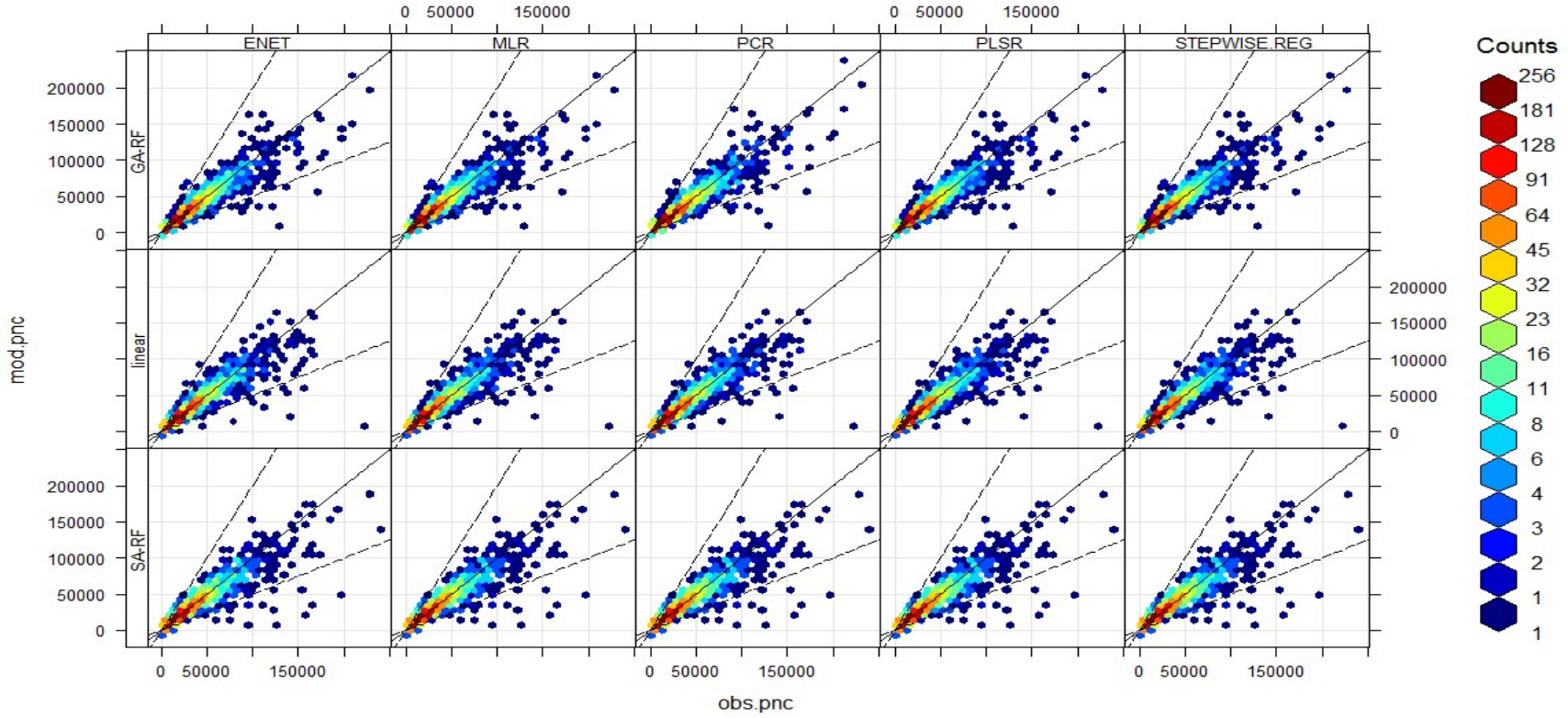


Figure E.4 Scatter plots comparing the performance of PNC models

Note: modPNC and obsPNC are modelled and observed PNC (number/cm^3) *concentrations respectively*

Appendix F Machine Learning Models

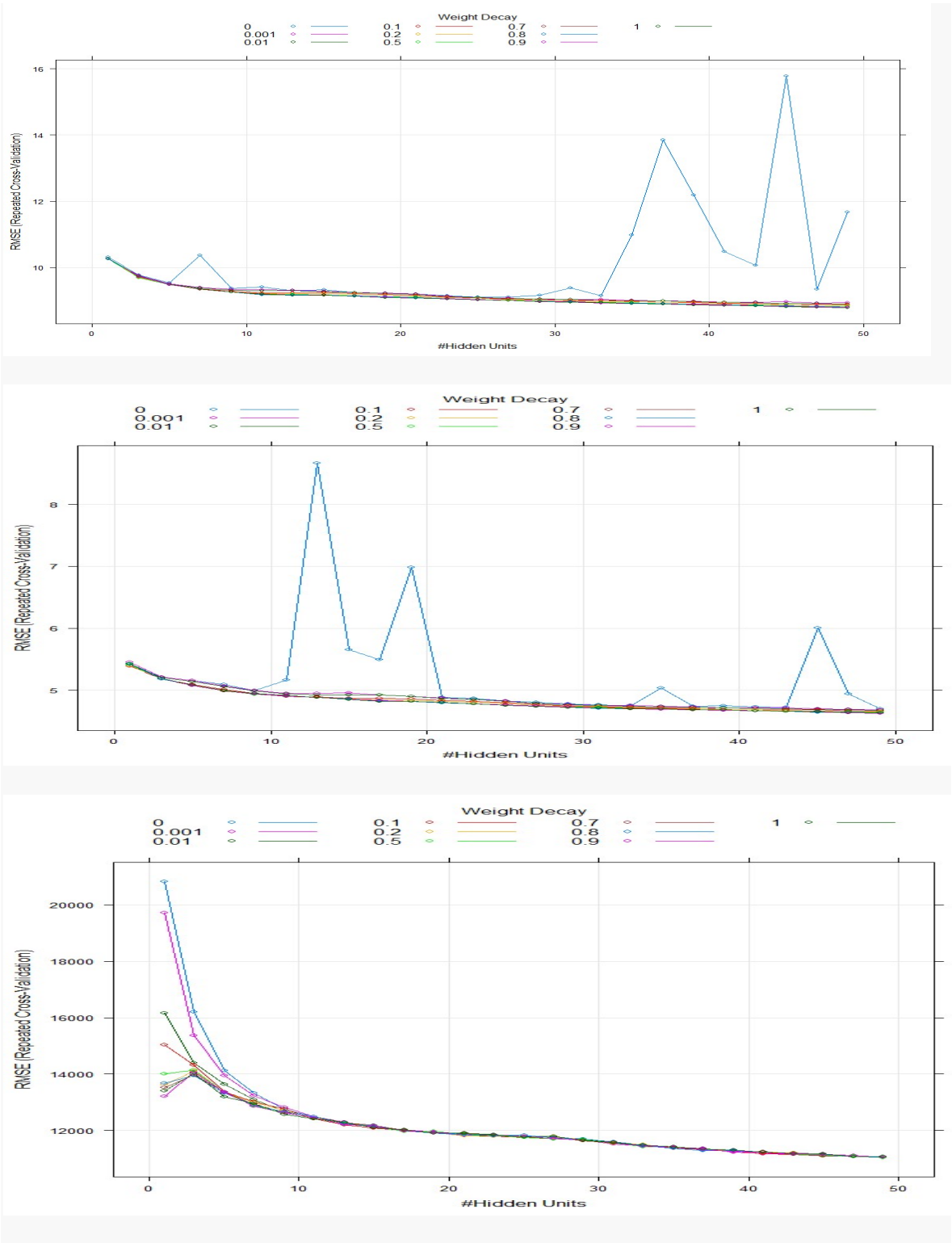


Figure F.1 Optimisation of AVG-MLP model parameters for PM_{10}(top), $PM_{2.5}$(middle), and PNC (bottom) respectively

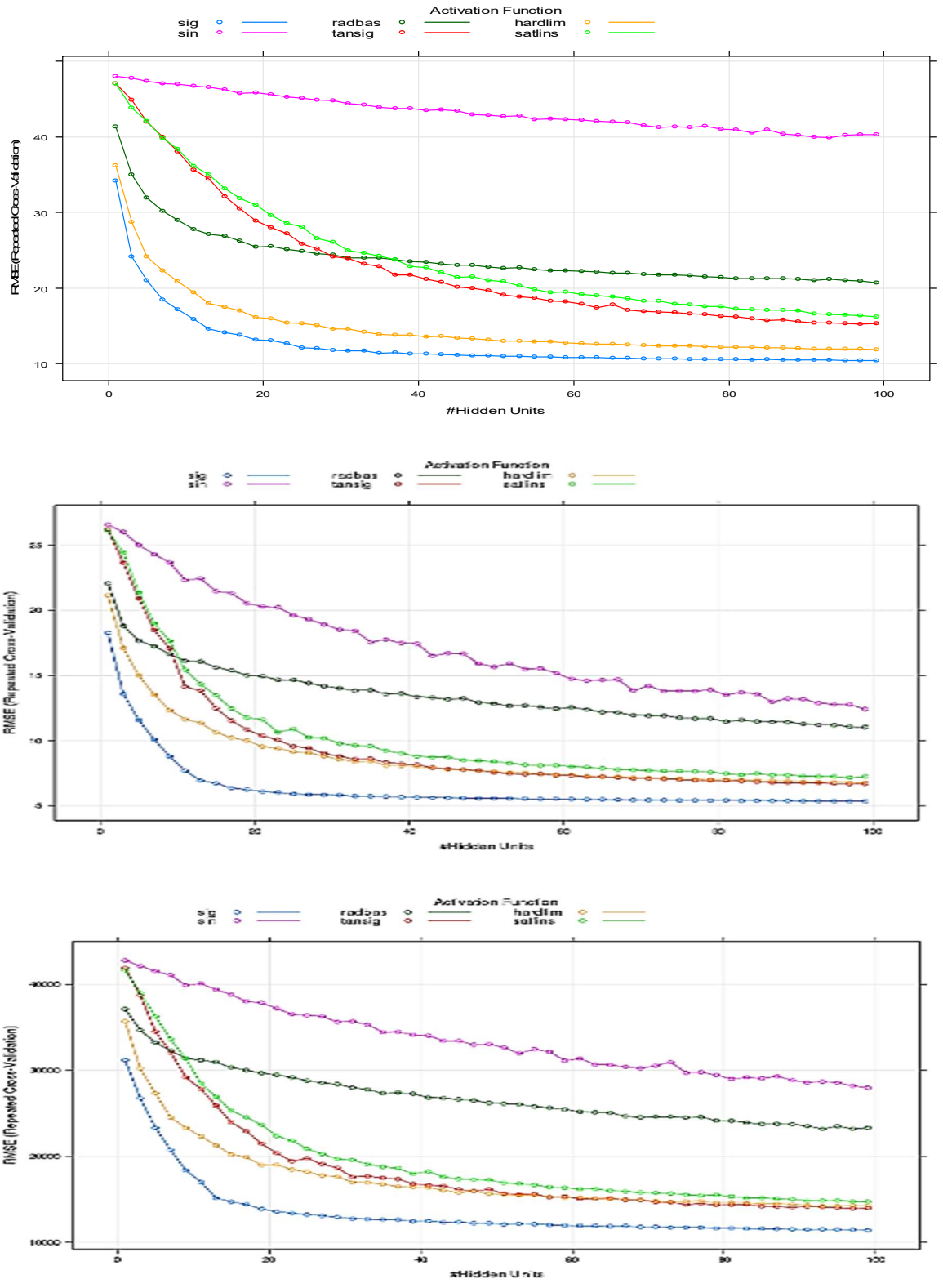


Figure F.2 Optimisation of ELM model parameters for PM_{10}(top), $PM_{2.5}$(middle), and PNC (bottom).

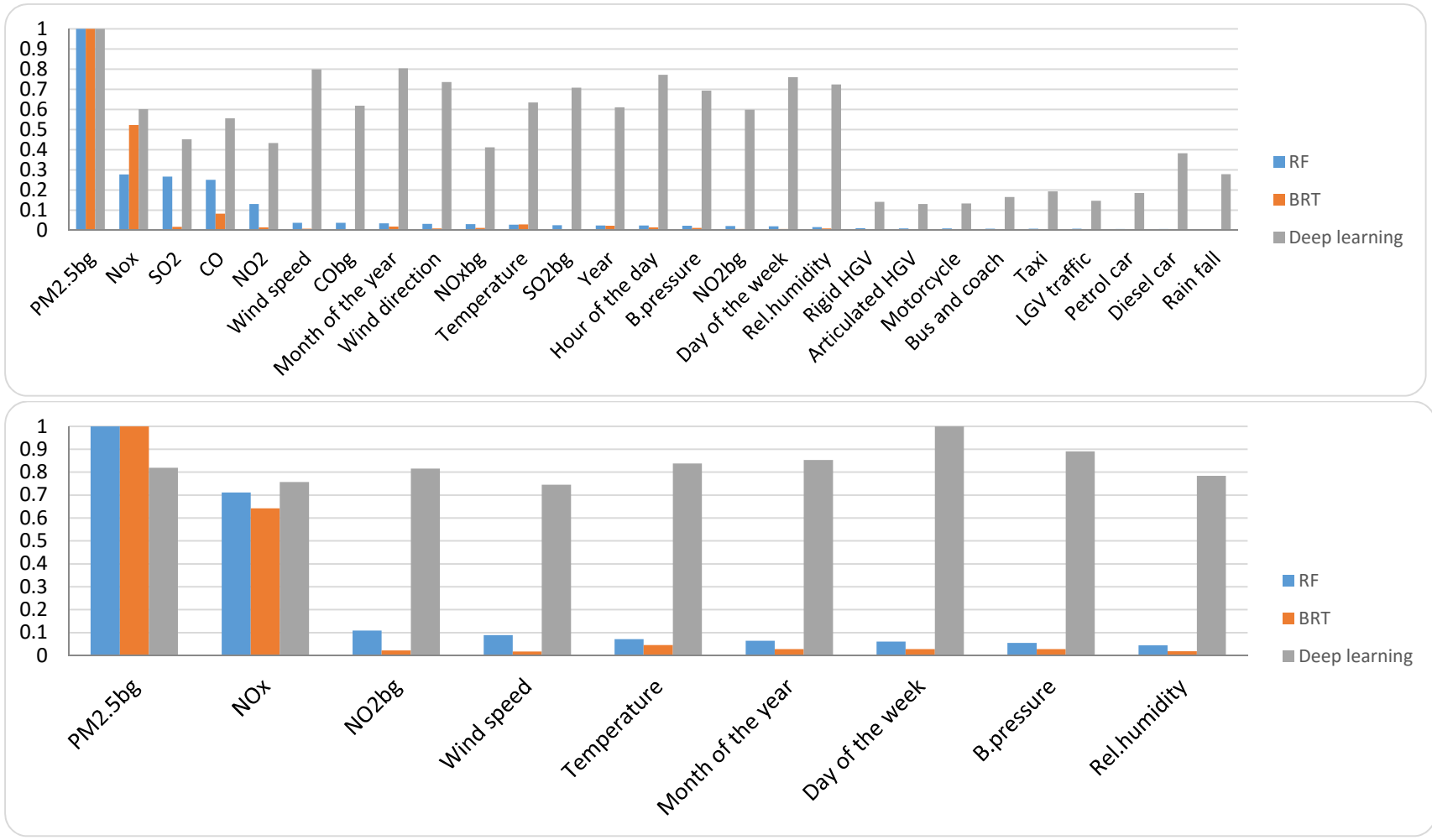


Figure F.3 Variable importance estimated by ML models for the prediction of $PM_{2.5}$

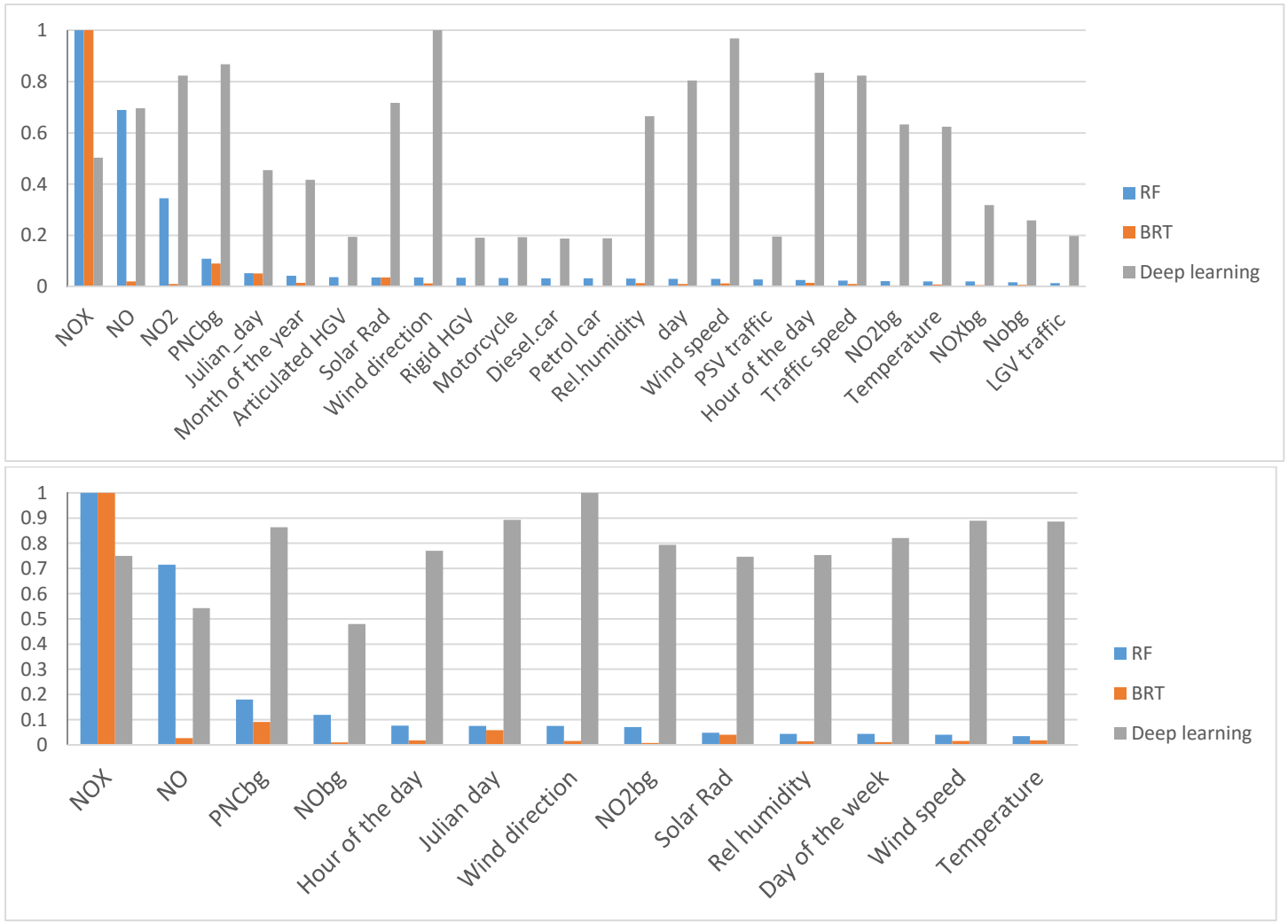


Figure F.4 Variable importance estimated by ML models for the prediction of PNC

Table F.1 Training Results for the ANN models

Row Labels	IOA	COE	R	RMSE	NMGE	NMB	MGE	MB	FAC2
$PM_{2.5}$									
all variables									
AVG-MLP	0.80	0.60	0.93	5.85	0.18	-0.01	4.20	-0.14	0.96
BRNN	0.85	0.71	0.95	4.35	0.13	0.00	3.01	-0.01	0.98
BRT	0.86	0.71	0.95	4.31	0.13	0.00	2.97	0.01	0.98
Deep learning	0.84	0.67	0.94	4.80	0.15	-0.03	3.38	-0.66	0.97
ELM	0.84	0.67	0.94	4.75	0.15	0.00	3.40	0.01	0.97
PCA-MLP	0.85	0.69	0.94	4.49	0.14	0.00	3.19	-0.02	0.98
RF	0.85	0.70	0.94	4.51	0.13	0.00	3.13	0.07	0.98
SVM linear	0.83	0.66	0.93	5.12	0.15	0.00	3.49	0.00	0.97
SVM radial	0.86	0.72	0.95	4.25	0.13	-0.01	2.91	-0.14	0.98
RF-GA									
AVG-MLP	0.85	0.71	0.95	4.32	0.13	0.00	3.04	0.06	0.98
BRNN	0.84	0.68	0.94	4.61	0.14	0.00	3.36	0.03	0.98
BRT	0.85	0.70	0.95	4.27	0.13	0.00	3.11	0.04	0.98
Deep learning	0.83	0.66	0.94	4.80	0.15	-0.02	3.51	-0.5	0.98
ELM	0.84	0.68	0.94	4.54	0.14	0.00	3.32	0.07	0.98
PCA-MLP	0.84	0.68	0.95	4.51	0.14	0.00	3.32	0.07	0.98
RF	0.85	0.70	0.95	4.27	0.13	0.01	3.11	0.14	0.98
SVM linear	0.83	0.66	0.93	5.00	0.15	-0.01	3.56	-0.23	0.98
SVM radial	0.85	0.70	0.95	4.33	0.13	0.00	3.12	-0.03	0.98

Table F.1 continues

Row Labels	IOA	COE	R	RMSE	NMGE	NMB	MGE	MB	FAC2
PNC									
all variables									
AVG-MLP	0.85	0.71	0.92	10344	0.17	0.01	5693	342	0.97
BRNN	0.87	0.74	0.94	9561	0.15	-0.01	5017	-243	0.98
BRT	0.88	0.77	0.94	8787	0.14	0.00	4532	2	0.99
Deep learning	0.87	0.74	0.94	9390	0.15	0.03	5072	905	0.99
ELM	0.85	0.70	0.93	10061	0.18	0.01	5899	227	0.96
PCA-MLP	0.83	0.65	0.90	11415	0.21	0.00	6760	95	0.94
RF	0.88	0.77	0.94	9032	0.14	0.00	4494	89	0.99
SVM linear	0.79	0.58	0.88	15119	0.25	-0.12	8193	-4058	0.96
SVM radial	0.88	0.76	0.93	9889	0.14	-0.04	4572	-1268	0.99
RF-GA									
AVG-MLP	0.86	0.73	0.94	9134	0.16	0.00	5225	97	0.97
BRNN	0.87	0.74	0.95	8771	0.15	0.01	4922	247	0.98
BRT	0.88	0.76	0.95	8150	0.14	0.00	4547	-128	0.99
Deep learning	0.87	0.73	0.95	8820	0.16	0.05	5131	1600	0.98
ELM	0.85	0.70	0.94	9040	0.17	0.01	5673	320	0.96
PCA-MLP	0.84	0.69	0.93	9999	0.18	0.01	5964	248	0.95
RF	0.89	0.77	0.96	7955	0.13	0.00	4316	101	0.99
SVM linear	0.86	0.72	0.93	10397	0.16	-0.04	5283	-1205	0.98
SVM radial	0.89	0.77	0.95	8310	0.13	-0.01	4392	-431	0.99

Table F.2 Test performance statistics for the ANN models

Row Labels	IOA	COE	R	FAC2	RMSE	NMGE.	NMB.	MGE.	MB.
$PM_{2.5}$									
all variables									
AVG-MLP	0.80	0.60	0.93	0.96	5.85	0.18	-0.01	4.20	-0.14
BRNN	0.85	0.71	0.95	0.98	4.35	0.13	0.00	3.01	-0.01
Deep learning	0.84	0.67	0.94	0.97	4.80	0.15	-0.03	3.38	-0.66
ELM	0.84	0.67	0.94	0.97	4.75	0.15	0.00	3.40	0.01
PCA-MLP	0.85	0.69	0.94	0.98	4.49	0.14	0.00	3.19	-0.02
RF_GA									
AVG-MLP	0.85	0.71	0.95	0.98	4.32	0.13	0.00	3.04	0.06
BRNN	0.84	0.68	0.94	0.98	4.61	0.14	0.00	3.36	0.03
Deep learning	0.83	0.66	0.94	0.98	4.80	0.15	-0.02	3.51	-0.50
ELM	0.84	0.68	0.94	0.98	4.54	0.14	0.00	3.32	0.07
PCA-MLP	0.84	0.68	0.95	0.98	4.51	0.14	0.00	3.32	0.07
PNC									
all variables									
AVG-MLP	0.85	0.71	0.92	0.97	10344	0.17	0.01	5693	342
BRNN	0.87	0.74	0.94	0.98	9561	0.15	-0.01	5017	-243
Deep learning	0.87	0.74	0.94	0.99	9390	0.15	0.03	5072	905
ELM	0.85	0.70	0.93	0.96	10061	0.18	0.01	5899	227
PCA-MLP	0.83	0.65	0.9	0.94	11415	0.21	0.00	6760	95
RF_GA									
AVG-MLP	0.86	0.73	0.94	0.97	9134	0.16	0.00	5225	97
BRNN	0.87	0.74	0.95	0.98	8771	0.15	0.01	4922	247
Deep learning	0.87	0.73	0.95	0.98	8820	0.16	0.05	5131	1600
ELM	0.85	0.70	0.94	0.96	9040	0.17	0.01	5673	320
PCA-MLP	0.84	0.69	0.93	0.95	9999	0.18	0.01	5964	248

Table F.3 Training Results for BRT models

Row Labels	R-squared	RMSE
$PM_{2.5}$		
All variables		
BRT		
Best Training incident	0.97	2.24
Cross validation	0.91	4.24
RF		
Best Training incident	0.90	4.44
Cross validation	0.90	4.46
RF-GA		
BRT		
Best Training incident	0.96	2.85
Cross validation	0.89	4.64
RF		
Best Training incident	0.88	4.69
Cross validation	0.88	4.72
PNC		
All variables		
BRT		
Best Training incident	0.99	3109
Cross validation	0.89	9027
RF		
Best Training incident	0.88	9528
Cross validation	0.88	9538
RF-GA		
BRT		
Best Training incident	0.98	3829
Cross validation	0.89	93169
RF		
Best Training incident	0.88	9392
Cross validation	0.88	9532

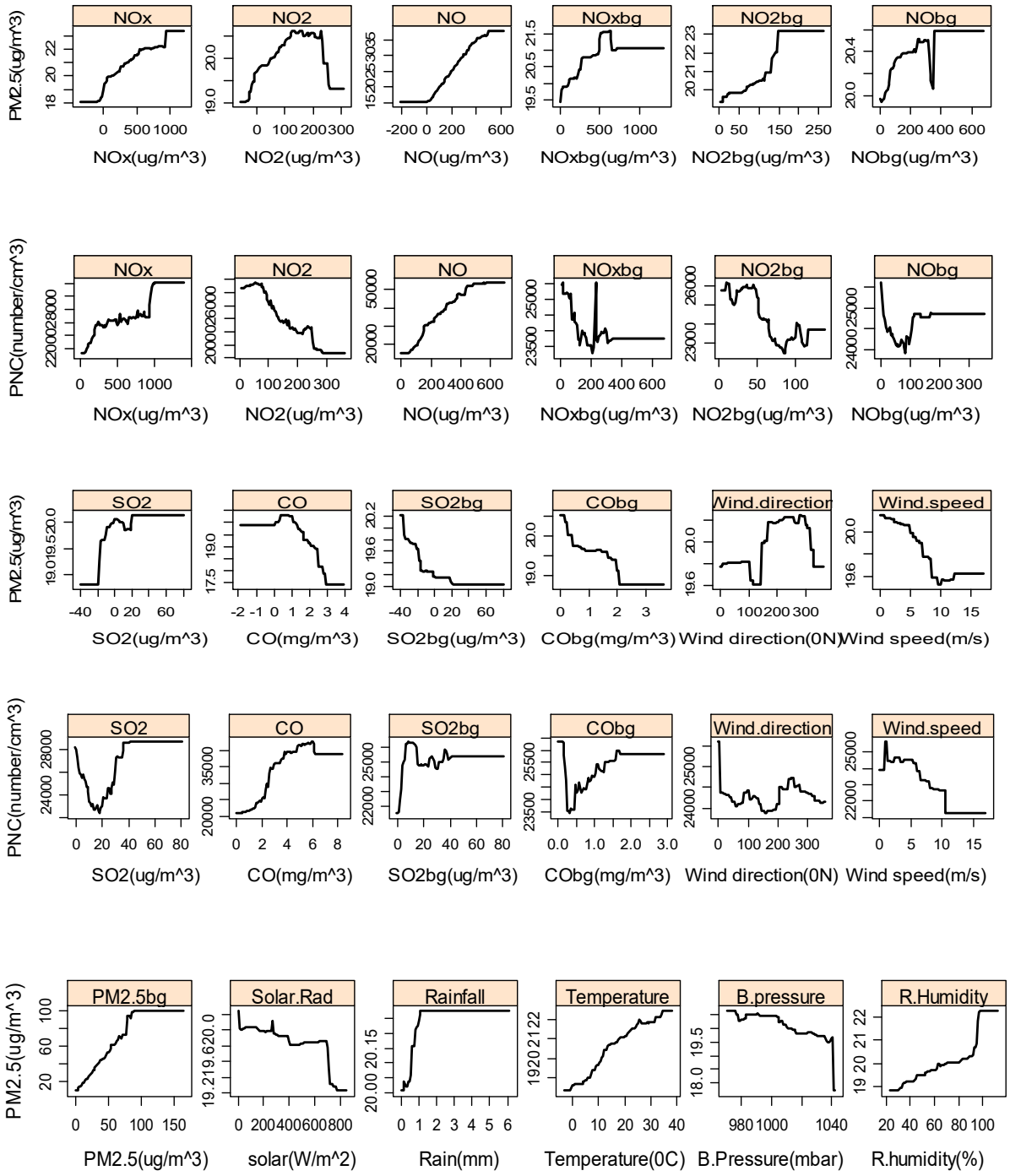


Figure F.5. Partial dependence plots showing the effects of pollutants and wind variables on the BRT model predictions of the roadside particle concentrations.

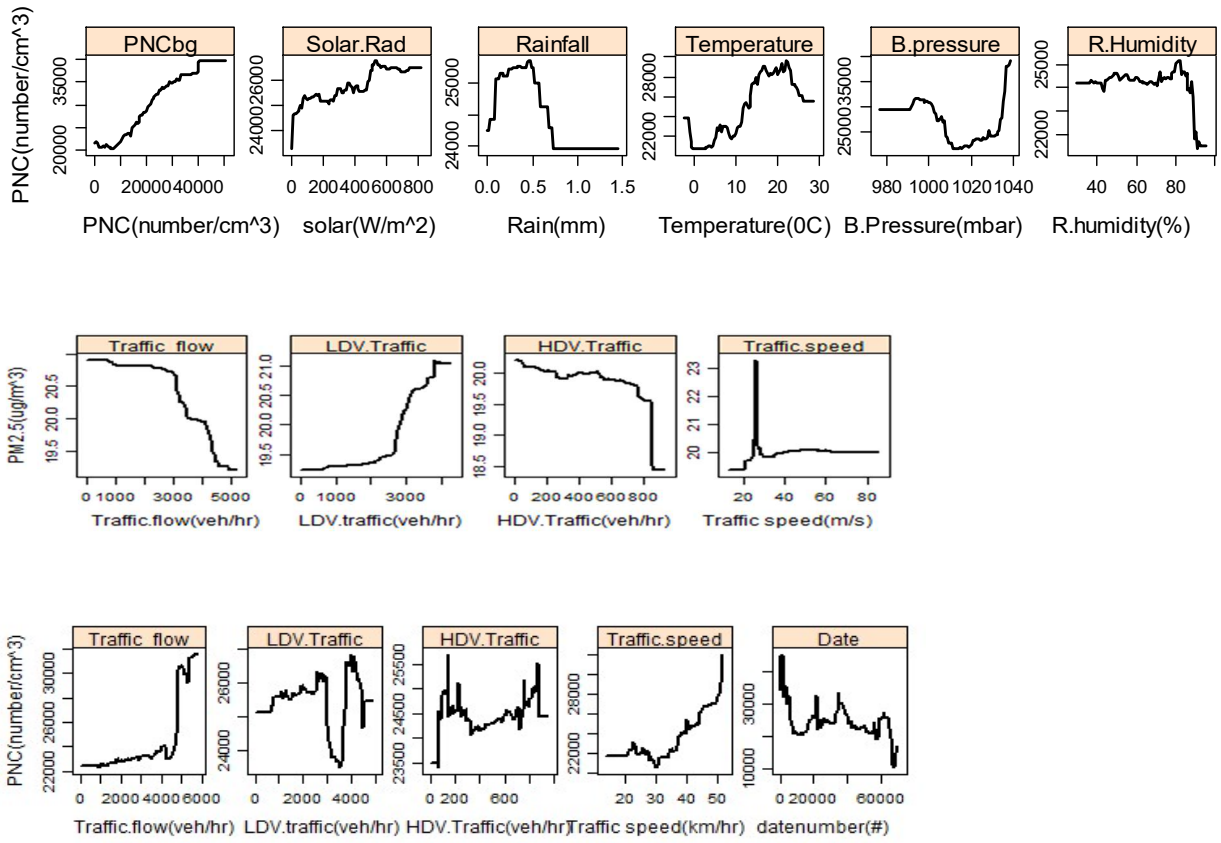


Figure F.5 *continued*

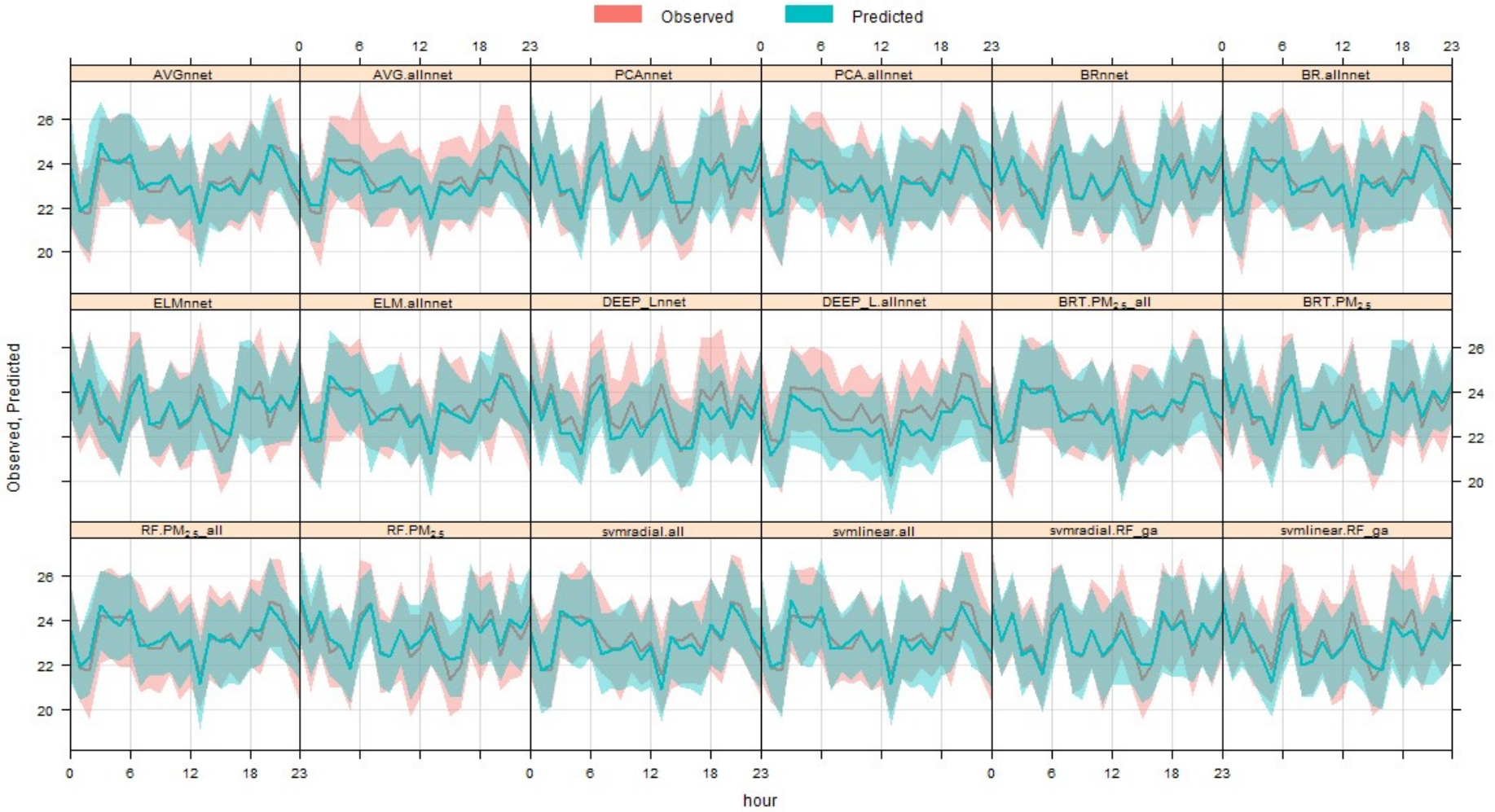


Figure F.6 Hourly variation plots comparing the pattern of the $PM_{2.5}$ (µg/m3) prediction of the ML models and the observed $PM_{2.5}$ concentrations.

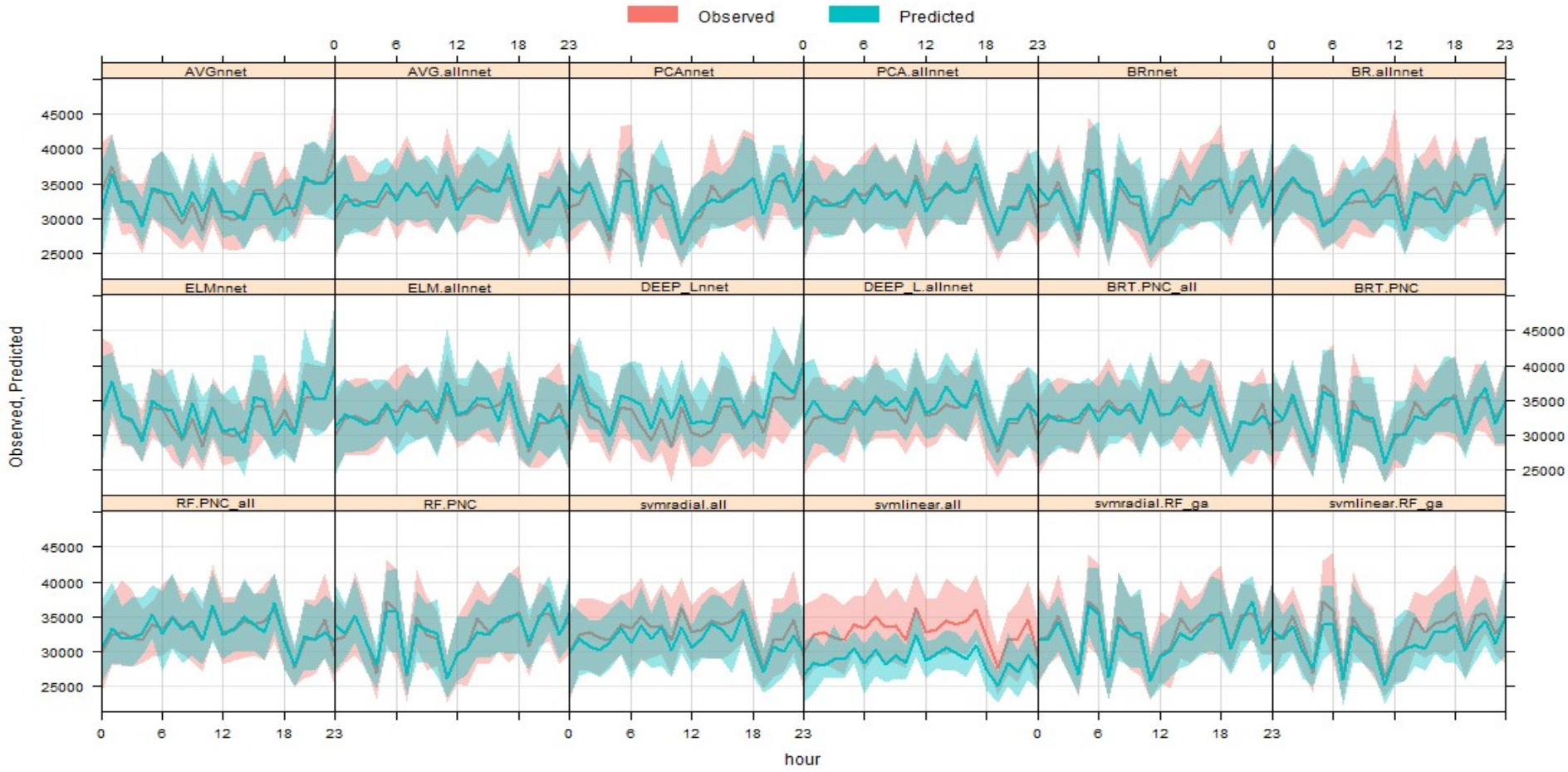


Figure F.7 Hourly variation plots comparing the pattern of the PNC (number/cm^3) prediction of the ML models and the observed PNC concentrations.

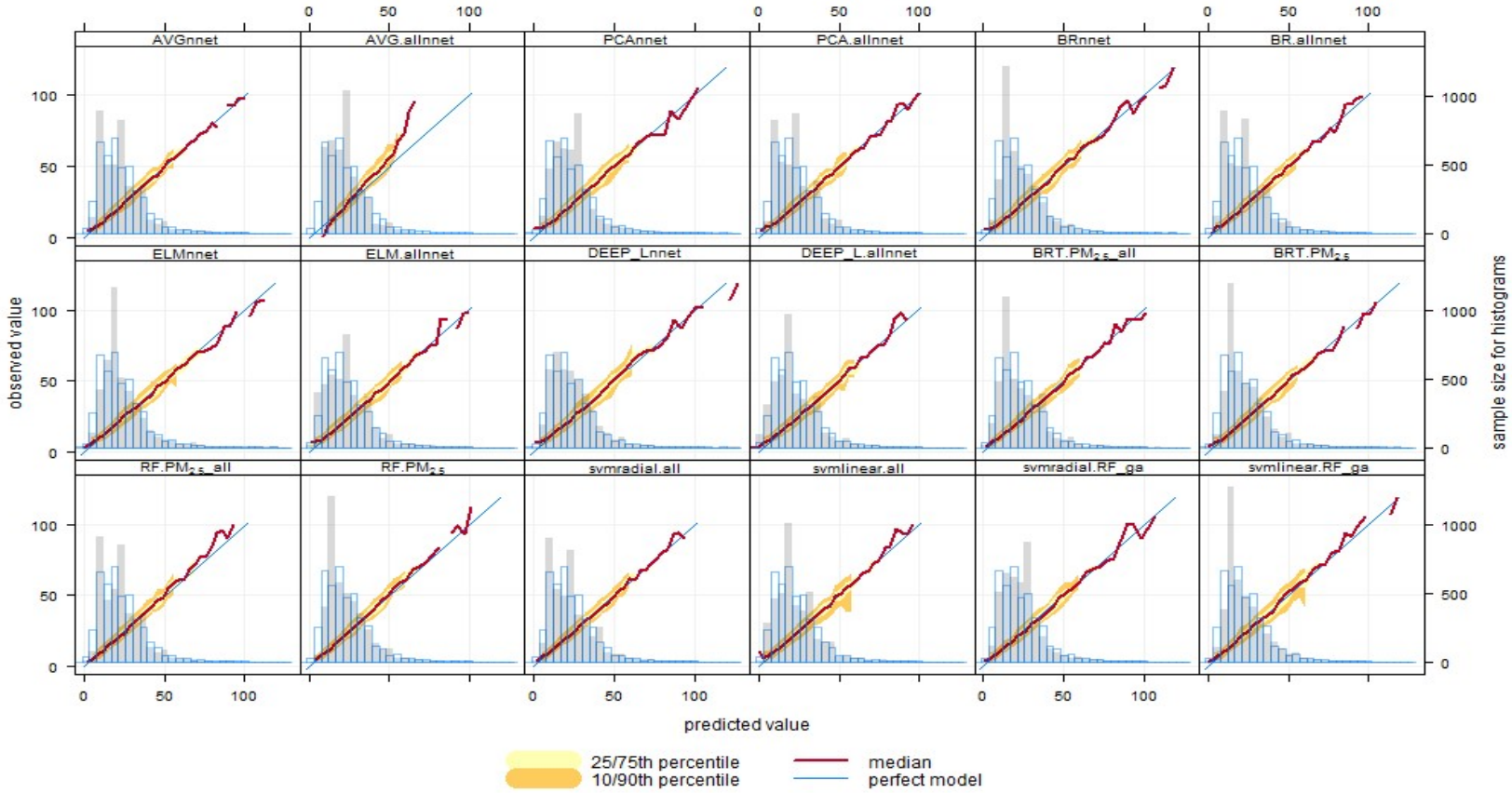


Figure F.8 conditional quantile plots showing the agreement between the observed and Machine learning predictions of the $PM_{2.5}$ ($\mu g/m^3$) concentrations.

Note: predicted value and observed value are modelled and observed $PM_{2.5}$ concentrations respectively

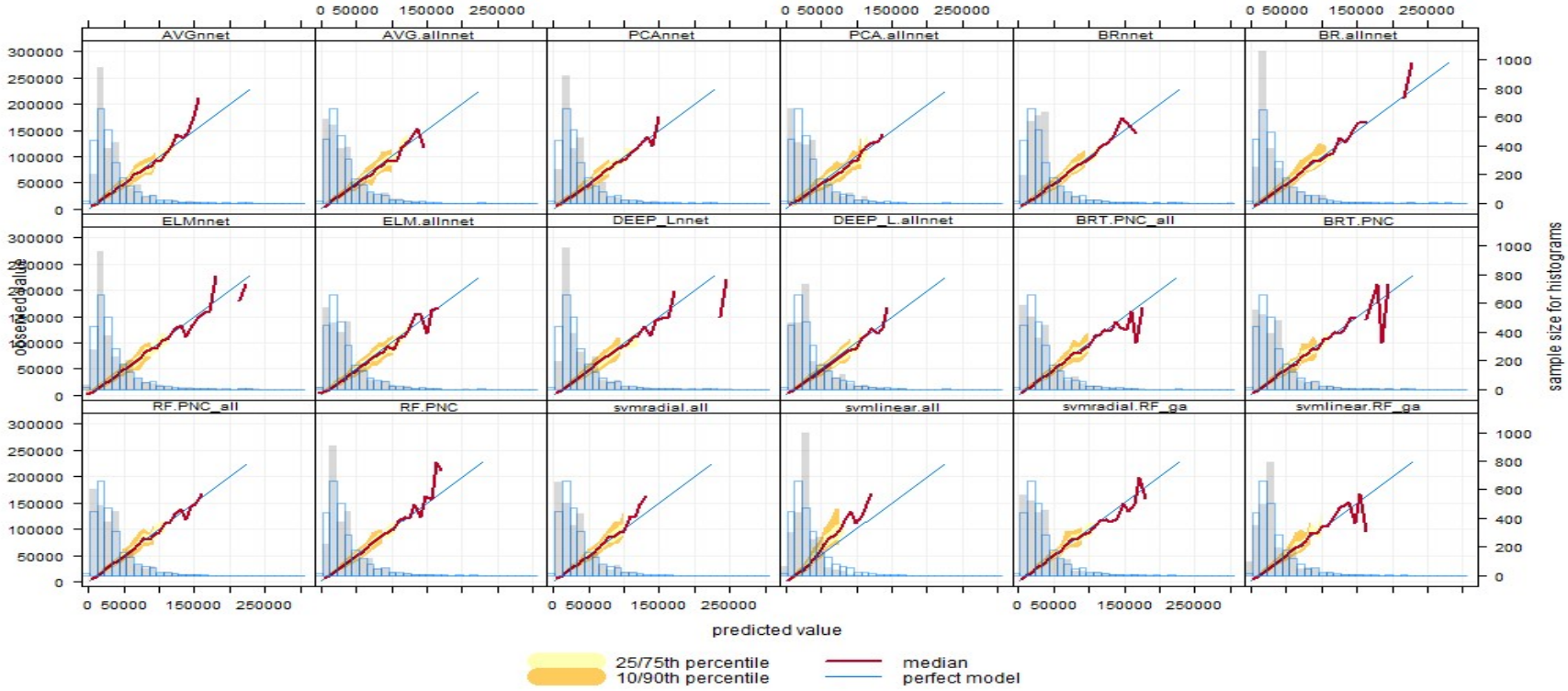


Figure F.9 conditional quantile plots showing the agreement between the observed and Machine learning predictions of the PNC (number/cm^3) concentration. *Note: predicted value and observed value are modelled and observed PNC concentrations respectively*

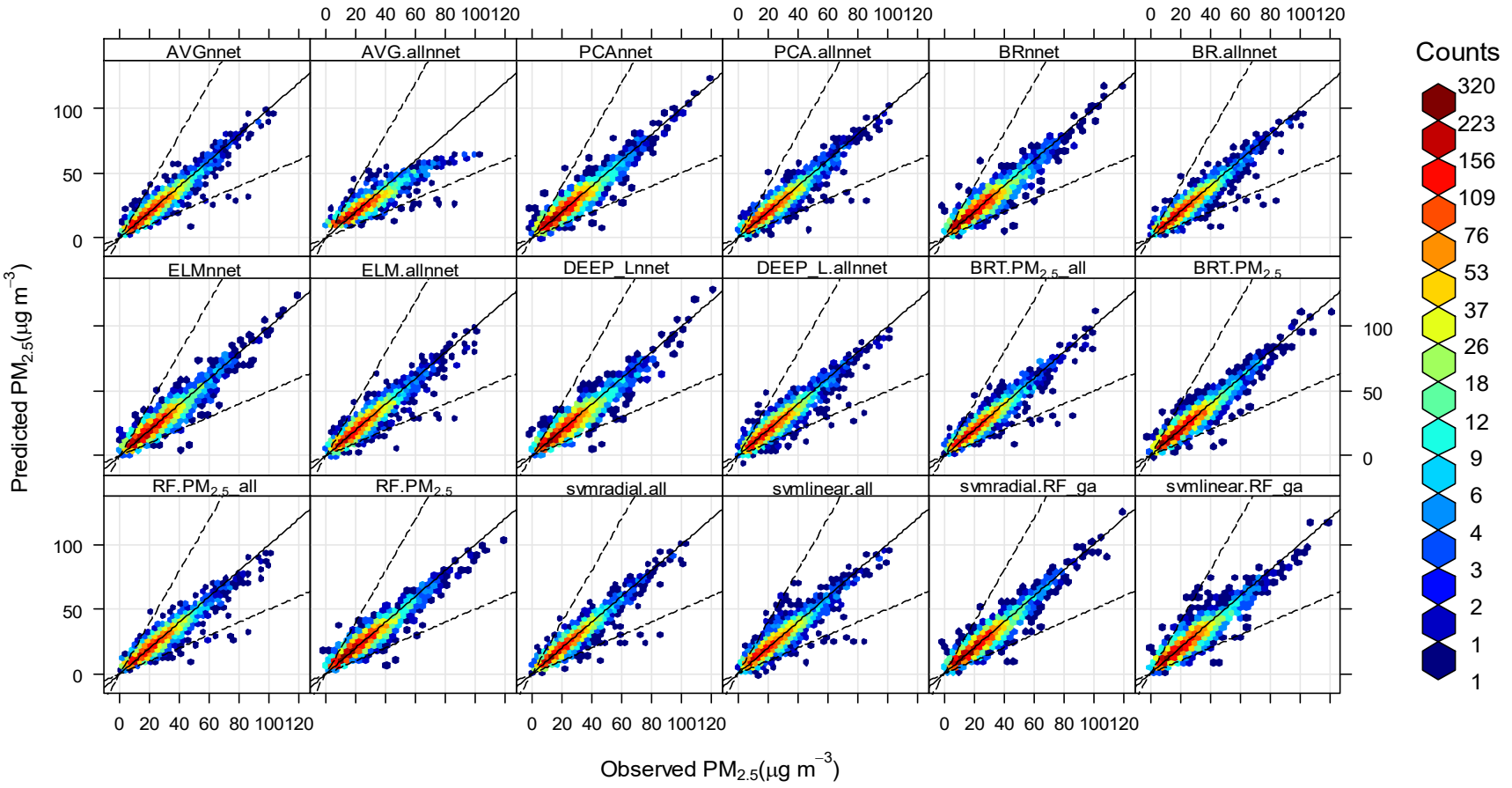


Figure F.4 Scatter plots comparing the prediction of the ML models and the observed $PM_{2.5}$ concentrations.

Note: mod$PM_{2.5}$ and obs$PM_{2.5}$ are modelled and observed $PM_{2.5}$ concentrations respectively

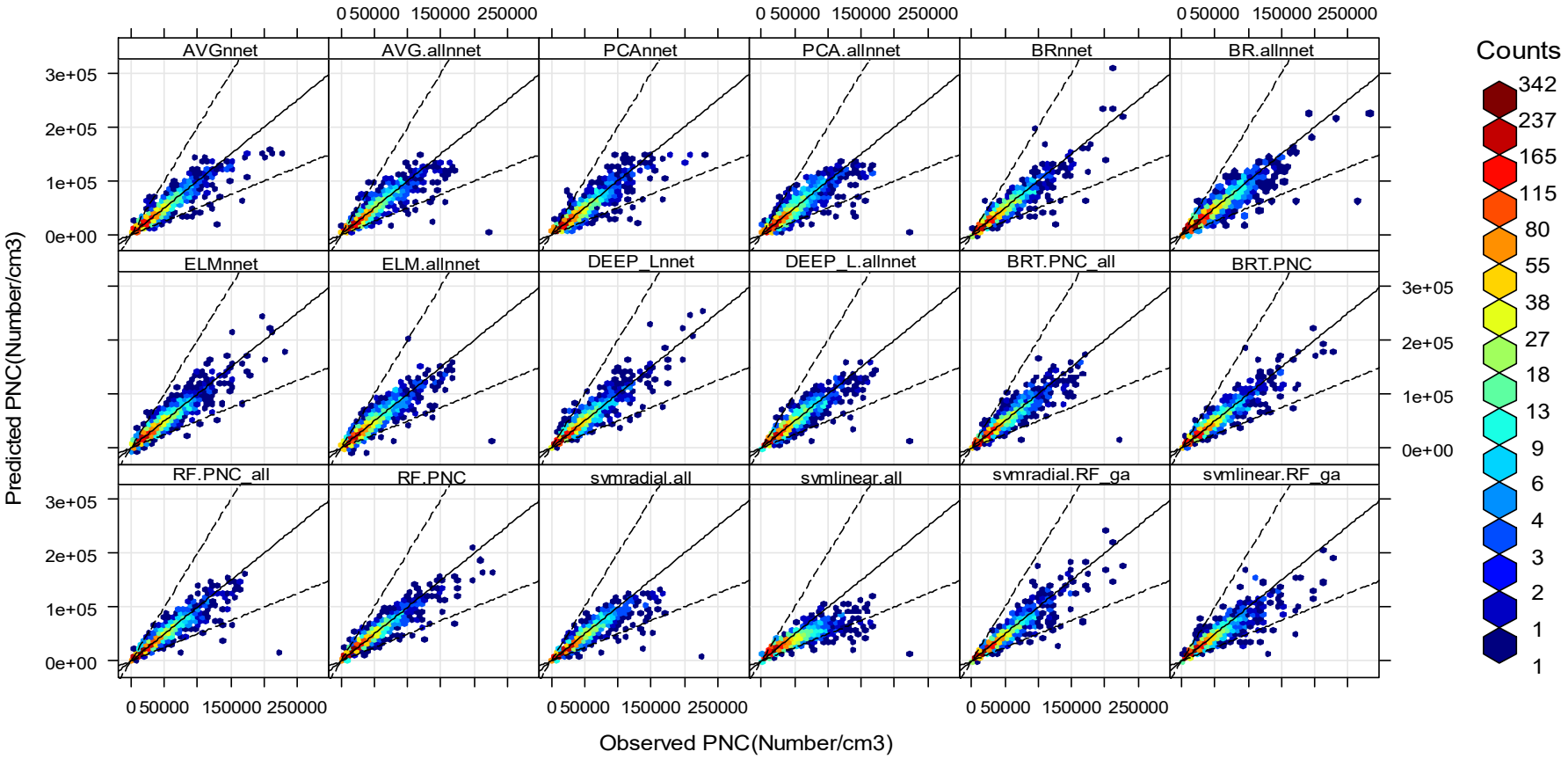


Figure F.5 Scatter plots comparing the prediction of the ML models and the observed PNC concentrations.

Note: modPM10 and obsPM10 are modelled and observed PNC concentrations respectively

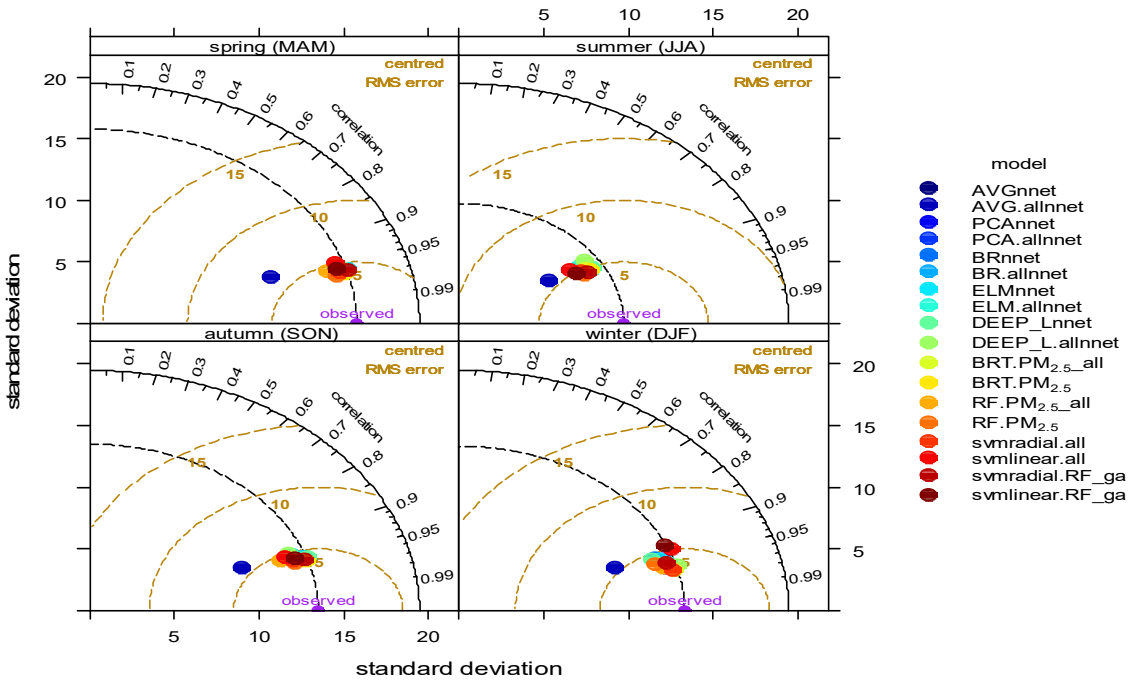


Figure F.6 Taylor's plot comparing the performance of Machine learning models for predicting $PM_{2.5}$

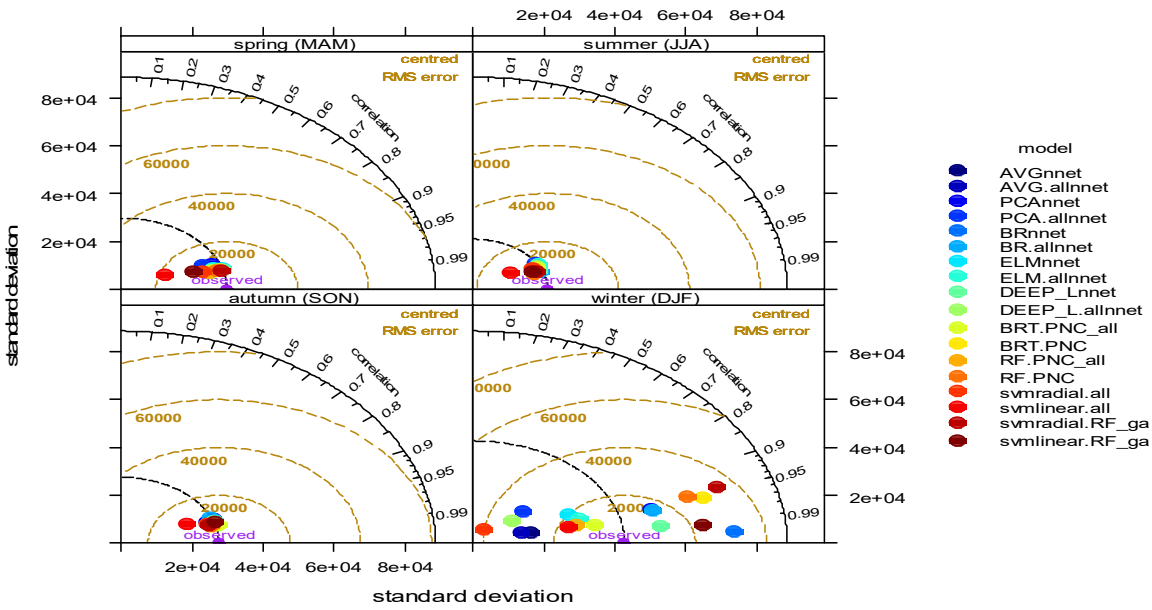


Figure F.7 Taylor's plot comparing the performance of Machine learning models for predicting PNC

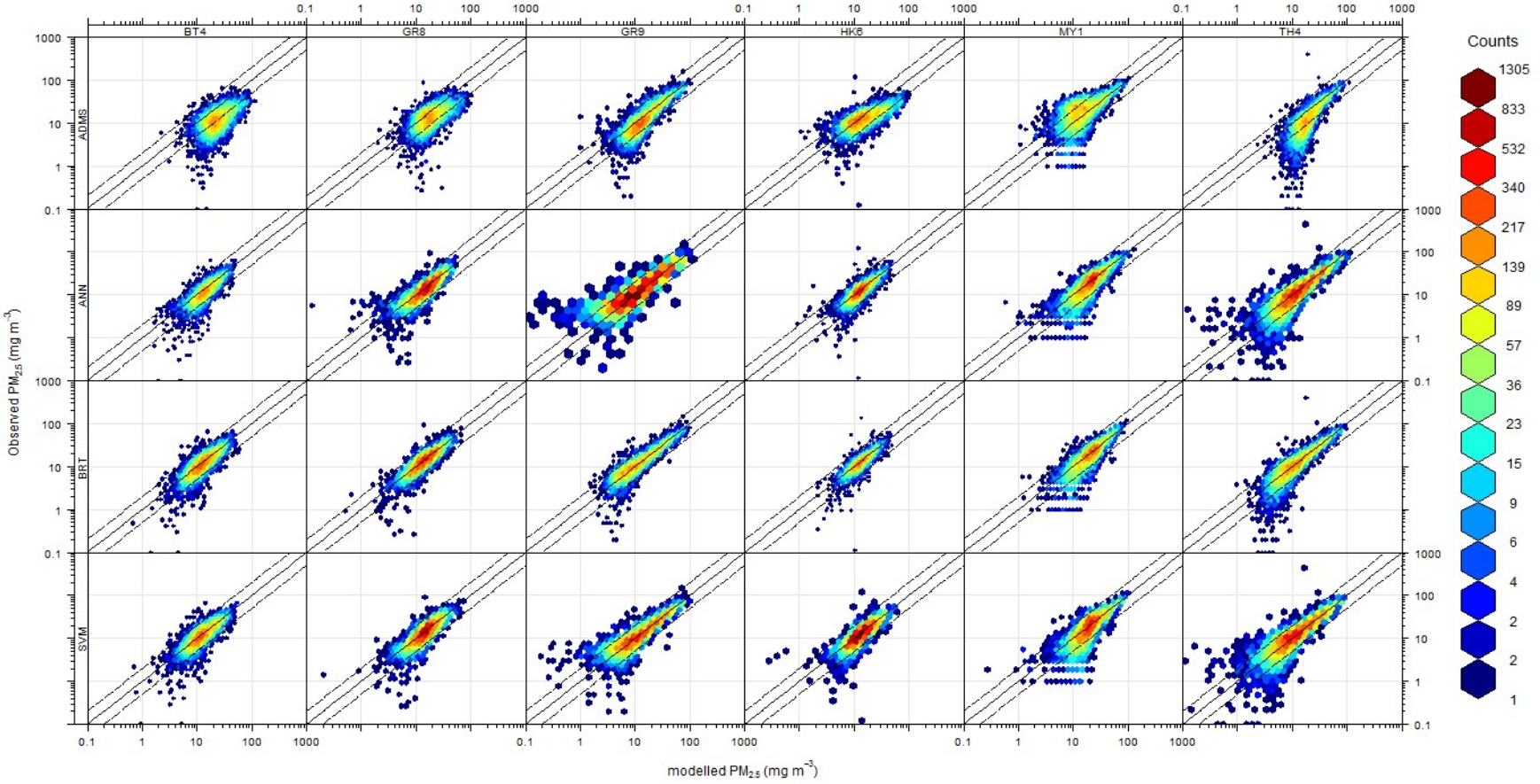


Figure F.8 Scatter plots showing the correlation between the predicted and observed $PM_{2.5}$ concentrations at 10 London sites

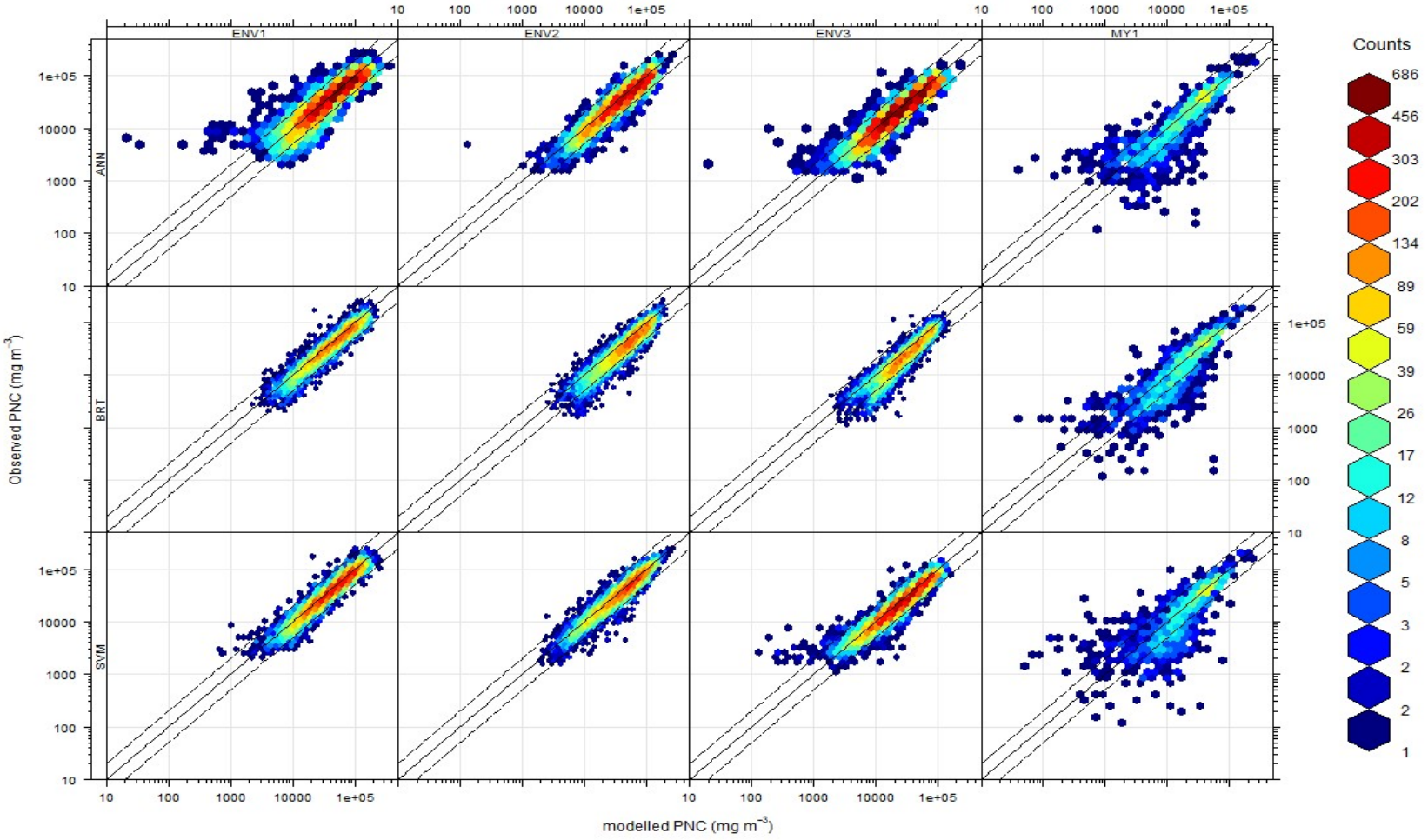


Figure F.9 Scatter plots showing the correlation between the predicted and observed PNC concentrations at 10 London sites

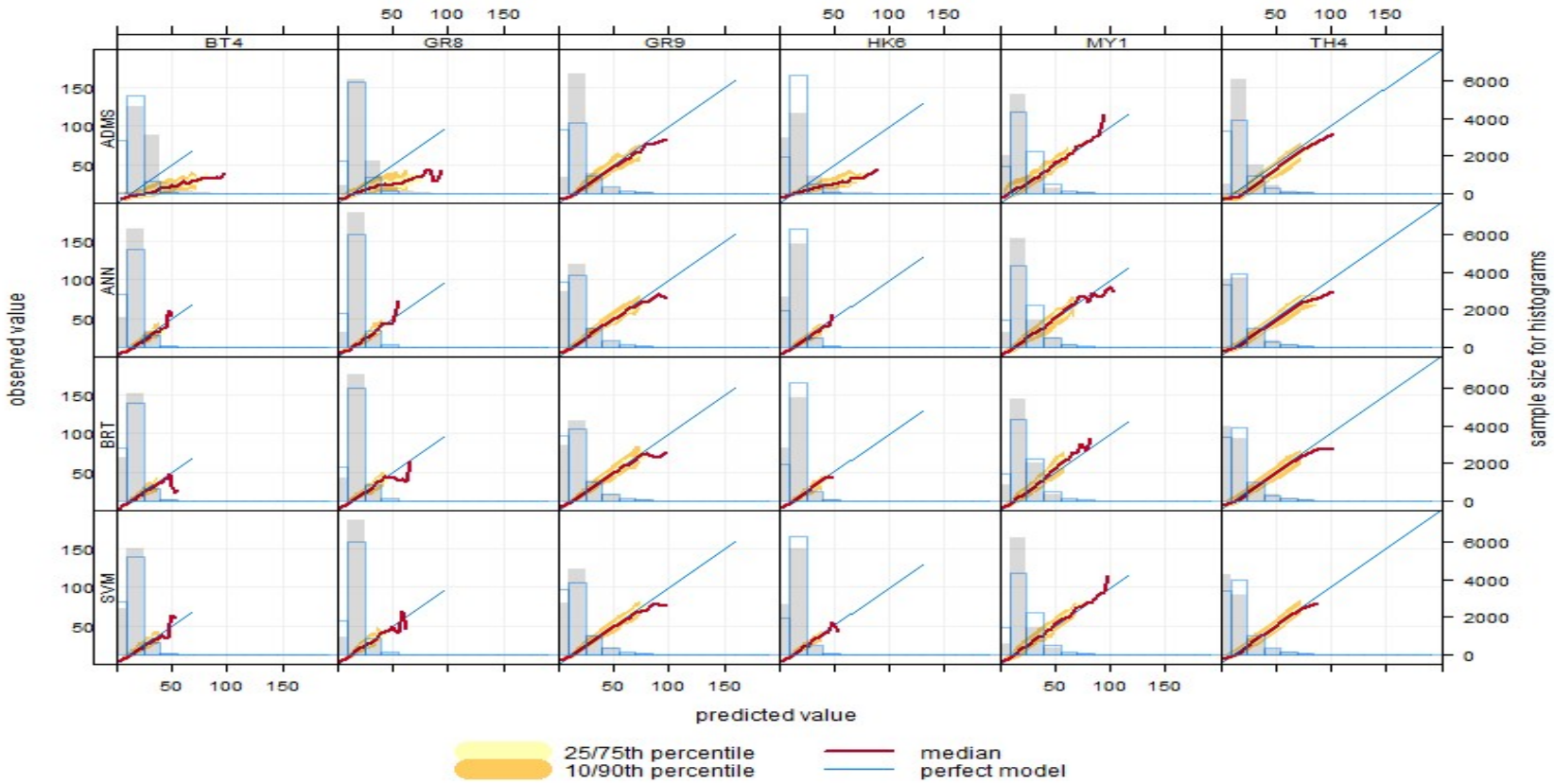


Figure F.10 Conditional quantile plots showing the prediction performance of the models at 6 $PM_{2.5}$ London monitoring sites.

Note: predicted value and observed value are modelled and observed $PM_{2.5}$ concentrations respectively

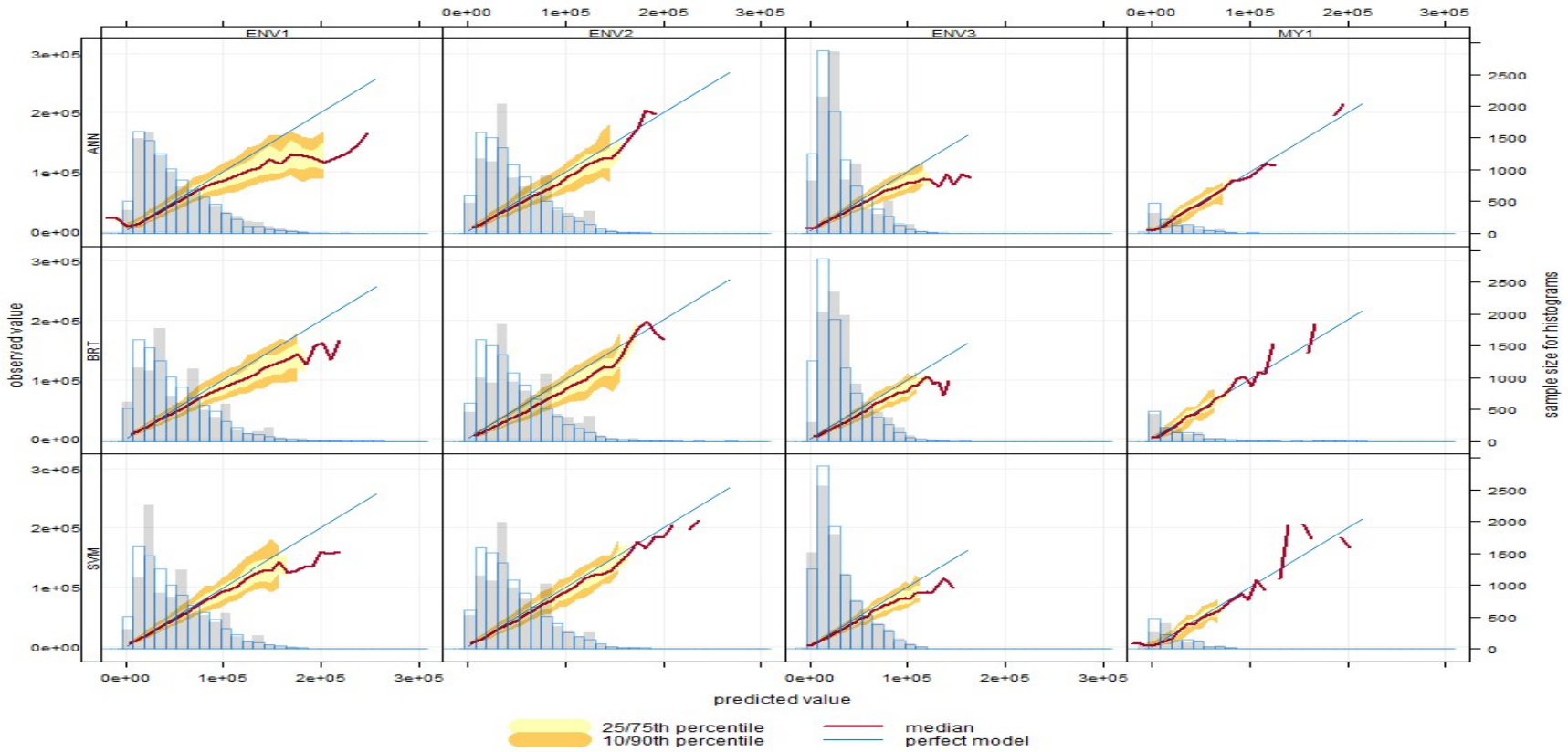


Figure F.11 Conditional quantile plots showing the prediction performance of the models at 6 PNC London monitoring sites.

Note: predicted value and observed value are modelled and observed PNC concentrations respectively

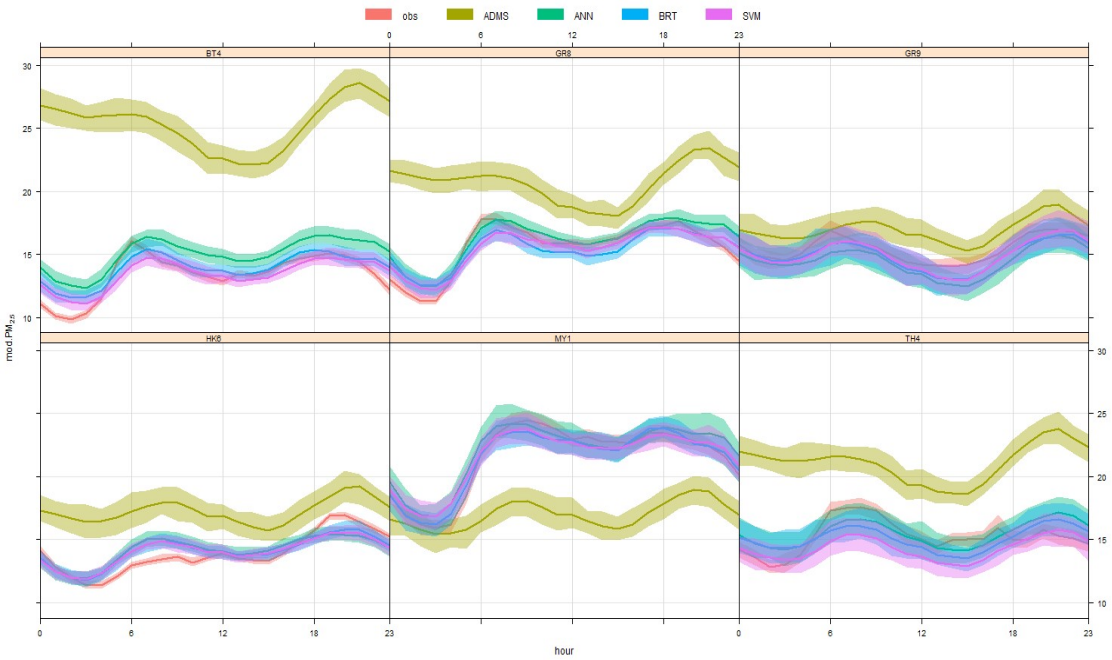


Figure F.12 Hourly time variation plots of the observed and predicted $PM_{2.5}$ concentrations

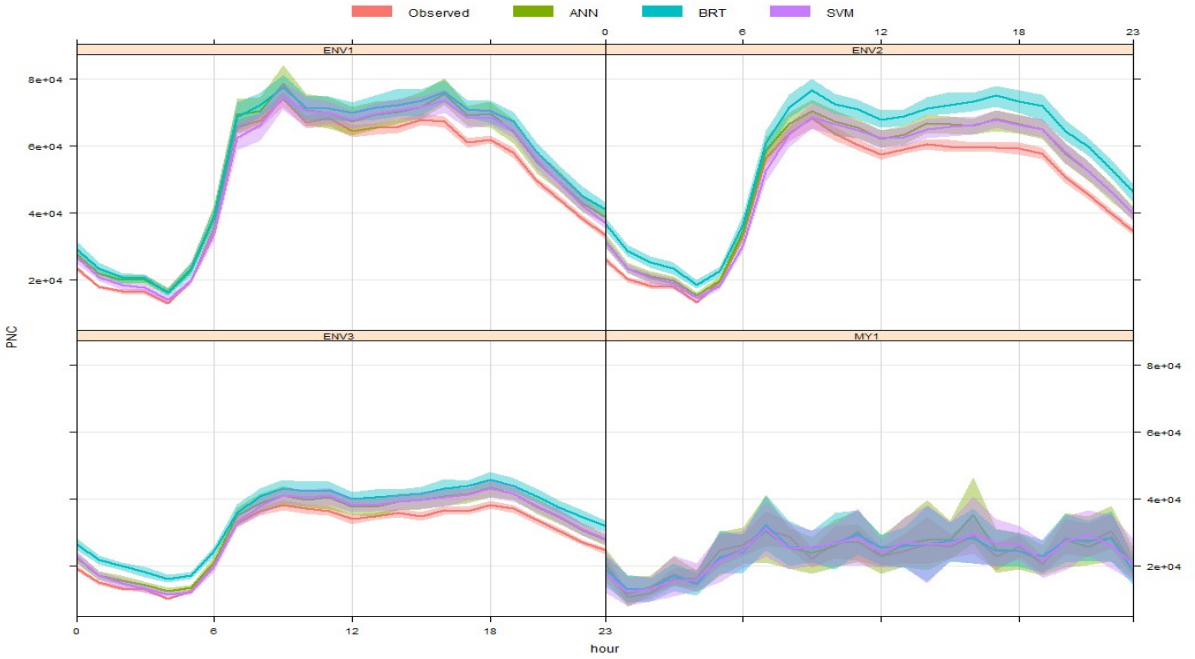


Figure F.13 Hourly time variation plots of the observed and predicted PNC concentrations

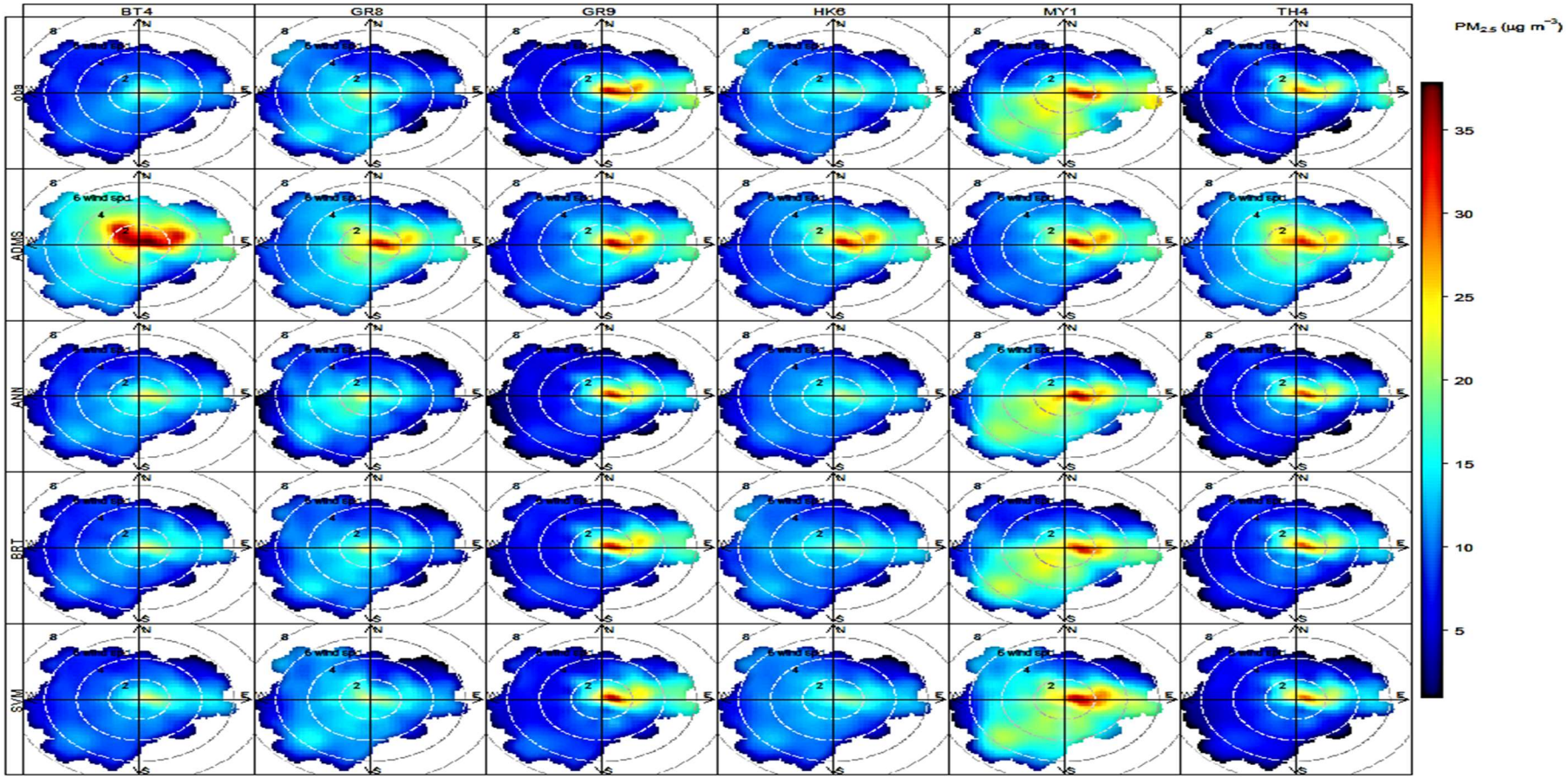


Figure F.14 Bivariate Polar plots showing the variation of the $PM_{2.5}$ concentrations with wind speeds and wind directions in the model predictions and the observations.

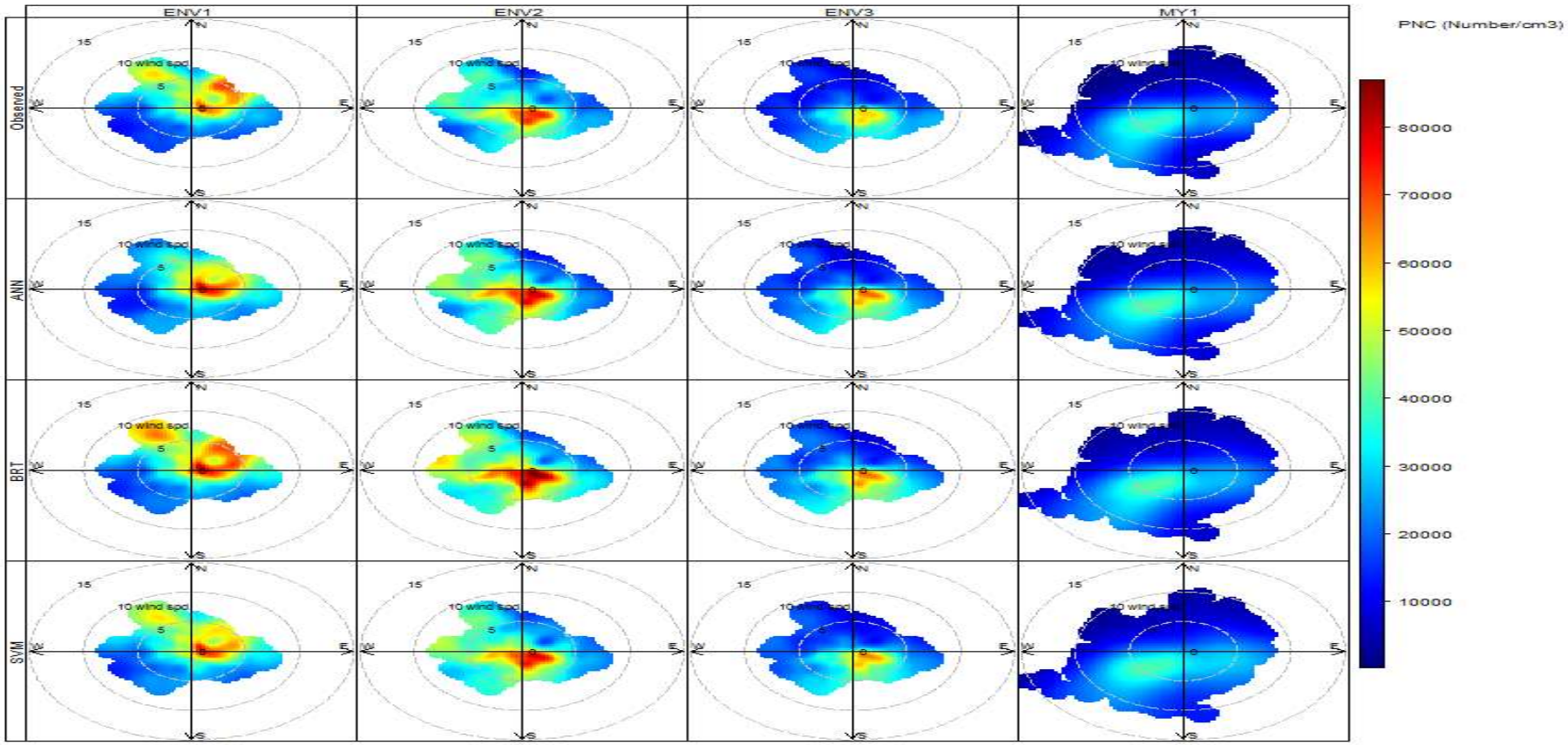


Figure F.15 Bivariate Polar plots showing the variation of the PNC concentrations with wind speeds and wind directions in the model predictions and the observations.

Appendix G Performance of the Models in Predicting Air Quality Statistics

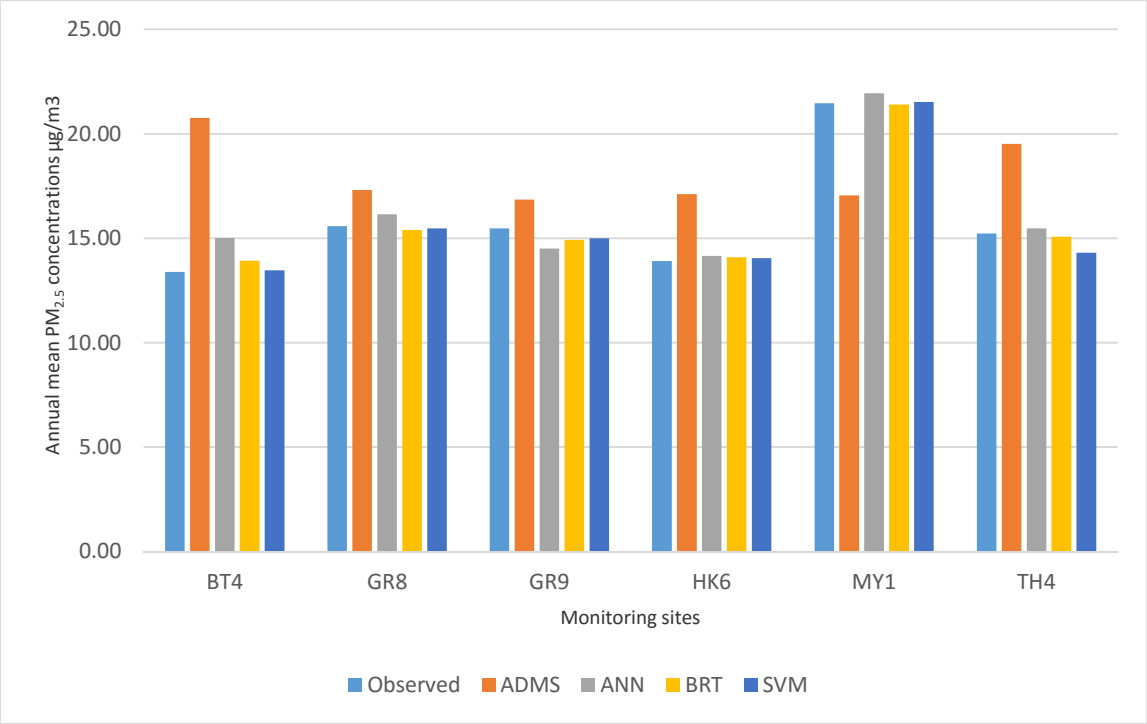


Figure G.1 Predicted and observed annual mean $PM_{2.5}$ concentrations

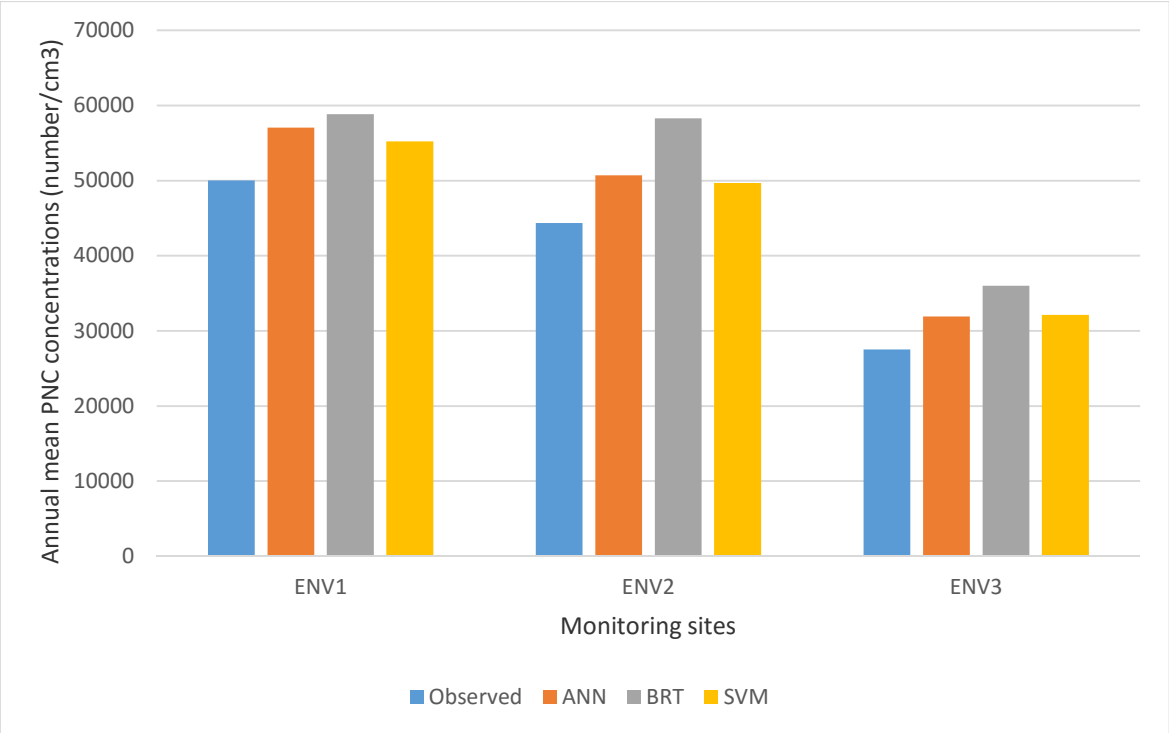
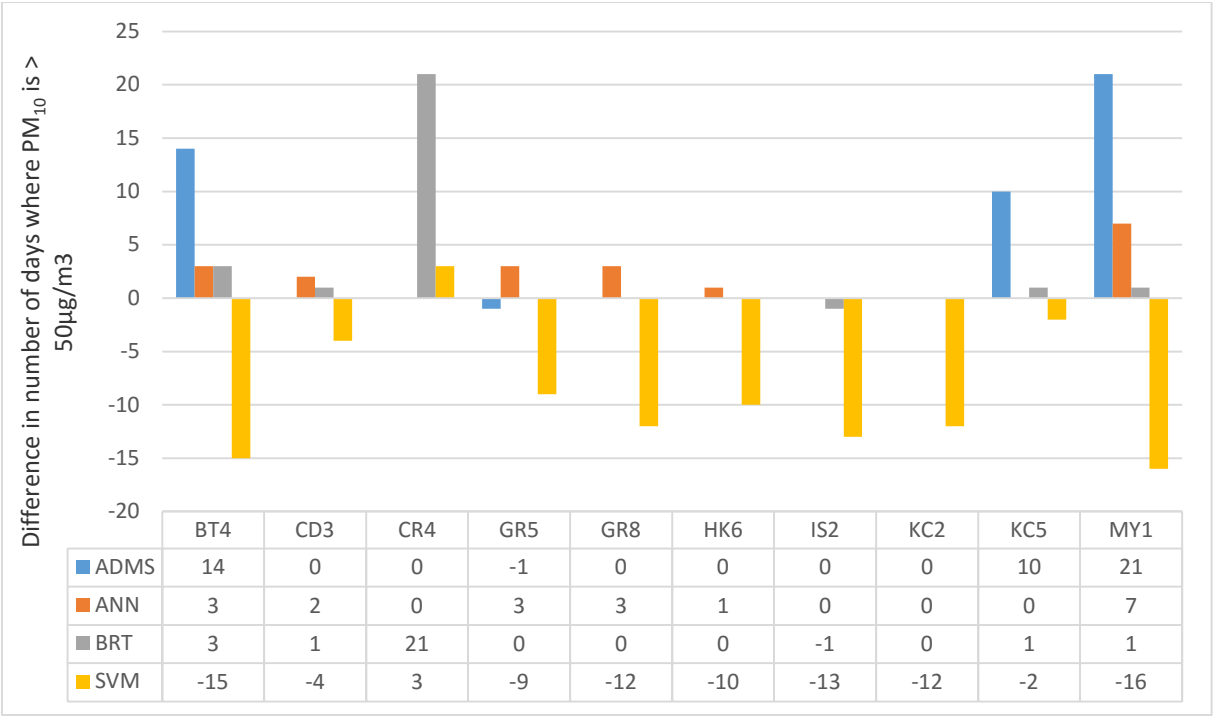
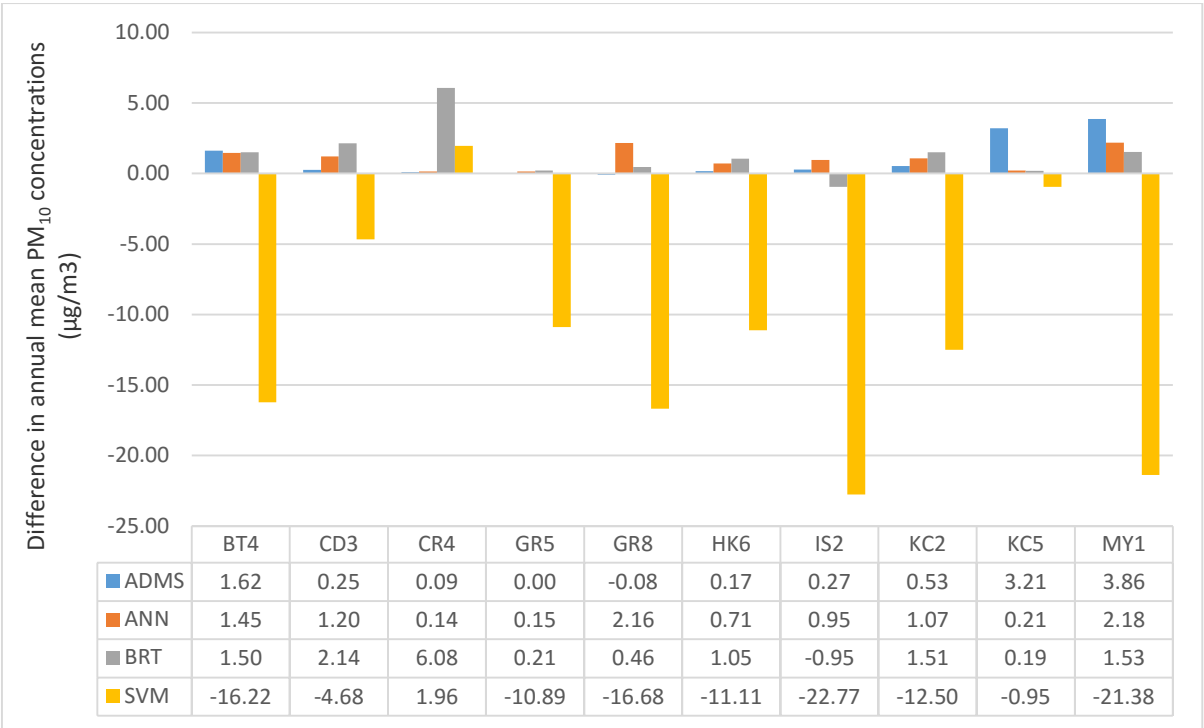


Figure G.2 Predicted and observed annual mean PNC concentrations



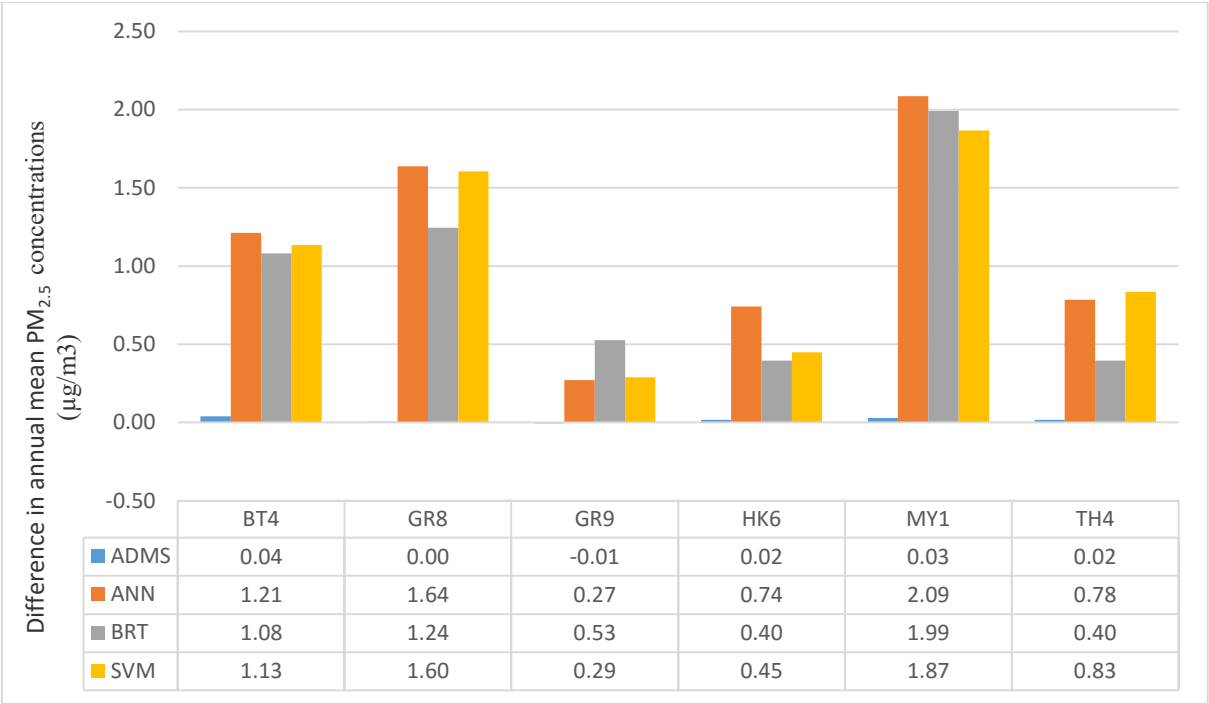
	BT4	CD3	CR4	GR5	GR8	HK6	IS2	KC2	KC5	MY1
ADMS	14	0	0	-1	0	0	0	0	10	21
ANN	3	2	0	3	3	1	0	0	0	7
BRT	3	1	21	0	0	0	-1	0	1	1
SVM	-15	-4	3	-9	-12	-10	-13	-12	-2	-16

Figure G.3 Predicted effects of Euro4/VI scenario on the days with $PM_{10} > 50\mu g/m^3$ in 2015



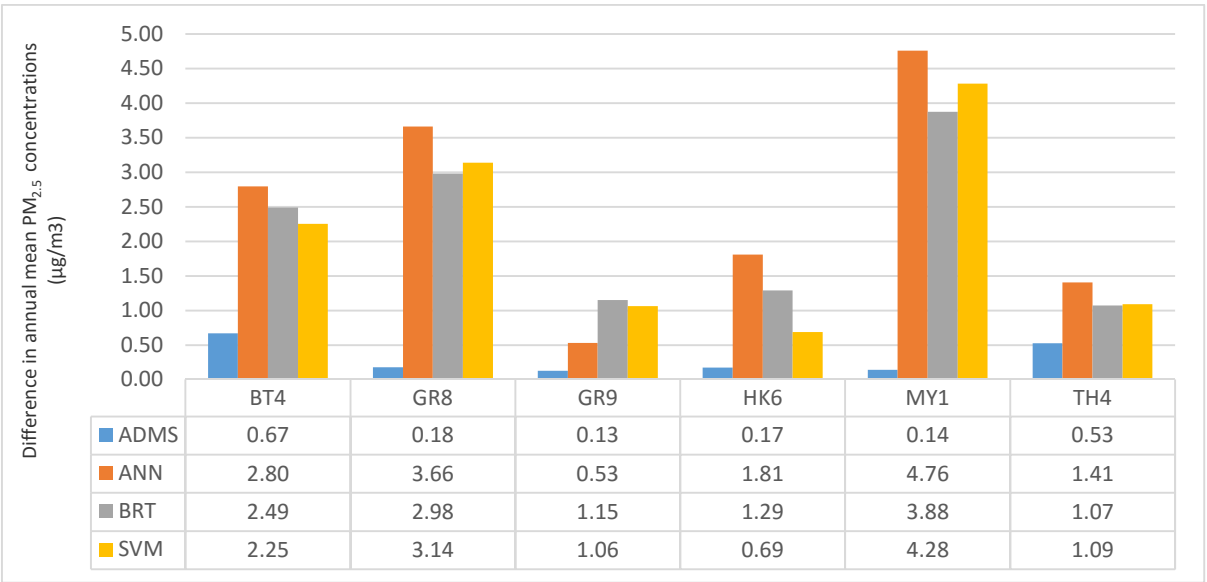
	BT4	CD3	CR4	GR5	GR8	HK6	IS2	KC2	KC5	MY1
ADMS	1.62	0.25	0.09	0.00	-0.08	0.17	0.27	0.53	3.21	3.86
ANN	1.45	1.20	0.14	0.15	2.16	0.71	0.95	1.07	0.21	2.18
BRT	1.50	2.14	6.08	0.21	0.46	1.05	-0.95	1.51	0.19	1.53
SVM	-16.22	-4.68	1.96	-10.89	-16.68	-11.11	-22.77	-12.50	-0.95	-21.38

Figure G.4 Predicted effects of Euro4/VI scenario on the annual mean PM_{10} concentrations in 2015



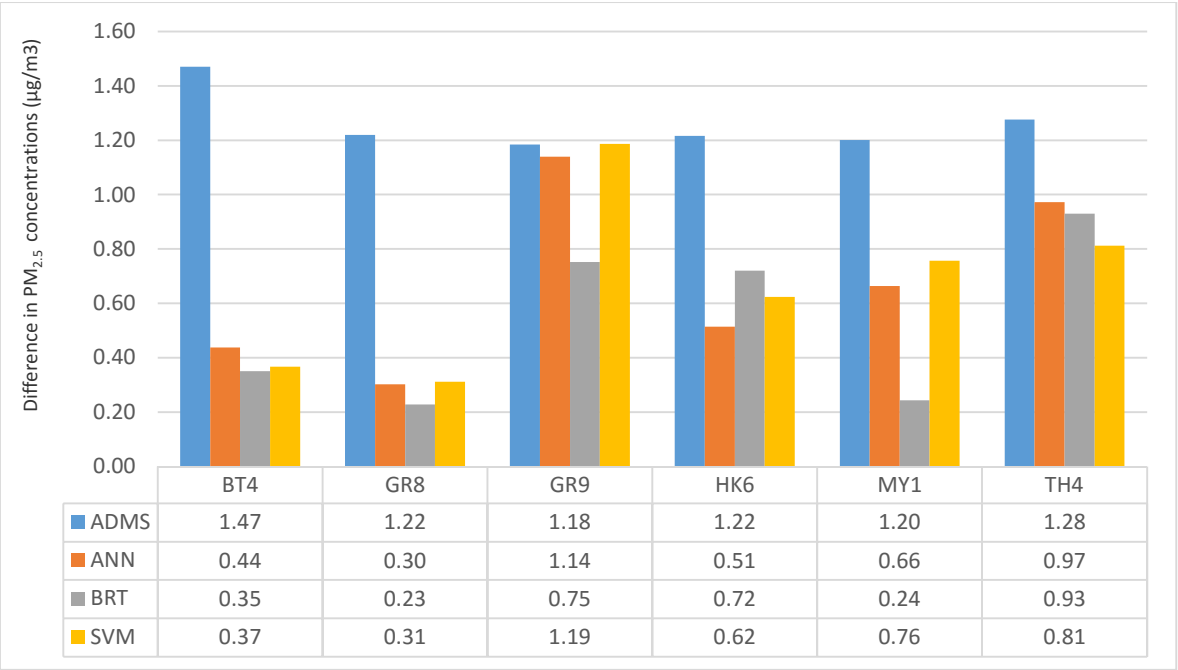
	BT4	GR8	GR9	HK6	MY1	TH4
ADMS	0.04	0.00	-0.01	0.02	0.03	0.02
ANN	1.21	1.64	0.27	0.74	2.09	0.78
BRT	1.08	1.24	0.53	0.40	1.99	0.40
SVM	1.13	1.60	0.29	0.45	1.87	0.83

Figure G.5 Predicted effects of Euro4/VI scenario on the annual mean $PM_{2.5}$ concentrations in 2012



	BT4	GR8	GR9	HK6	MY1	TH4
ADMS	0.67	0.18	0.13	0.17	0.14	0.53
ANN	2.80	3.66	0.53	1.81	4.76	1.41
BRT	2.49	2.98	1.15	1.29	3.88	1.07
SVM	2.25	3.14	1.06	0.69	4.28	1.09

Figure G.6 Predicted effects of Euro4/VI scenario on the annual mean $PM_{2.5}$ concentrations in 2015



	BT4	GR8	GR9	HK6	MY1	TH4
ADMS	1.47	1.22	1.18	1.22	1.20	1.28
ANN	0.44	0.30	1.14	0.51	0.66	0.97
BRT	0.35	0.23	0.75	0.72	0.24	0.93
SVM	0.37	0.31	1.19	0.62	0.76	0.81

Figure G.7 Predicted change in the annual mean $PM_{2.5}$ concentrations from 2012 to 2015

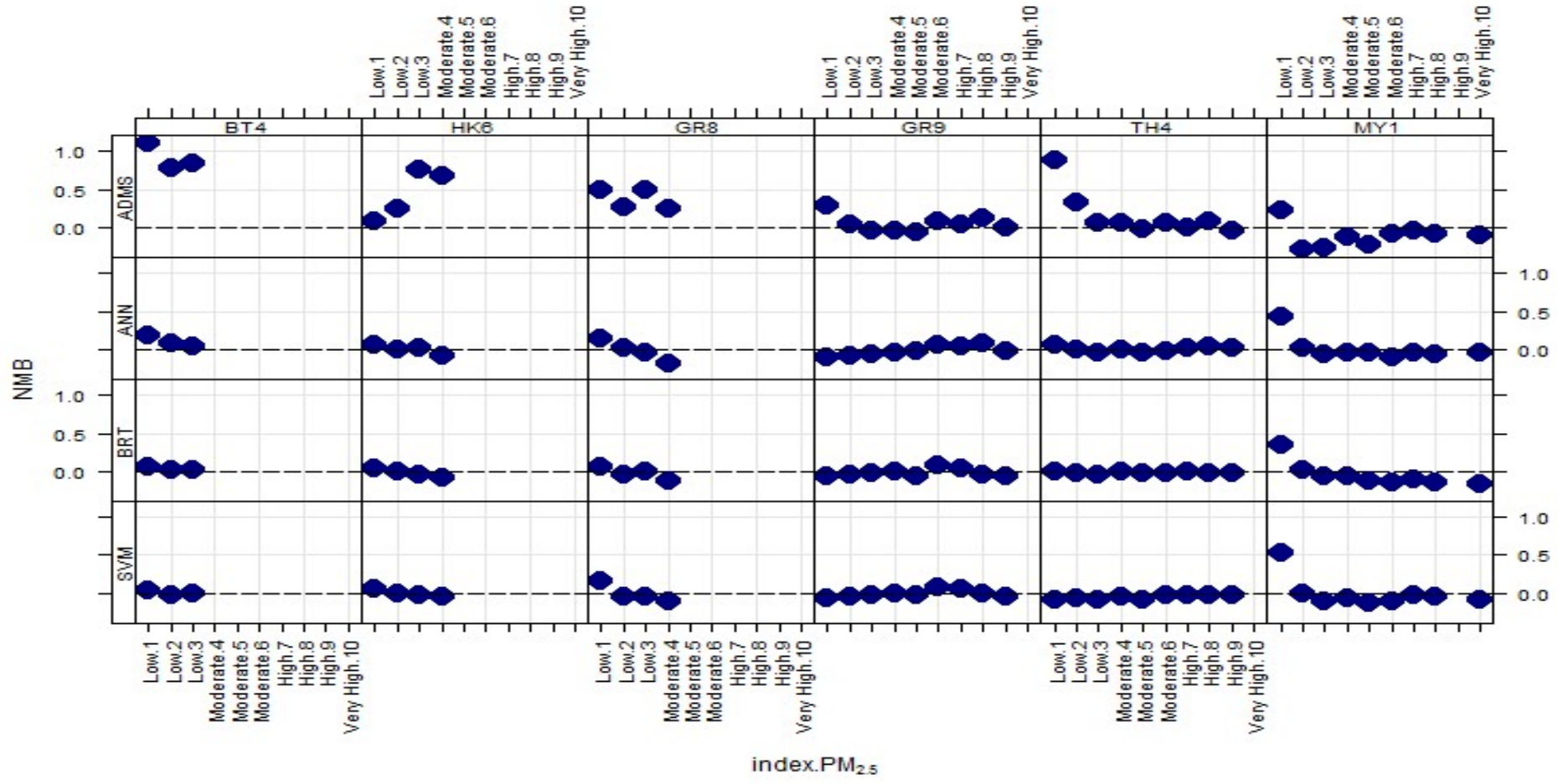


Figure G.8 Graphical comparison of model performance (normalised mean bias) against daily air quality index for PM2.5

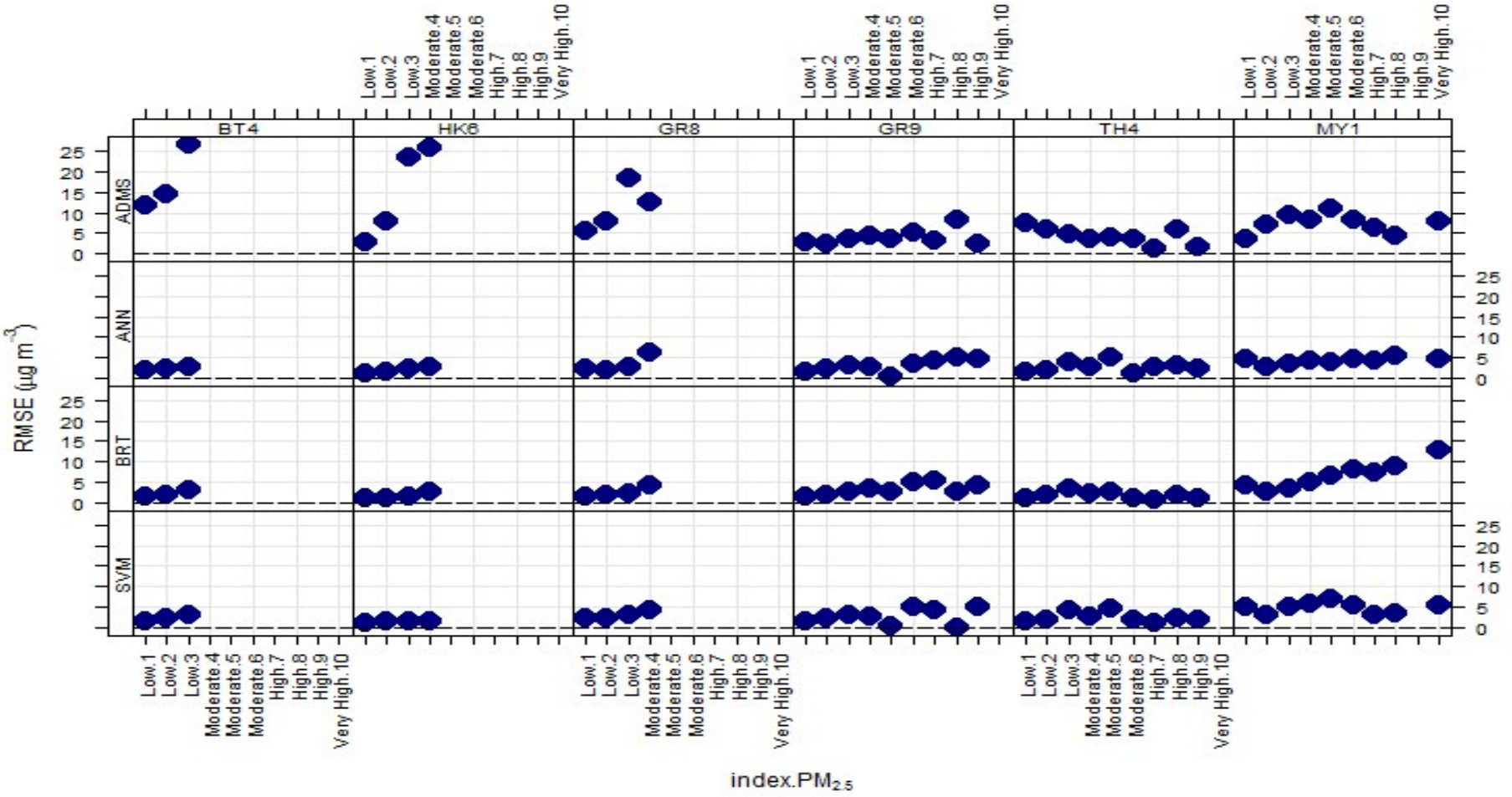


Figure G.9 Graphical comparison of model performance (RMSE) against daily air quality index for $PM_{2.5}$

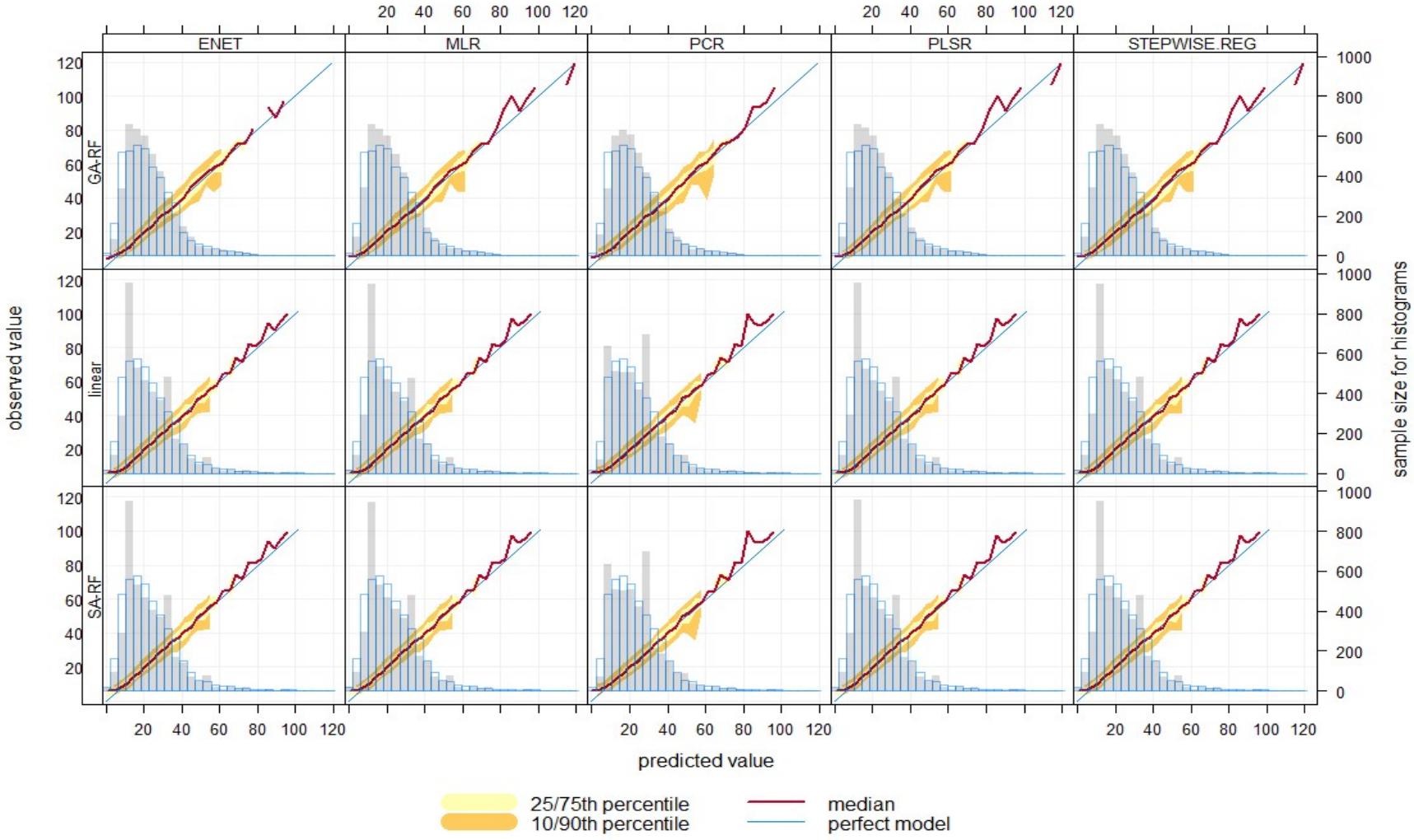


Figure G.10 Conditional Quantile plots comparing the performance of $PM_{2.5}$ models

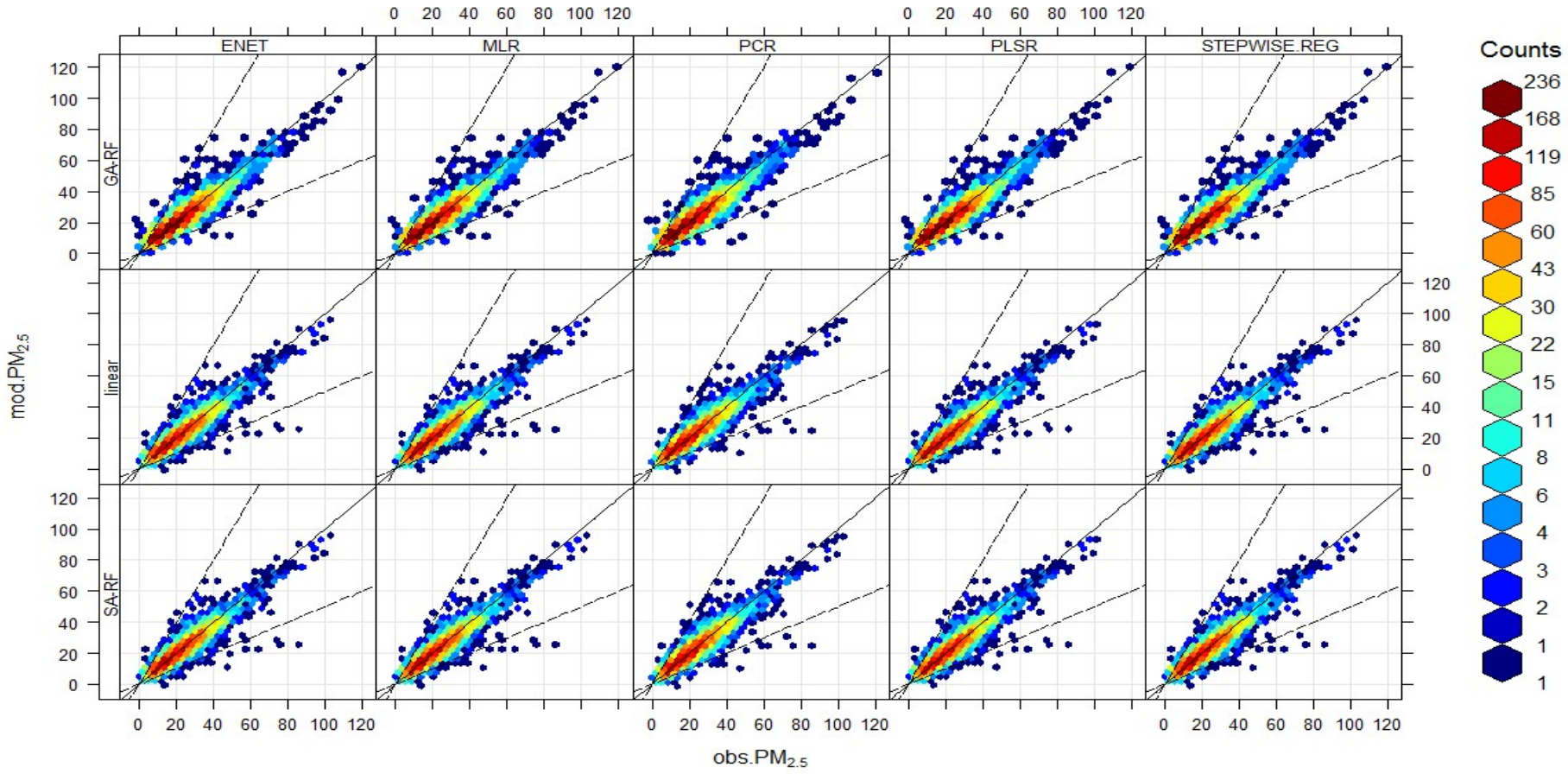


Figure G.11 Scatter plots comparing the performance of $PM_{2.5}$ models

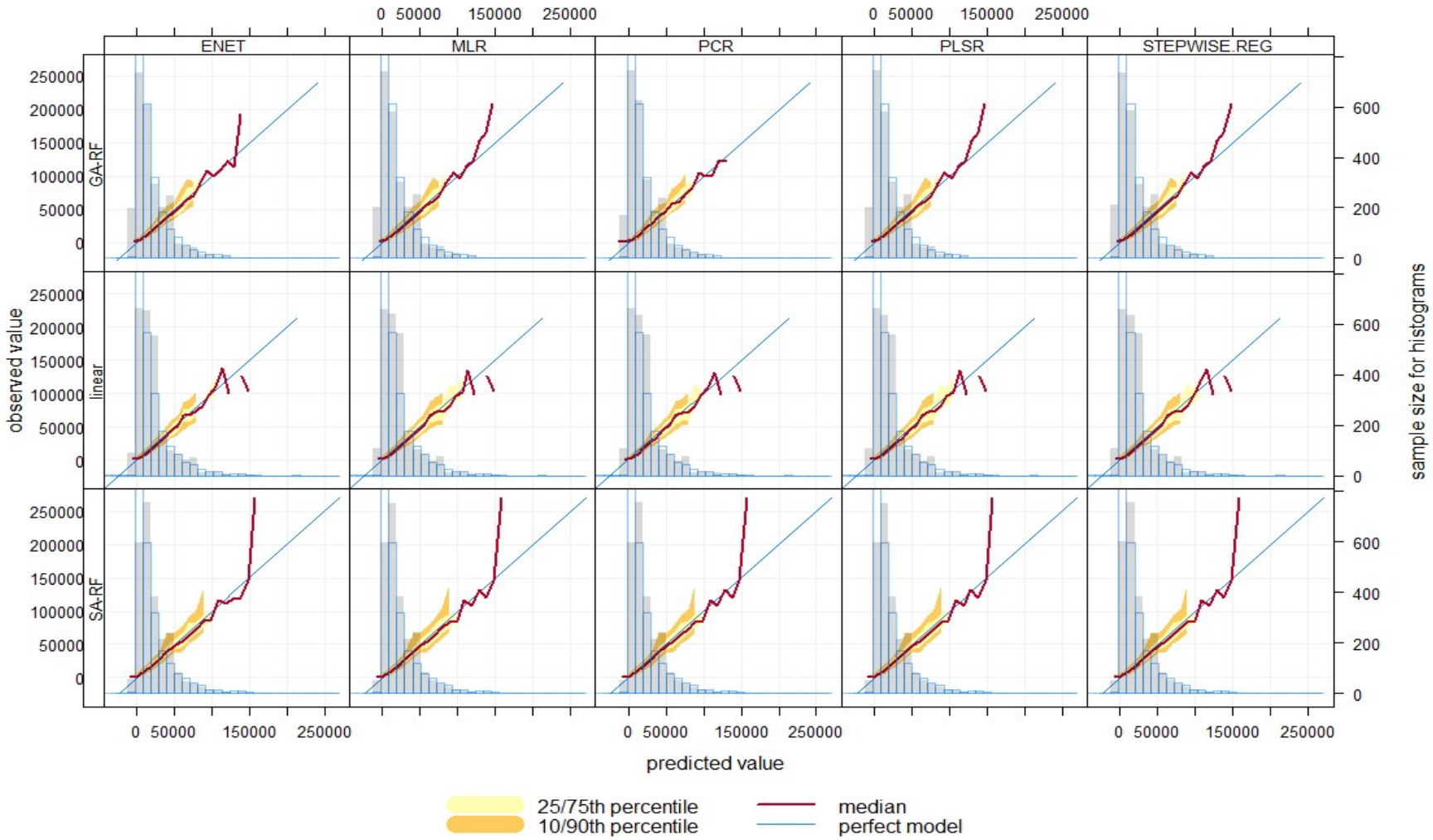


Figure G.12 conditional quantile plots comparing the performance of PNC models

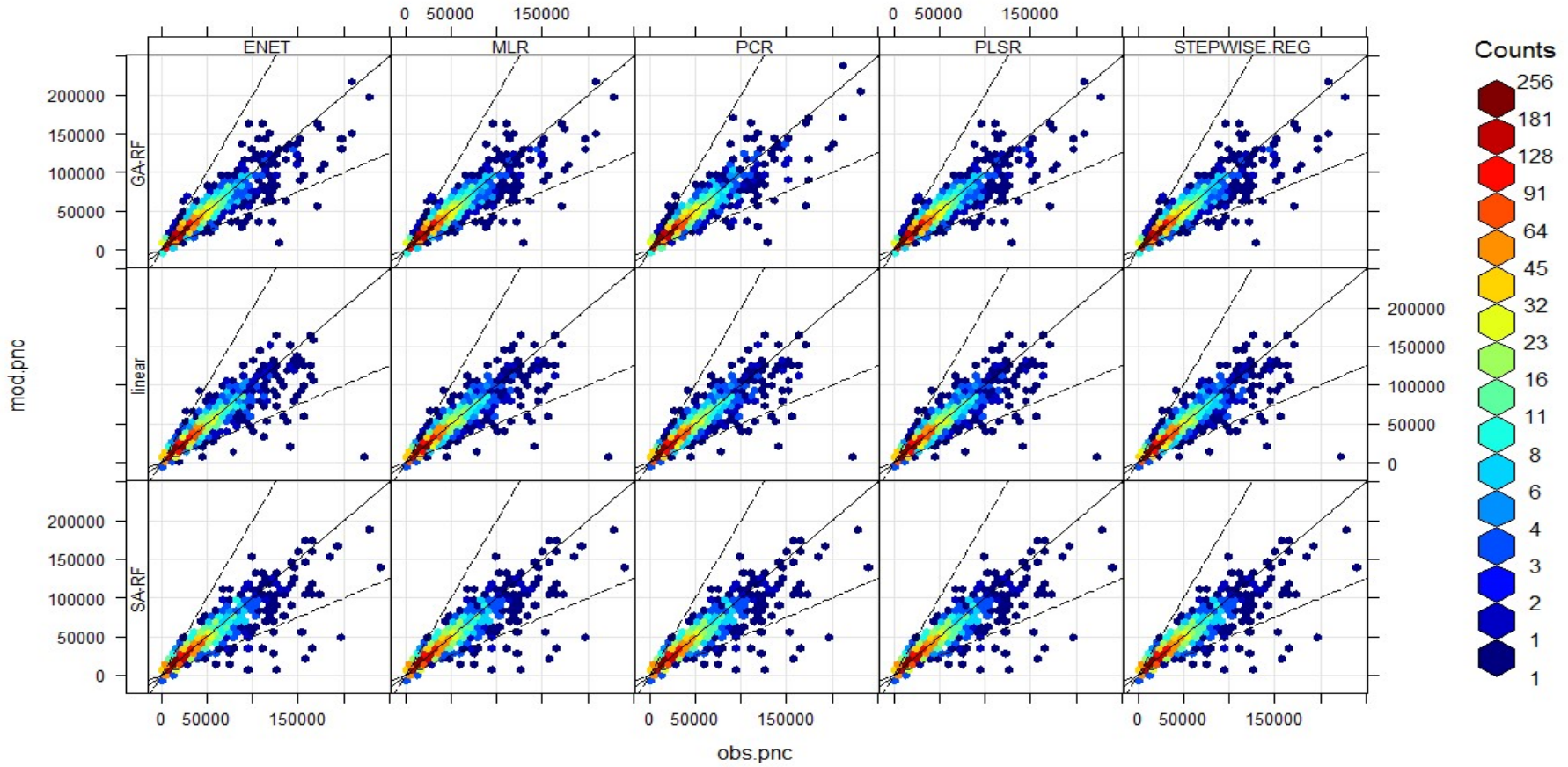


Figure G.13 Scatter plots comparing the performance of PNC model

Appendix H Statistical Performance of the Models in Spatiotemporal Predictions

Table H.1 Statistical performance of the models in spatiotemporal predictions (PM_{10})

Row Labels	FAC2	NMB	NMGE	RMSE	R	COE	IOA
BT4							
ADMS	0.89	0.12	0.35	16.58	0.70	0.18	0.59
ANN	0.98	-0.03	0.20	10.04	0.87	0.54	0.77
BRT	0.98	-0.02	0.19	9.54	0.88	0.56	0.78
SVM	0.96	-0.17	0.22	11.72	0.87	0.48	0.74
CD3							
ADMS	0.81	-0.24	0.34	14.99	0.71	0.10	0.55
ANN	0.98	-0.02	0.17	8.95	0.84	0.56	0.78
BRT	0.98	0.01	0.18	9.29	0.83	0.53	0.77
SVM	0.98	-0.06	0.16	9.11	0.84	0.58	0.79
CR4							
ADMS	0.80	-0.05	0.42	20.38	0.42	0.08	0.54
ANN	0.84	0.01	0.38	19.17	0.45	0.17	0.58
BRT	0.82	0.15	0.44	20.91	0.43	0.03	0.52
SVM	0.84	-0.06	0.37	19.17	0.43	0.17	0.59
GR5							
ADMS	0.93	0.00	0.25	9.16	0.84	0.42	0.71
ANN	0.99	-0.01	0.13	4.69	0.95	0.71	0.86
BRT	1.00	0.01	0.12	4.48	0.95	0.73	0.86
SVM	0.99	-0.08	0.13	4.91	0.95	0.70	0.85
GR8							
ADMS	0.84	-0.22	0.34	19.19	0.55	0.12	0.56
ANN	0.97	0.07	0.23	14.12	0.72	0.40	0.70
BRT	0.98	0.02	0.21	14.02	0.72	0.46	0.73
SVM	0.90	-0.26	0.29	17.06	0.69	0.25	0.63
HK6							
ADMS	0.82	-0.24	0.33	14.23	0.74	0.11	0.56
ANN	0.99	-0.07	0.16	8.50	0.87	0.58	0.79
BRT	0.99	-0.03	0.14	7.89	0.88	0.62	0.81
SVM	0.99	-0.09	0.16	8.65	0.87	0.57	0.78

Table H.1 *continued*

Row Labels	FAC2	NMB	NMGE	RMSE	R	COE	IOA
IS2							
ADMS	0.88	-0.17	0.30	13.92	0.68	0.18	0.59
ANN	0.99	-0.01	0.14	8.54	0.84	0.62	0.81
BRT	1.00	-0.01	0.13	8.37	0.85	0.64	0.82
SVM	0.96	-0.23	0.25	11.25	0.83	0.33	0.67
KC2							
ADMS	0.84	-0.18	0.32	12.62	0.75	0.14	0.57
ANN	0.99	0.01	0.17	7.00	0.88	0.54	0.77
BRT	0.99	0.01	0.16	6.74	0.89	0.57	0.78
SVM	0.98	-0.11	0.19	7.87	0.87	0.50	0.75
KC5							
ADMS	0.95	0.14	0.28	13.33	0.76	0.14	0.57
ANN	0.98	0.11	0.23	9.58	0.80	0.31	0.65
BRT	0.98	0.06	0.21	9.35	0.80	0.35	0.68
SVM	0.98	-0.04	0.20	9.41	0.79	0.39	0.70
MY1							
ADMS	0.85	-0.10	0.35	18.64	0.56	0.02	0.51
ANN	0.98	-0.06	0.17	10.61	0.84	0.52	0.76
BRT	0.99	-0.03	0.16	9.91	0.85	0.56	0.78
SVM	0.94	-0.24	0.27	15.24	0.79	0.25	0.63

Table H.2 Statistical performance of the models in spatiotemporal predictions ($PM_{2.5}$)

Row Labels	FAC2	NMB	NMGE	RMSE	R	COE	IOA
BT4							
ADMS	0.70	0.56	0.63	11.49	0.66	-0.55	0.22
ANN	0.95	0.12	0.26	4.48	0.82	0.37	0.69
BRT	0.97	0.04	0.22	4.16	0.83	0.45	0.73
SVM	0.97	0.01	0.23	4.25	0.81	0.44	0.72
GR8							
ADMS	0.86	0.12	0.42	9.41	0.57	-0.11	0.44
ANN	0.98	0.04	0.19	4.15	0.85	0.49	0.74
BRT	0.99	-0.01	0.18	4.06	0.86	0.51	0.76
SVM	0.97	-0.01	0.21	4.47	0.82	0.45	0.73
GR9							
ADMS	0.92	0.10	0.26	5.66	0.91	0.57	0.79
ANN	0.95	-0.06	0.18	4.21	0.95	0.70	0.85
BRT	0.98	-0.04	0.17	4.06	0.95	0.71	0.86
SVM	0.95	-0.03	0.18	4.08	0.95	0.70	0.85
HK6							
ADMS	0.91	0.24	0.40	8.81	0.74	-0.18	0.41
ANN	0.98	0.02	0.17	3.61	0.83	0.48	0.74
BRT	0.99	0.01	0.16	3.47	0.85	0.51	0.76
SVM	0.98	0.01	0.17	3.50	0.84	0.50	0.75
MY1							
ADMS	0.84	-0.05	0.32	8.67	0.78	0.32	0.66
ANN	0.93	0.02	0.22	6.30	0.88	0.52	0.76
BRT	0.94	0.00	0.22	6.33	0.88	0.52	0.76
SVM	0.91	0.00	0.24	6.74	0.86	0.48	0.74
TH4							
ADMS	0.76	0.29	0.39	8.28	0.85	0.34	0.67
ANN	0.93	0.02	0.20	6.05	0.89	0.66	0.83
BRT	0.94	-0.01	0.19	5.93	0.90	0.68	0.84
SVM	0.93	-0.06	0.19	6.00	0.90	0.68	0.84