URBAN AIR QUALITY MONITORING, MAPPING AND MODELLING TO DETERMINE THE MAIN DRIVERS OF AIR POLLUTION

By

Said Munir

A thesis submitted in partial fulfilment of the requirements for the Degree of
Doctor of Philosophy

Department of Civil and Structural Engineering, Faculty of Engineering, University of Sheffield.

November 2022

To
My family for their unconditional love and support

ABSTRACT

Air pollution is a growing concern for human health, biodiversity and natural environment in large urban areas. It is, therefore, vital to monitor and model air quality (AQ) in urban areas to understand its spatiotemporal variabilities and its main drivers. Traditionally it was not possible to develop high-resolution AQ maps in urban areas due to sparse reference network. However, since the emergence of low-cost sensors (LCS), it has become possible to structure a dense network of sensors and develop high-resolution AQ maps. This is what this PhD project intends to achieve by: (a) analysing the suitability of LCS for urban AQ monitoring and how their measurements can be further improvement using advance calibration techniques, (b) deploying a dense network of AQ sensors based on multiple criteria and using sensors of different grades, (c) employing various AQ modelling and mapping techniques including geostatistical interpolations, land-use regression (LUR) and dispersion modelling, and (d) using data fusion approaches to fuse measured and estimated pollutant concentrations. A multi-criteria Air Quality Monitoring Network (AQMN) was structured based on economic, social and environmental indicators. The network was made of several layers of sensors including reference sensors, LCS (e.g., AQMesh pods and Envirowatch E-MOTEs) and IoT (internet of things) sensors. The data from the designed AQMN was used in AQ mapping, models validations, analysing spatiotemporal variability of pollutants and sensor calibration. Reference sensors were used as standard to calibrate measurements of the LCS employing multiple linear regression and generalised additive model. LUR models were developed for the first time in Sheffield using several land-use and emission related variables. In contrast to previous studies that mostly used linear techniques, here nonlinear regression approaches were also used for developing LUR models, which outperformed the linear counterparts. LUR models were trained and validated using annual average NO_2 concentrations from diffusion tubes as well as from LCS. The models were cross-validated by comparing estimated and measured concentrations. LUR model demonstrated that among predictor variables altitude had negative significant effect, whereas major roads, minor roads and commercial areas had positive significant effect on NO_2 concentrations. Furthermore, an Airviro dispersion model was developed and several emission scenarios were tested, which showed that NOx concentrations were mainly controlled by road traffic, whereas PM_{10} concentrations were controlled by point sources. To further improve the AQ maps, modelled and measured concentrations were fused (integrated) to produce high-resolution maps in Sheffield using data fusion technique known as Universal Kriging, which estimated realistic (based on priori expectations) NO_2 concentration maps that inherited spatial patterns of the pollutant from the model estimations and adjusted the modelled values against the measured concentrations. The methodology was successful in demonstrating the spatial variability and highlighting the hotspots of NO_2 concentrations in Sheffield. The main findings of the project are: (a) The project proposed a nonlinear generalised additive model for low-cost sensors calibrations in outdoor environment. Low-cost sensors are a cheaper source of AQ data, however, they require robust outfield calibrations. (b) A formal approach was proposed for structuring an AQMN in urban areas, which was based on multi-criteria. (c) It was shown that LUR model based on nonlinear machine learning approach outperformed the dispersion modelling approach. (d) Data fusion techniques (such as Universal krigging) were employed to integrate model estimations with measured concentrations. Such data fusion approaches are useful tools for improving data quality and producing high-resolution AQ maps. (e) Time series modelling ARIMA with exogenous variables (ARIMAX) outperformed other linear and nonlinear time series models, and is proposed as an early warning tool for predicting potential pollution episodes in order to be proactive in adopting precautionary measures. Limited data was available on particulate matter, especially on fine and ultrafine particulates, therefore, further work is required on particulate matter monitoring, modelling and management in urban areas.

Keywords: air quality modelling, air quality mapping, air quality monitoring network, land-use regression, sensors calibration, Airviro dispersion model, low-cost sensor, data fusion, urban air pollution.

RESEARCH CONTRIBUTION

This PhD thesis is composed in the alternative format ‘publication format thesis’ incorporating a collection of papers that are already published in peer review journals as open access. From the research work conducted as a part of this PhD, the following papers were published:

1. Munir, S., Mayfield, M. 2021. Application of Density Plots and Time Series Modelling to the Analysis of Nitrogen Dioxides Measured by Low-Cost and Reference Sensors in Urban Areas. Nitrogen, 2, 167–195.
2. Munir, S., Mayfield, M., Coca, D., 2021. Understanding Spatial Variability of NO_2 in Urban Areas Using Spatial Modelling and Data Fusion Approaches. Atmosphere, 2021, 12(2), 179.
3. Munir, S., Mayfield, M., Coca, D., Mihaylova, L.S., 2020. A Nonlinear Land-Use Regression Approach for Modelling NO_2 Concentrations in Urban Areas – Using Data from Low-Cost Sensors and Diffusion Tubes. Atmosphere, 2020, 11 (7), 736.
4. Munir, S., Mayfield, M., Coca, D., Mihaylova, L.S., Osammor, O., 2020. Analysis of Air Pollution in Urban Areas with Airviro Dispersion Model - A Case Study in the City of Sheffield, United Kingdom. Atmosphere, 11 (3), 285.
5. Munir, S., Mayfield, M., Coca, D., Jubb, S.A., 2019. Structuring an Integrated Air Quality Monitoring Network in Large Urban Areas – Discussing the Purpose, Criteria and Strategy. Atmospheric Environment: X 2 (2019): 100027.
6. Munir, S., Mayfield, M., Coca, D., Jubb, S.A., Osammor, O., 2018. Analysing the Performance of Low-Cost Air Quality Sensors, Their Drivers, Relative Benefits and Calibration in Cities - A Case Study in Sheffield. Environmental Monitoring and Assessment, 191 (2): 94.

I (Said Munir) confirm that I am the primary contributor to the writing of these jointly authored publications. My contribution amounts at least 90 to 95 % of the total work in each publication, including designing, analysis, writing up, interpretation of the results, visualisation, presenting it in the required format and editing and revision. However, I do acknowledge the contribution of my supervisors Professor Martin Mayfield (first supervisor) and Professor Daniel Coca (second supervisor) in terms of supervision, minor editing and general guidance. Dr. Ogo Osammor helped me in how to use Airviro model and provided NO_2 diffusion tubes data, whereas Steve Jubb provided technical help in sensors purchase and their installation. Professor Lyudmila Mihaylova revised the papers which she is a co-author of and did minor editing of the language.

ACKNOWLEDGEMENTS

I would like to express my greatest thanks to my supervisors Professor Martin Mayfield (First Supervisor) and Professor Daniel Coca (Second Supervisor) for their help and guidance, without their support I would not have been able to produce this thesis.

I wish to acknowledge EPSRC (grant number - EP/R512175/1) and Siemens plc for funding this PhD project. All statistical data analysis in this project was carried out using the Statistical Software R programming language and several of its packages, therefore I am thankful to the R Development Core Team and to the authors of different R packages used in this project, especially openair, mgcv, MASS, and automap.

Last but not least, I am thankful to my family for their unconditional love and support not only during this PhD, but always, and I would have not made it even to the start line of writing this thesis without their help. My wife Meher, my son Talha and daughter Ukashah probably suffered the most during my PhD as I was busy in my study and could not give them the time they deserved. Thank you very much.

Said Munir
Sheffield, UK
November 2022

Table of Contents

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1. Motivation of the project

Air pollution has become a growing issue for public health especially in urban areas. Air pollution is causing numerous human health and environmental problems. Polluted air, especially with the high levels of NO_2 and particulate matter (PM_{10} and $PM_{2.5}$), is considered the most serious environmental risk to public health in urban areas in the UK (Manisalidis et al., 2020; DEFRA, 2015). Air pollutants penetrate the respiratory system via inhalation and cause respiratory diseases, cardiovascular diseases, reproductive and central nervous system dysfunctions, and cancer (Manisalidis et al., 2020). It is reported that atmospheric pollution remains responsible for approximately 9 million deaths per year, corresponding to one in six deaths worldwide (Fuller et al., 2022), showing the seriousness of the air pollution issue. Air pollution mainly affects those living in large urban areas, where road emissions contribute the most to the degradation of air quality (Manisalidis et al., 2020). Therefore, it is important to understand emission sources, spatiotemporal variability and main drivers of air pollution employing various monitoring and modelling tools. Air pollution is mainly considered a serious issue in urban environments due to the following reasons:

- urban environments experience higher volumes of road traffic,
- Urban environments have more point sources including residential, commercial and industrial emissions,
- In urban areas narrow roads are surrounded by tall buildings which hinder the dispersion of locally emitted air pollutants, and
- Urban areas are densely populated, which expose more people to air pollution.

Air pollutant levels demonstrate spatial heterogeneity everywhere, but especially more so in urban areas because of spatial variability in emission sources and dense and tall buildings, which hinder locally emitted pollutants. Therefore pollution levels vary from street to street due to differences in emission strength and factors influencing dispersion of air pollutants. It is challenging to capture microlevel (from house to house and street to street) spatial heterogeneity in air pollution levels in urban areas using the traditional AQ monitoring network. Although sufficient for current regulatory need, the AQ monitoring network (AQMN) operated by DEFRA and Sheffield city council is not dense enough to account for microlevel spatial variations in air pollutant concentrations. In Sheffield there are three AQMS operated by DEFRA and five operated by Sheffield city council. This suggests a need for further improvement in AQMN in Sheffield. One of the motivation of this project is to structure a dense network of AQ sensors in Sheffield, made of several layers of sensors including reference sensors, LCS (e.g., AQMesh and Envirowatch E-MOTEs) and IoT (internet of things) sensors. These sensors are deployed in different parts of the city (static network), mounted on vehicles and carried by people (mobile network) (Figure 1.1). In addition, there is a network of over 180 NO_2 diffusion tubes in Sheffield. It should be noted here that NO_2 diffusion tubes and continuous air quality monitoring sensors have totally different temporal resolution, and therefore, for comparison in this study we will use annual average NO_2 concentrations for both diffusion tubes and other sensors. The proposed dense AQMN will produce AQ data, which will help develop high-resolution spatial maps in Sheffield. Historically AQMN are based on a single criterion, whereas the proposed AQMN will be based on multiple criteria using social, economic and environmental indicators. In this project data from static network is used only.

Furthermore, using data from the designed network the project intends to investigate how LCS can be used to render high quality data. For this purpose, air pollution measurements of LCS and reference sensors will be compared using both linear and nonlinear modelling techniques. Measurements of the LCS will be adjusted using the developed calibration methods.

AQ data collected by various sensor types will be used to produce high-resolution maps employing advanced modelling and mapping approaches including geostatistical interpolation, land-use regression and dispersion modelling techniques. Furthermore, measured and estimated concentrations will be fused using data fusion techniques, e.g. Universal Kriging. This will further support the monitoring network in improving spatial and temporal coverage in the city. The final goal is to provide a cheaper AQ monitoring network to produce data of acceptable quality, which will be achieved with the help of (a) structuring a purpose designed AQ network, (b) sensors calibrations, (c) data analysis and modelling, and (d) data fusion techniques.

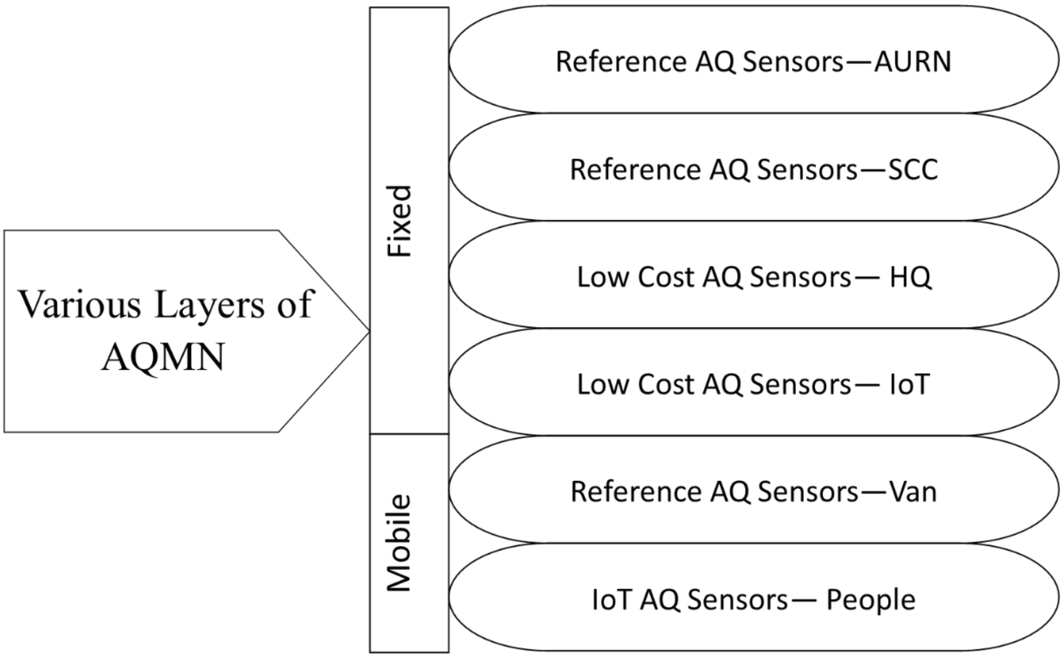


Figure 1.1. Showing different layers (each type of sensors is making a layer) of the proposed AQMN in Sheffield.

1.2. Main drivers of air pollution

Air pollution levels in urban areas are driven by two main factors:
(i) Emission sources
(ii) Influencing factors

Types and strength of emission sources are the main drivers of air pollutants in urban areas. Air pollutant emissions include line sources like road traffic, point sources like incinerators and area sources like residential emissions. Road traffic is considered the main emission source in large urban areas. In this project the main sources of NO_2 and PM_{10} will be determined using monitoring and modelling techniques.

Influencing factors are responsible for the dispersion of air pollutants both horizontally and vertically and include meteorological parameters, e.g., wind speed, wind direction and temperature, topographical and land-use characteristics, e.g., green spaces, large water bodies, commercial areas, altitude and hilly areas, and building density and height, e.g., tall buildings on both sides of a road in urban areas can act as a street canyon, which traps pollutants and hinders their dispersion. In addition, pollutant transformation (e.g., NO transformation to NO_2 and O_3, secondary particulates formation, and sinks also affect pollutant levels. LUR models will be used to determine the main influencing factors in Sheffield.

1.3. Aims and objectives

The core aims of this PhD project are: (a) to understand spatiotemporal variability of AQ; and (b) to determine the main drivers of air pollution in urban areas using Sheffield as a case study. The aims will be achieved by deploying a dense AQMN based on multiple criteria and employing various modelling and mapping techniques for AQ analysis and developing high-resolution maps. The aims of the project can be subdivided into the following specific objectives:

I To define criteria for structuring a multipurpose AQMN and to structure a dense network made of several layers of AQ sensors including reference, LCS and IoT sensors in Sheffield.

II To deploy LCS for rendering high quality measurements of NO_2 concentrations. This objective assesses and improves the quality of LCS data using calibrations models.

III To model the spatial variability of AQ in urban areas using various modelling approaches including dispersion and land-use regression models.

IV To explore the use of novel data fusion techniques to improve both model estimation and LCS measurements and further improve the quality of AQ maps.

V To analyse temporal variability of AQ using both graphical and time series analysis.

1.4. How different chapters (papers) address the objectives?

This PhD thesis is composed in the alternative format ‘publication format thesis’ incorporating a collection of papers that are already published. This section discusses how different papers/chapters are linked with the objectives of this PhD project and what each chapter intends to achieve. Main aim of each chapter is briefly described below (section 1.4.1 to 1.4.9). . Table 1.1 present the structure of the thesis and how different chapters are linked with the objectives of the project.

Table 1.1. Thesis structure and linking objectives with chapters

Chapter number	Brief description
Chapter 1	Introduction
Chapter 2	Literature review
Chapter 3	Objective 2 (LCS calibration)
Chapter 4	Objective 1 (Structuring an AQMN)
Chapter 5	Objective 3 (Spatial variability - dispersion modelling)
Chapter 6	Objective 3 (Spatial variability - land-use regression)
Chapter 7	Objective 4 (Data fusion)
Chapter 8	Objective 5 (Temporal variability)
Chapter 9	Conclusions

Below a brief description of each chapter is provided.

Chapter 1: Introduction

Chapter 1 briefly outlines the main aims, objectives and motivations of the project. 'How different chapters (papers) address the objectives' are also discussed in this chapter. This chapter introduces the project, outlines the main purpose of the project and briefly discusses how the project was carried out.

Chapter 2: What is the state of the art for air quality monitoring and modelling?

Chapter 2 presents background materials and resources of air quality in the UK, especially in Sheffield. In chapter 2 a detailed literature review is carried out covering various air quality modelling approaches, low-cost sensors and data fusion techniques.

Chapter 3: Analysing the performance of low-cost air quality sensors, their drivers, relative benefits and calibration in cities - a case study in Sheffield

This chapter is based on the published paper: *Munir, S., Mayfield, M., Coca, D., Jubb, S.A., Osammor, O., 2018. Analysing the performance of low-cost air quality sensors, their drivers, relative benefits and calibration in cities - a case study in Sheffield. Environmental Monitoring and Assessment, 191(2):94.* The paper addresses **objective # 2** of the project: "To deploy LCS for rendering high quality measurements of NO_2 concentrations".The paper analyses and compares NO_2, NO and CO concentrations measured by Envirowatch E-MOTEs and reference sensors for a year to determine how good the LCS are for measuring these pollutants. A calibration approach is recommended for improving the quality of LCS measurements. Chapter 4: Structuring an integrated air quality monitoring network in large urban areas – Discussing the purpose, criteria and deployment strategy

This chapter is based on the published paper: *Munir, S., Mayfield, M., Coca, D., Jubb, S.A., 2019. Structuring an Integrated Air Quality Monitoring Network in Large Urban Areas – Discussing the Purpose, Criteria and deployment Strategy. Atmospheric Environment: X 2 (2019): 100027.* The paper addresses **objective # 1** of the project: 'To define criteria for structuring a multipurpose AQMN and to structure a dense network made of several layers of AQ sensors including reference, LCS and IoT sensors in Sheffield'. In this paper, a methodology is proposed which is supported by numerical, conceptual and GIS frameworks for structuring an air quality monitoring network using social, environmental and economic indicators as a case study in Sheffield, UK. The main factors used for air quality monitoring station selection are population-weighted pollution concentration (PWPC) and weighted spatial variability (WSV) incorporating population density (social indicator), pollution levels and spatial variability of air pollutant concentrations (environmental indicator). The total number of sensors is decided on the basis of budget (economic indicator), whereas the number of sensors deployed in each output area is proportional to WSV. The purpose of AQ monitoring and its role in determining the location of air quality monitoring stations is analysed.

Chapter 5: Analysis of air pollution in urban areas with Airviro dispersion model - A case study in the City of Sheffield, United Kingdom

This chapter is based on the published paper: *Munir, S., Mayfield, M., Coca, D., Mihaylova, L.S., Osammor, O., 2020. Analysis of air pollution in urban areas with Airviro dispersion model - A case study in the City of Sheffield, United Kingdom. Atmosphere, 11(3), 285*. The paper addresses **objective # 3** of the project: 'To model the spatial variability of AQ in urban areas using various modelling approaches including dispersion and land-use regression models '. In this paper two air pollutants, oxides of nitrogen (NOx) and particulate matter (PM_{10}) are monitored and modelled employing Airviro air quality dispersion modelling system in Sheffield. The aim is to determine the most significant emission sources and analyse the spatial variability of these two pollutants.

Chapter 6: A nonlinear land-use regression approach for modelling NO_2 concentrations in urban areas – using data from low-cost sensors and diffusion tubes

This chapter is based on the published paper: *Munir, S., Mayfield, M., Coca, D., Mihaylova, L.S., 2020. A nonlinear land-use regression approach for modelling NO_2 concentrations in urban areas – using data from low-cost sensors and diffusion tubes. Atmosphere, 2020,* 11*(7), 736*. This paper addresses **objective # 3** of the project: 'To model the spatial variability of AQ in urban areas using various modelling approaches including dispersion and land-use regression models'. In this paper, a nonlinear generalised additive model was proposed for LUR and its performance was compared to a linear LUR model in Sheffield, UK for the year 2019. Pollution models were estimated using NO_2 measurements obtained from 188 diffusion tubes and 40 low-cost sensors. Performance of the models was assessed by calculating several statistical metrics including correlation coefficient (R) and root mean square error (RMSE). High-resolution (100 m x100 m) maps were created for NO_2 concentrations. Chapter 7: Understanding spatial variability of NO_2 in urban areas using spatial modelling and data fusion approaches

This chapter is based on the published paper: *Munir, S., Mayfield, M., Coca, D., 2021. Understanding spatial variability of NO_2 in urban areas using spatial modelling and data fusion Approaches. Atmosphere, 2021, 12(2), 179.* This paper addresses **objective # 4**: 'To explore the use of novel data fusion techniques for improving both model estimations and LCS measurements and further improving the quality of AQ maps'. The aim in this chapter was to analyse small-scale spatial variability in NO_2 concentrations with the help of pollution maps. Maps of NO_2 were produced using geostatistical interpolation, Airviro dispersion model and LUR model. Finally, NO_2 concentrations estimated by Airviro and LUR models were fused with the measured NO_2 concentrations. Chapter 8: Application of density plots and time series modelling to the analysis of nitrogen dioxides measured by low-cost and reference sensors in urban areas

This chapter is based on the published paper: *Munir, S., Mayfield, M. 2021. Application of density plots and time series modelling to the analysis of nitrogen dioxides measured by low-cost and reference sensors in urban areas. Nitrogen, 2, 167–195.* This paper addresses **objective # 5**: "To analyse temporal variability of AQ using both graphical and time series analysis". In this chapter temporal variability of NO_2 concentration was analysed at different temporal scales. In the first part of chapter 8, NO_2 concentrations measured by twenty-eight Envirowatch E-MOTEs, thirteen AQMesh pods, five reference sensors operated by Sheffield City Council and three reference sensors operated by DEFRA were analysed. NO_2 concentrations measured by different sensors were compared using density plots and time variation plots for over a year (August 2019 to September 2020). Long-term trend of NO_2 concentration was also determined for over twenty years (2000 – 2020). In the second part of the chapter, NO_2 concentrations were analysed employing univariate linear and nonlinear time series models based on persistence, and their performance was compared with a more advanced time series model using two exogenous variables (NOx and O_3).

Chapter 9: Conclusions and future work

Chapter 9 briefly discusses the main findings and summarises the results of this PhD project. Furthermore, the chapter provides suggestions for future work to further improve AQ monitoring and modelling in Sheffield. Main future works include understanding indoor-outdoor interface of air quality and characterising chemical composition of particulate matter in Sheffield.

1.5. Summary of the chapter

Air pollution is a growing concern especially in large urban areas causing various negative impacts on the surrounding environment and human health. Air pollutant levels have declined in the UK during the last couple of decades, however the reductions are not sufficient and therefore air pollutant levels especially NO_2 and PM_{10} still exceed AQ standards in large urban areas in the UK. Therefore, further interventions are required for effectively managing air quality to reduce exposure and negative impacts. In this project, different air quality monitoring and modelling approaches are deployed to understand the spatial variability of air quality in a typical urban city using Sheffield as a case study. This chapter described the main aims and objectives of this project and linked the various objectives of the study with different chapters of the thesis. In the next chapter a detailed literature review is carried out to put this project in the context of what is already done and identify a research gap, which will be addressed in the later chapters.

1.6. References

1. DEFRA, 2015. Improving air quality in the UK Tackling nitrogen dioxide in our towns and cities, UK overview document, December 2015. Available on: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/486636/aq-plan-2015-overview-document.pdf (accessed 09/10/2017).
2. WHO, 2013. Health effects of particulate matter, policy implications for countries in eastern Europe, Caucasus and central Asia. Publications of WHO Regional Office for Europe UN City, Marmorvej 51 DK-2100 Copenhagen, Denmark.
3. Landrigan, P.J., 2016. Air pollution and health. The Lancet Public Health, 2(1): 4 – 5. DOI: https://doi.org/10.1016/S2468-2667 (16)30023-8.
4. Daly, A., Zannetti, P., 2007. Air Pollution Modelling – An Overview. Chapter 2 of AMBIENT AIR POLLUTION (P. Zannetti, D. Al-Ajmi, and S. Al-Rashied, Editors). Published by The Arab School for Science and Technology (ASST) (http://www.arabschool.org.sy) and The EnviroComp Institute (http://www.envirocomp.org/).
5. Manisalidis, I., Stavropoulou, E., Stavropoulos, A., Bezirtzoglou, E., 2020. Environmental and Health Impacts of Air Pollution: A Review. Frontiers in Public Health, 8:14. doi: 10.3389/fpubh.2020.00014.
6. Fuller, R., Landrigan, P.J.,Balakrishnan, K., Bathan, G., 2022. Pollution and health: a progress update. Lancet Planet Health 2022; 6: e535–47. Published Online May 17, 2022 https://doi.org/10.1016/ S2542-5196(22)00090-0.

CHAPTER 2: WHAT IS THE STATE OF THE ART FOR AQ MONITORING AND MODELLING?

2.1. Introduction

This chapter provides a detailed literature review of various air quality (AQ) modelling and monitoring approaches. The aim is to find out as to what is the state of the art for AQ modelling, data fusion and low-cost sensors (LCS) in light of the current literature. This will help identify the research gap, which will shape this project.

Firstly, it is important to understand why air quality is important and how it affects human health. Air pollution is one of the most serious environmental threats to health, killing 6.4 million people in 2015 worldwide both in developed and less-wealthy nations (Landrigan, 2016). Out of these, 2.8 million deaths were caused by indoor air pollution and 4.2 million deaths by outdoor air pollution. Air pollution is causing various health problems including respiratory problems, cardiovascular diseases, lung cancer and asthma (WHO, 2013). Walters and Ayres (2001) have also reported that air pollution especially particulate matter and nitrogen dioxide (NO_2) pollution may cause premature deaths and hospital admissions for conditions such as cardiovascular problems, allergic reactions and lung cancer. The UK Committee on the Medical Effects of Air Pollutants (COMEAP, 2009 & 2010) have investigated the negative effects of long-term exposure to air pollution on human health and considered $PM_{2.5}$ as the pollutant responsible for increased risk of mortality. It is reported that exposure to air pollution is particularly harmful for children, people with existing health problems and the elderly (Khallaf, 2011). Evidence suggests that the negative impacts of air pollution are dependent on the levels of air pollutants and length of exposure, where higher levels and longer exposure result in more severe adverse effects (Khallaf, 2011; Walters and Ayres, 2001). Furthermore, air pollution may reduce visibility, damage historical buildings and monuments, affect vegetation and reduce crop yield and quality (Khallaf, 2011; Ivaskova et al., 2015).

It is important to mention that air pollution is a serious issue anywhere, however, it is considered more of a serious problem in large urban areas (Brunt et al., 2016; DEFRA, 2017). This is due to the fact that urban areas possess greater number of emission sources and have densely built-up areas including tall buildings and street canyons, which hinder dispersion of locally emitted air pollutants (Wu et al., 2017). Urban areas have various emission sources, e.g., road traffic, point emissions and area emissions emitting high volume of both gaseous (e.g., NO, NO_2, CO and SO_2) and particle pollutants (e.g., PM_{10} and $PM_{2.5}$) (DEFRA, 2017). In addition, more people live per unit area in urban areas, this causes the exposure of more people to air pollution. These factors like presence of high pollution concentrations and high population density make air pollution a more serious issue in urban areas. Therefore, it is vital to characterise the spatial variability of air pollution in urban areas.

Understanding spatial variability of air pollution in urban areas has been a serious challenge due to the sparse air quality network operated by DEFRA and local authorities in large urban areas in the UK. This is also the case in Sheffield, where only three air quality monitoring stations (AQMS) are installed by DEFRA and 5 by the Sheffield City Council. The reason is that reference sensors are expensive to purchase and maintain, therefore historically it was not possible to design a dense air quality monitoring network in urban areas. However, with the emergence of the low-cost sensors (LCS), now structuring a dense network has become feasible. As mentioned in Chapter 1, in this PhD project the intention is to structure a dense air quality monitoring network made of several layers of sensors including reference sensors, LCS and IoT sensors. Data collected by the dense network will be analysed using different data analysis and modelling techniques to create high-resolution maps. Sensors calibration and data fusion techniques will be employed to improve data quality by comparing and fusing air quality measurements with estimated pollution concentrations. Below a detailed review of the current air quality modelling and monitoring techniques is provided to understand the state of the art of such techniques.

2.2. Urban air quality modelling

Air pollution models are numerical tools for describing the causal relationship of atmospheric pollutants with emissions, meteorology, deposition, chemical transformation, and other factors like topography and land-use (Daly and Zannetti, 2007). Air quality modelling is carried out for several purposes including air quality prediction/forecasting, quantifying the impacts of air pollution (e.g., health impacts), modelling the impacts of various factors on air pollution, analysing the relationship between different pollutants, modelling pollution processes and transport, quantifying pollutant concentrations, deposition and environmental fate, running and testing emission scenarios, predicting the dispersion of air pollutants in the atmosphere based on emissions and meteorological parameters, quantifying the emissions of air pollutants from various emission sources, determining long-term trend in air pollutant concentrations, and producing high resolution spatiotemporal maps of air pollution (Aldrin and Haff, 2005; Andersen et al., 2006; Arnold et al., 2005; Baur et al., 2004; Berastegi et al., 2001; Brasseur et al., 1998; Munir et al., 2013; Westmoreland et al., 2007). Dispersion models are also used for emergency planning of accidental chemical releases (Wilkening and Baraldi, 2007).

Modelling studies are carried out on local levels like modelling of pollutants dispersion from a road or a point source, regional scale smog modelling and global circulation modelling. Air pollutant concentration in the atmosphere is determined by four main processes: emissions, transport, chemistry and deposition. Air quality modelling try to quantify the effect of these processes and determine atmospheric concentrations. In this review air quality modelling is divided into two main branches: Dispersion modelling and Statistical Modelling. These are further divided into several sub-branches as shown in Figure 2.1. Each type of these modelling approaches is reviewed below.

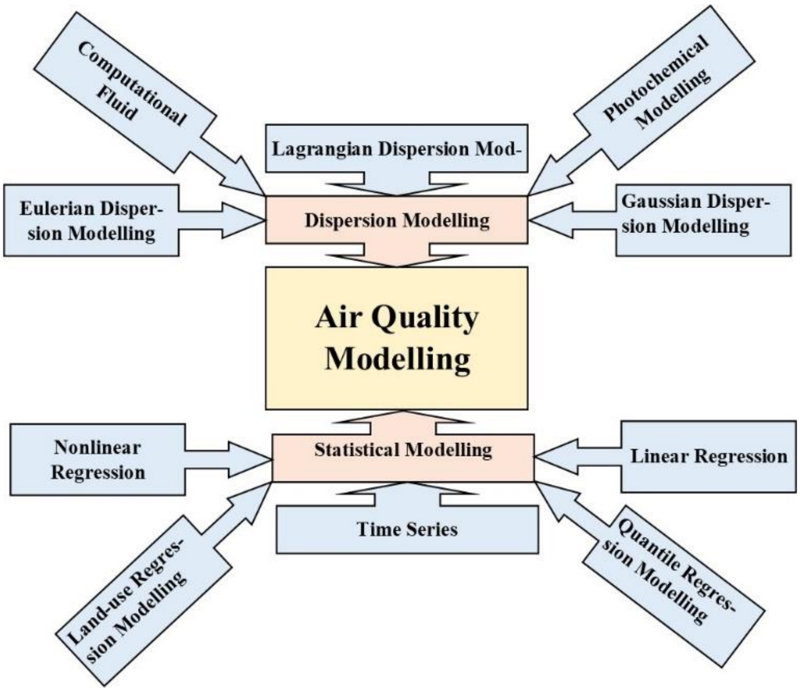


Figure 2.1. Various types of air quality modelling

2.2.1. Dispersion air quality modelling

Dispersion air quality modelling techniques are effective tools for determining downwind concentrations of air pollutants at a given time and space emitted by a known emission source for example an industrial plant or road traffic. These models are mathematical representation of the atmospheric processes determining the rate at which pollutants are mixed with clean air. Dispersion models are run with computer programs, which are able to solve the mathematical equations and algorithms to simulate the pollutant dispersion in the atmosphere. Dispersion models have the potential to incorporate both temporal and spatial variations to replace the need for air pollutant monitoring. However, high cost of purchase, maintenance and high input demands limit their application.

Dispersion of air pollutants in the atmosphere is mainly controlled by (e.g., Daly and Zannetti, 2007):

(a) Advection/transport,

(b) Diffusion,

(d) Chemical reaction, and

(e) Deposition of air pollutants.

Atmospheric dispersion model requires various inputs to produce the required outputs. The main inputs of dispersion model include (CERC, 2017): (a) Meteorological inputs, e.g., wind speed and direction, cloud cover, solar radiation, air temperature and stability classes for defining atmospheric turbulence, (b) Detailed emission data of pollutants, (c) Emissions release parameters which include source location, height, type of source (line source, area source and point source), exit velocity, exit temperature and mass flow rate or release rate, (d) Terrain characteristics at both source and receptor locations, and (e) Surface roughness, defined in terms of rural or urban terrain.

A number of techniques exist for modelling the dispersion of air pollutants, which vary in sophistication but all include some sort of simplification of the dispersion processes. Selecting a particular modelling approach depends on several factors such as the temporal and spatial scale of the modelling, resolution of the data available to run the model, the purpose of the modelling, skill of the modeller, time available and financial and computer resources available (El-Harbawi, 2013).

There are three main types of dispersion modelling techniques: Gaussian modelling, Eulerian grid modelling and Lagrangian trajectory modelling (Salmond et al., 2006). El-Harbawi (2013) reviewed dispersion modelling techniques and divided them into five main types: Gaussian models, Box models, Lagrangian models, Eulerian Models and Dense gas models. Modi et al. (2013) reviewed dispersion modelling techniques in which in addition to the above five types they mentioned two other models: Computational Fluid Dynamic (CFD) and Aerosol Dynamic Models.

2.2.1.1. Gaussian plume modelling

Gaussian modelling was introduced by Pasquill (1961, 1962 and 1974) and Briggs (1965). Gaussian modelling technique is simple and computationally efficient and require simple input data. Gaussian modelling is normally used for fast screening type calculation of the pollutant dispersion from point sources, line sources or area sources. Urban areas can be modelled as a sum of area sources (e.g., domestic emission), point sources (e.g., factory or power stations), and line sources (e.g., road traffic). Gaussian modelling treats dispersion as a statistical process rather than representing individual turbulent motions of the atmosphere. Well known examples of Gaussian dispersion modelling are Airviro (Gauss model), AERMOD - EPA, CTDM (Complex Terrain Dispersion Model – EPA), ADMS – CERC UK, CALINE3 for highway air pollution and Offshore and Coastal Dispersion (OCD) model for coastal areas.

In the Gaussian plume the expanding plume has a Gaussian or normal distribution of concentrations in the vertical (z) and lateral (y) directions. The concentration (C) (µg/m^3) at any point (x, y, z) is given by (William, 2000):

$$C(x,y,z,t) = \frac{Q}{2\pi . u . \sigma_y . \sigma_z} . \exp\left(-\frac{y^2}{2\sigma_y^2}\right) . \left[\exp\left(-\frac{(z-H_{eff})^2}{2\sigma_z^2}\right)\right] + \exp\left(-\frac{(z+H_{eff})^2}{2\sigma_z^2}\right) \qquad (2.1)$$

Where,

C (x, y, z) pollutant concentration at point (x, y, z)

x, y, and z are the along wind, crosswind, and vertical distance

µ is wind speed (in the x downwind direction, m/s)

σ represents the standard deviation of the concentrations in the y and z direction

Q is the pollutant mass emission rate (g/s)

H_{eff} is the effective stack height (height of the stack plus the plume rise)

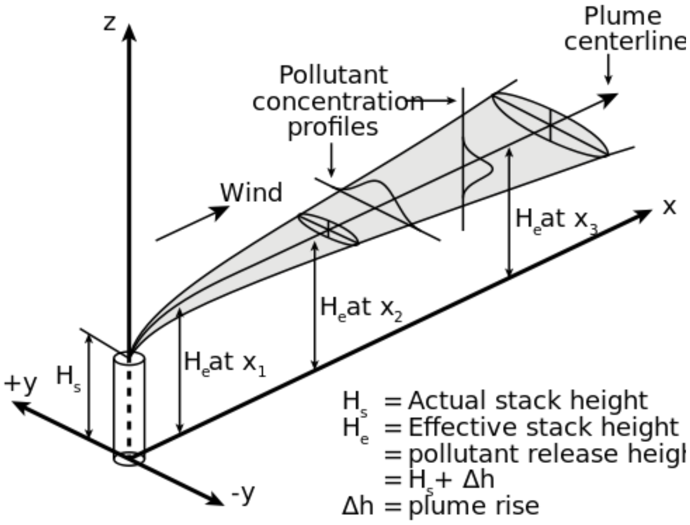

z
Plume centerline
Pollutant concentration profiles
Wind
x
H_e at x_3
H_e at x_2
H_e at x_1
+y
H_s
-y
H_s = Actual stack height
H_e = Effective stack height
= pollutant release heigl
= H_s + Δh
Δh = plume rise

Figure 2.2. Schematic diagram of a Gaussian plume (Lagzi et al., 2013)

Gaussian plume models are steady state model, assuming that conditions remain similar over the averaging period. Therefore, Gaussian models cannot capture the random fluctuations in a real plume. Gaussian models are suitable for modelling average concentrations of non-reactive pollutants, and therefore are commonly used for primary pollutants emitted by a road traffic and industrial chimneys.

2.2.1.2. Lagrangian and Eulerian Dispersion Modelling

Eulerian and Lagrangian models are state-of-the-art tools of recent atmospheric dispersion simulations (Dacre et al., 2011). In Lagrangian modelling an air parcel (puff) is followed along a trajectory and is assumed to keep its identity during its path, whereas in Eulerian modelling the area under investigation is divided into grid cells both in vertical and horizontal directions (Daly and Zannetti, 2007) and mathematical equations are solved for every grid point in the domain at each time step.

Lagrangian models are simpler in concept and application than Eulerian models, which require greater complexity of inputs and computing (William, 2000). In contrast to Gaussian, these models are run for long range transport (>100 km). These models incorporate dry and wet deposition as well as chemical reaction of pollutants. Furthermore, they take into account large scale meteorological parameters like specifying the movement of air mass on a synoptic scale. Lagrangian models treat the atmosphere as a series of air parcels moved around within a wind field. In this way, it follows the trajectories of a single or multiple air parcels as they are transported by the wind over long distances. The Eulerian approach treats atmosphere as a grid made up of volumes or boxes, whose properties change with time. Perfect mixing is assumed within each grid. In this case the mathematical equations representing transport and chemical transformation have to be solved at each grid point. Eulerian models are better for modelling dispersion from area sources or secondary pollutants (e.g., ozone). However, Eulerian models are computationally expensive and require large parallel computers to model smog or global climate models.

Example of Lagrangian model: European UNECE/EMEP (European Monitoring and Evaluation Program) model, UK meteorological Numerical atmospheric dispersion modelling environment (NAME) model, UK meteorological STOCHEM model, Hybrid Single Particle Lagrangian Integrated Trajectory Model (HYSPLIT) developed at NOAA's Air Resources Laboratory. Example of Eulerian dispersion modelling: Urban Airshed Model (UAM) iv and v, the California Grid Model (CALGRID), Comprehensive Air Quality Model with Extensions (CAMx), the Regional Oxidant Model (ROM), the Community Multiscale Air Quality Model (CMAQ), Airviro (Grid model), the Regional Modelling System for Aerosols and Deposition (REMSAD) and the Community Multiscale Air Quality (CMAQ-UK) model.

Eulerian and Lagrangian models are used for simulating air pollutant dispersion on large scale, e.g., meso- to macro-scale. These models depend on numerical weather prognostic (NWP) for meteorological data. However, the grid resolution of NWP models is normally from 1 to 10 km, which is too coarse for urban air pollution modelling where source and receptor points are often located within a few hundred meters from each other, surrounded by a very complex geometry. Therefore, Eulerian and Lagrangian models cannot be used for urban scale modelling of air quality. For urban scale modelling three approaches can be applied: Gaussian dispersion modelling (section 2.2.1.1), Computational Fluid Dynamics (CFD) modelling (section 2.2.1.3) and statistical modelling (2.2.2).

2.2.1.3. Computational Fluid Dynamics (CFD) Model

CFD are vital tools that can model airflow and pollutant dispersion within up to 1 m resolution in urban areas (Sanchez et al, 2017). CFD models provide a tool to solve various partial differential equations (PDEs) that define the governing equations of transport and dispersion of air pollutants. The basic equations of motion are known as the Navier–Stokes equations, which relate the conservation of mass, energy and momentum. The basic form of the Navier–Stokes equation for turbulent incompressible fluids is given below:

$$\frac{\partial \vec{v}}{\partial t} + \left(\vec{v}\nabla\right)\vec{v} = -\frac{1}{\rho}\nabla p - \vec{g} + \nu_T \nabla^2 \vec{v} \tag{2.2}$$

where $\vec{v}$ is the wind field, ρ is density, p is pressure, ν_T is the eddy viscosity and $\vec{g}$ is the gravitational acceleration vector. CFD are flexible and can be applied at microscale. CFD models consist of four main parts (Lagzi et al., 2013): (i) A mesh generator, which splits the spatial domain into a user defined resolution cells; (ii) A PDE solver, which solves the Navier–Stokes and other equations; (iii) Turbulence model, which characterise the turbulence in the complex geometry; and (iv) A visualization tool, which represents the outputs of the model.

To predict a pollutant concentrations, CFD requires traffic emission and meteorological data. Emission data are obtained from an emission model coupled to a microscale traffic simulation system, whereas wind speed and direction are obtained from a microscale meteorological model. In CFD modelling longer periods of time, for example several weeks or months, is generally a problem due to huge computational time. To overcome this problem, Parra et al. (2010) and Santiago et al. (2017) suggested steady CD-RANS (Computational Fluid Dynamics - Reynolds-averaged Navier-Stokes) modelling approach. Sanchez et al. (2017) applied CFD model in Madrid, Spain to model the spatial distribution of NOx concentrations in a heavily trafficked urban area. They followed a weighted average CFD simulations to compute the time evolution of NOx dispersion considering the actual atmospheric conditions. Pollutant emissions were estimated from the traffic emission model and meteorological parameters were derived from a mesoscale meteorological model. The predicted NOx concentrations were compared with measured concentrations. The estimated concentrated correlated well with measured concentrations in the research area in Madrid. The German MISKAM and the French MERCURE are examples of CFD software designed for atmospheric dispersion and wind engineering studies.

In addition to studying the atmospheric dispersion of pollutant in urban areas, the CFD models have good potential for studying anthropogenic ventilation sources of air dynamic and circulation. The main shortcoming is that due to the complexity of its mathematics, CFD simulations are processor-intensive, requiring substantial computational time and modelling cost.

2.2.1.4. Photochemical air quality models

The photochemical air quality models are large-scale models which simulate changes in pollutant concentrations using mathematical equations that describes the chemical and physical processes in the atmosphere (Daly and Zanetti, 2007). Photochemical models can be applied at all levels ranging from local to global scale. Photochemical models can play a vital role in understanding how pollutants evolve in the atmosphere and in determining the levels of various photochemical pollutants, mainly O_3 and secondary NO_2 which are formed in the atmosphere from the chemical reactions of primary pollutants.

Photochemical models try to simulate the ways in which air pollutants form, accumulate, and dissipate in the atmosphere (Yarwood et al., 2014). They do this by describing the process which form O_3 molecules in the atmosphere. For instance, emission of pollutants from a point source or a line source that can contribute to O_3 formation. Another example is the simulation photochemical reactions that lead to O_3 pollution. Photochemical reactions are chemical reactions which are caused by solar radiation. Photochemical reactions may result in the formation of radicals which react with volatile organic compounds (VOCs) and NOx to form O_3 molecules (Yarwood et al., 2014).

Examples of photochemical air quality models include Community Multiscale Air Quality (CMAQ), Comprehensive Air Quality Model with Extensions (CAMx), Regional Modelling System for Aerosols and Deposition (REMSAD), and Urban Airshed Model (UAM) (El-Harbawi, 2013).

2.2.1.5. Hybrid models

Several models have been developed which used more than one modelling techniques known as hybrid models. For example, Airviro modelling system uses Guass model to calculate concentrations of pollutants above ground (open landscape) or rooftop (buildings) on the local/urban-scale, and it uses the Grid model to calculate concentrations on a regional scale.

2.2.2. Statistical modelling

Statistical modelling utilises mathematical equations based on sampled data to approximate the underlying phenomena (population). In other words, statistical modelling is the formalisation of relationships between variables in the form of mathematical equations. Inferential statistics uses observed data to deduce properties of an underlying probability distribution. In contrast, descriptive statistics characterises only the sampled data without the assumption that it comes from a population.

Statistical modelling techniques are widely applied to analyse air pollutant data around the globe. A brief review of these studies are provided below:

2.2.2.1. Time series modelling

Time series modelling is the act of predicting the future by understanding the past (Adhikari and Agrawal, 2013). Time series is a set of observations in sequence over successive times. Mathematically time series is a set of vectors x (t), where t represents the time elapsed and its value could be 0, 1, 2….n, and x is a random variable. The data points in a time series should be arranged in a chronological order. Time series could be either univariate (made of a single variable) or multi-variate (made of several variables). Time series is generally made of four components: Trend, Seasonal and Irregular (random) components (Adhikari and Agrawal, 2013).

To account for the four components time series models could be either multiplicative or additive.

$$\text{Multiplicative time series model: } Y(t) = T(t)\times S(t)\times C(t)\times I(t) \qquad (2.3)$$

$$\text{Additive time series model: } Y(t) = T(t) + S(t) + C(t) + I(t) \qquad (2.4)$$

Where Y(t) is the time series variable (observations) and T(t), S(t), C(t) and I(t) are trend, seasonal, cyclical and irregular components. Multiplicative time series model assumes that the components of a time series are not necessarily independent and might affect each other; whereas the additive time series model assumes that the four components are independent of each other. A time series is non-deterministic in nature, which in simple words means that the future prediction cannot be certain. The probability structure of a time series is termed as a stochastic process (Hipel and McLeod, 1994).

Time series modelling is a useful tool for data analysis and have several benefits, which include data cleaning, data understanding and forecasting (Bush, 2020). Time series modelling can help us filter out the noise and reveal the true signal in a dataset. Once dataset is cleaned and the time series is divided into its different components, it helps us understand the true nature of the dataset. Finally, like other modelling approaches, time series modelling helps us predict future levels with the help of present and past levels of the time series (here NO_2 concentrations) (Bush, 2020). To build a time series model the time series must be stationary. For the time series to be stationary the mean, variance and covariance of the series should not be a function of time (Srivastava, 2015). If the time series is not stationary, it should be stationarised first and then fit a stochastic model (Srivastava, 2015). Differencing and power transformations can be used to detrend and deseasonalise the time series to make it stationary (Adhikari and Agrawal, 2013).

Time series model could be linear or nonlinear depending on the relationship between current and past observations (Adhikari and Agrawal, 2013). Autoregressive (AR) and moving average (MA) are the two widely used linear time series models (Hipel and McLeod, 1994). The AR and MA are combined to form the Autoregressive Moving Average (ARMA) and Autoregressive Integrated Moving Average (ARIMA) (Cochrane, 1997). To model and predict a seasonal time series the Seasonal Autoregressive Integrated Moving Average (SARIMA) model is used which is a variation of ARIMA (Hipel and McLeod, 1994). ARIMA along with its various variations are also known as the Box-Jenkins models because they are based on Box-Jenkins principle (Hipel and McLeod, 1994). Kadiyala and Kumar (2012) have reported three general categories of time series linear modelling: (a) the univariate time series models (e.g., AR, MA, ARMA, ARIMA); (b) the multivariate time series models (e.g., Autoregressive Moving Average model with eXogenous inputs (ARMAX) and Autoregressive Integrated Moving Average model with eXogenous inputs (ARIMAX)); and (c) the multiple time series models (e.g., Vector Autoregressive with eXogenous inputs (VARX), Vector Autoregressive Moving Average (VARMA), Seasonal Vector Autoregressive (SVAR), Vector Error Correction (VEC) model). Further details provided in chapter 8, Figure 8.1.

Pisoni et al. (2009) have developed a Nonlinear Autoregressive model with eXogenous variable (NARX) for predicting peak ozone concentrations in Italy. Pisoni et al. (2009) discussed the possible advantages of using polynomial NARX instead of artificial neural network. The model was able to predict ozone concentrations with the model performance similar to neural network-based NARX. However, in addition polynomial-based NARX model identified the best set of regressors for ozone prediction which is an extra advantage of the model. Diaz-Robles et al. (2008) and Goyal et al. (2006) have applied ARMA and ARIMA for predicting PM_{10} in urban areas. The models were unable to capture the nonlinear behaviour and extreme PM_{10} concentrations. Both ARMA and ARIMA require continuous past AQ data, which is sometimes an issue for employing these models. More advance time series modes like ARMAX and ARIMAX have the ability to include external variables (explanatory variables). ARMA and ARIMA models can be coupled with ANN (e.g., Diaz-Robles et al., 2008) and polynomial (e.g., Pisoni et al., 2009) to make hybrid models, which are more powerful. To build a hybrid model, firstly ARIMA or ARIMAX model is developed and then ANN or polynomial model is employed to model the residuals of the time series model. To predict daily maximum PM_{10} concentrations four models: ARMAX, ANN, Multiple Linear Regression (MLR) and a hybrid model (ARIMAX couple with ANN) were employed in Chile. The hybrid model produced significantly better performance capturing 100 % alert and 80 % pre-emergency episode. Goyal et al. (2006) developed ARIMA, MLR and hybrid of ARIMA-MLR to predict respirable particulate matter in India and Hong Kong. Again, the hybrid model demonstrated significantly better than the other individual two models in terms of particulate matter predictability. Rahman et al. (2016) employed two models for forecasting air pollution index in Malaysia. They employed seasonal autoregressive integrated moving average (SARIMA) and fuzzy time series (FTS). AQ data from three monitoring stations were used which included industrial, residential and sub-urban areas. SARIMA which used data from urban areas perform better, whereas FTS perform better when using data from suburban areas.

This indicates that classical time series (SARIMA) models can perform as good as the modern techniques (FTS).

This review shows that time series models have made continuous progress during the last decade or so. The updated time series modelling techniques not only can handle multivariate time series but also address nonlinearities and include exogenous variables. The most recent version is the NARMAX (Chen and Billings,1989), which has been applied in several engineering studies, for example Zito and Landau (2005) applied polynomial NARMAX model to diesel engine in which a practical identification procedure was developed. However, no example of NARMAX application to air quality was found in the current literature.

2.2.2.2. Linear regression models

Linear regression models are probably the most widely used models for air quality modelling. They can be represented as below:

$$Y_i = \beta_o + \beta_i X_i + \varepsilon i \qquad (2.5)$$

Where Y is the response (dependent) variable, X is the explanatory (independent or predictor variable), β_o is the intercept, β is the slope (coefficient) of the equation and ε is the error term. The regression parameters or coefficients (β_o and β_i) are generally calculated by minimising the sum of square errors (least square method).

Regression techniques correlate air pollutants such as NOx concentrations to independent (explanatory) variables (e.g., road traffic characteristics, meteorological parameters and the concentrations of other air pollutants). These statistical models are used: (a) to understand the association between response and explanatory variables; (b) to predict and forecast air pollutant levels; (c) investigate temporal and spatial variations in air pollutant levels; and (d) to analyse the health impacts of air pollutants (Aldrin and Haff, 2005; Baur et al., 2004; Gardner and Dorling, 2000). Traditional statistical models such as simple linear regression and multiple linear regression models usually show inferior performance in terms of predicting future air pollutant concentrations. These models are probably the most commonly used approaches for air quality modelling. However, as reported previously by several researchers (e.g., Baur et al., 2004; Gardner and Dorling, 2000; Munir et al., 2014), air pollutant data are non-normally distributed and the association between explanatory and response variables is non-linear, therefore classic statistics which assume normality and linearity of the data should not be applied to air quality data, otherwise it can result in erroneous results (Reimann et al., 2008).

Linear regression models were employed to predict several air pollutants (NO_2, PM_{10}, and O_3) using six pollutants (NO, NO_2, O_3, SO_2, CO, PM_{10}) and meteorological parameters (wind speed, solar radiation, humidity and temperatures) as predictors in Surabaya City, Indonesia (Syafei et al., 2015). Serial error correlations in the predicted concentration was addressed to improve the model prediction. Alternative, variables selection based on independent component analysis (ICA) and principal component analysis (PCA) were used to obtain subsets of the predictor variables to be imputed into the linear model. The former models using original variables showed much better performance in comparison to the model using ICA and PCA for predictors extraction.

Pires et al. (2008) compared the performance of five linear models for predicting the daily concentrations of PM_{10}. The models were: (i) multiple linear regression; (ii) principal component regression; (iii) independent component regression; (iv) quantile regression; and (v) partial least squares regression. Several air pollutants (SO_2, CO, NO, NO_2 and PM_{10}) and meteorological parameters (wind speed, wind direction, temperature and relative humidity) were used as predictors in the models. Using the training data set quantile regression model demonstrated the best, whereas independent component regression demonstrated the worst performance. Using the testing dataset multiple linear regression, principal component regression and partial least squares regression models showed similar results to each other and had better performance compared to the other approaches. Pires et al. (2008) reported that the model performance was affected by the data size and the fact whether the data were included in the training dataset or not.

2.2.2.3. Quantile regression model

Ordinary linear regression analyses the average relationship between the response and explanatory variable, whereas quantile regression describes the relationship at different points in the conditional distribution of response variable (y). These points could be median (50th percentile or 0.5 quantile), 1st quartile (25th percentile or 0.25 quantile), 3rd quartile (75th percentile or 0.75 quantile) or any other point like 0.1, 0.2, …. 0.9 quantile (Hao and Naiman, 2007). Quantile regression model is given by the following equation:

$$Yi = \beta_o^{(p)} + \beta_{(i)}^{(p)} Xi + \varepsilon_{(i)}^{(p)} \qquad (2.6)$$

In the above equation 'p' is the p^{th} quantile and its value lies between 0 and 1.

In the quantile regression models the full distribution of response variable (e.g., NOx) as a function of several important explanatory variables, e.g., other air pollutants like CO, O_3, SO_2 and meteorological variables like wind speed, direction, temperature and relative humidity. Standard regression models provide an incomplete picture of the relationship between response and explanatory variables as they consider only the mean response (Cade and Noon (2003). Quantile regression, which was developed by Koenker and Bassett (1978), provides a more complete picture of the conditional distribution by considering the whole distribution of the response variable.

Several researchers have applied quantile regression to the analysis of air quality data. Baur et al. (2004) utilised quantile regression model to analyse the nonlinear association between ground level ozone and local meteorology. Baur et al. (2004) demonstrated that quantile regression model produced better results than linear regression model in terms of R^2-value. Also, it was demonstrated that the effect of explanatory variables on ozone concentrations varied significantly at various quantiles of ozone distribution. Carslaw et al. (2013) used a quantile regression technique to analyse the air pollutant emissions of petrol and diesel passenger cars aiming to explore the effects of high power vehicles on exhaust emission. Carslaw et al. (2013) noted that significantly more insight was gained into the pollutant emissions by using quantile regression model which described and modelled the full distribution of vehicle emissions as a function of several important explanatory variables. Moreover, Sayegh et al. (2014) compared the performance of five modelling techniques, including quantile regression, multiple linear regression, generalised additive model, and 1-

way and 2-way Boosted Regression Trees models. They used several statistical metrics including root mean squared error (RMSE), correlation coefficients (r), the fraction of prediction within a factor of two (FAC2), Index of agreement (IA), and mean bias error (MBE) to assess the models performance. Sayegh et al. (2014) concluded that the performance of quantile regression model was better than the other models included in the study. Quantile regression model predicted PM_{10} concentrations with minimum error and high accuracy. The better performance of quantile regression model was believed to be due to the model's approximation behaviour which is based on a number of quantiles used in the model, in contrast to other models which is based on average association. Furthermore, Martin et al. (2009), Sousa et al. (2008), Munir et al. (2012) and Munir (2014) have employed quantile regression modelling techniques to model various air pollutants and reported significantly improved prediction as compared to some other regression approaches, mostly multiple linear regression.

It is important to note that quantile regression model has the ability to model pollutant concentration using different quantiles, for example 0.05 quantile (5th percentile), median or 0.5 quantile (50th percentile), 0.75 quantile (75th percentile) and so on. There are two ways to assess the performance of quantile regression model: (a) local performance that considers only the prediction of a single quantile, let say median; and (b) global performance which considers the prediction of several quantiles at the same time. It is not mentioned whether Pires et al. (2008) considered local or global performance of quantile regression model. Quantile regression model has the potential to show much better performance when global performance is assessed, which is reported by Baur et al. (2004), Sayegh et al. (2014), and Munir et al. (2014).

Several authors (e.g. Duenas et al. 2002) have reported that air pollutants data are not normally distributed. Furthermore, it is reported that air pollutants have nonlinear association with their predictors (e.g. Gardner and Dorling2000; Baur et al. 2004), which means that the contributions of the explanatory variables (e.g. meteorological variables) vary significantly at different levels of the response variable. Ordinary linear regression modelling techniques assume linearity of the relationship between response and explanatory variables and normal distribution of the error terms, therefore, when parametric linear regressions are applied to air quality data, they may result in biased results. It is, therefore, advised to apply more robust nonlinear statistical approaches for air quality data analysis.

2.2.2.4. Nom-linear regression model

Nonlinear models relax the assumption of linearity between response and explanatory variables. Polynomial regression and step functions are simple examples of nonlinear regressions, whereas regression splines, smoothing spline, local regression and generalized additive models are more sophisticated example of nonlinear regression modelling (James et al., 2013).

Singh et al. (2012) compared the performance of linear and nonlinear modelling for predicting air quality in Lucknow, India. The modelling approaches they employed were partial least squares regression (PLSR), multivariate polynomial regression (MPR) and artificial neural network (ANN). Three different ANN models were developed viz. multilayer perceptron network (MLPN), radial-basis function network (RBFN) and generalized regression neural network (GRNN). They modelled SO_2, NO_2 and particulate matter (PM) using several explanatory variables including meteorological parameters (air temperature, relative humidity, wind speed) and air pollutants (PM, NO_2, SO_2). Nonlinear models (MPR, ANNs) performed better than the linear PLSR models. Among nonlinear models, the performance of the ANN models was better than the low-order nonlinear MPR models. Furthermore, among ANN, GRNN outperformed the other two ANN showing high correlation between modelled and observed concentrations.

Nieto and Anton (2014) applied a nonparametric regression algorithm known as multivariate adaptive regression splines (MARS) in norther Spain to determine the controlling factors of air pollution. The MARS model produced results which were simple and easy to interpret. The model is computationally efficient and has the ability to accurately estimate the contributions of the input variables. The three fitted MARS models for NO_2, SO_2 and PM_{10} had coefficients of determination equal to 0.84, 0.90 and 0.77 and correlation coefficients equal to 0.92, 0.95 and 0.88, respectively, indicating an acceptable goodness of fit.

Examples of non-linear regression models are Generalised Additive Models (GAM) and Boosted Regression Trees (BRT) models, below a brief review of these models is provided.

Generalised Additive Models (GAM)

Generalised Additive Model (GAM) which was proposed by Hastie and Tibshirani (1990), relaxes some of the assumptions which are assumed by the classical statistical tests, such as normality of the error term and linear association between response and explanatory variables (Wood, 2006).

To permit nonlinear relationship between explanatory and response variables GAM replaces the linear components of multiple linear regression (βjxij) with a nonlinear smooth function fj (xij). GAM model can be expressed as given below:

$$\text{yi} = \beta_0 + \sum_{j=1}^{p} fi(Xij) + \varepsilon\text{i} \qquad (2.7)$$

In the above equation 'f' is a basic function known as smoothing function and can be parametric, semi-parametric or non-parametric in nature. Several options for selecting the smoothers are available including regression splines, natural spline and smoothing spline (Woods, 2017; James et al., 2013). GAM is additive because separate 'fj' for each Xj are calculated and then their total contributions are determined by adding them up.

GAM has the properties of both generalised linear models and additive models. GAM does not assume the response probability distribution to be normal and can allow it to be any member of the exponential family, such as exponential, gamma, Poisson or any other (Wood, 2006). To model the concentrations of NO, NO_2 and PM_{10} with the help of meteorological parameters and traffic characteristics, Aldrin and Haff (2005) applied GAM and assessed the model performance by comparing the predicted values with measured ones. On the other hand, Carslaw et al. (2007) employed a GAM for quantifying temporal trends in air pollutants mainly emitted by road traffic, such as NO, NO_2, CO, benzene and 1,3-butadiene at Marylebone Road London. Also, Westmoreland et al. (2007) used GAM to model NO_2 levels in a busy street canyon in Gillygate York. They compared the outcomes of GAM and ADMS-Urban dispersion model with measured concentrations of NO_2. The results showed that NO_2 concentrations predicted by GAM were in much better agreement with measured concentrations than those predicted by ADMS-Urban.

Li et al. (2013) modelled Polycyclic Aromatic Hydrocarbons (PAH), particle number count (PNC), NO_x and $PM_{2.5}$ using traffic, meteorology and elevation variables in Southern California. They compared the performance of a GAM and MLRM, where GAM explained 57-89% of the variance and performed significantly better than MLRM. Furthermore, wind speed and direction were found to be the most important predictors for all pollutants used as response variables, whereas traffic-related variables played a significant role for PAH, PNC, and NOx, and air temperatures and relative humidity for $PM_{2.5}$. Generally GAM performs better that the linear counter parts.

Boosted Regression Trees (BRT)

In statistical models recently more advanced approaches are available which have the capability to handle non-normal distributed data, explain interaction between various variables and provide better prediction. Such approaches include tree-based methods for classification and regression. Tree-based methods involve stratifying (segmenting) the explanatory variables into a number of simple regions (James et al., 2013). These methods are known as decision tree methods because the rules which are used for splitting the predictor variables are generally summarised in a tree form. Tree-based methods are simple and easy to interpret. Different techniques are available to produce multiple trees viz. bagging, random forests, and boosting (James et al., 2013).

Classification and regression trees methods also include Boosted Regression Trees (BRT), which is a type of additive regression model. BRT combines the strengths of two algorithms: Regression trees and Boosting techniques, where the former relates a response variable to its predictors by recursive binary splits and the latter combines many simple models to give improved predictive performance (Elith et al., 2008). Bagging creates multiple copies of the original training dataset using the bootstrap technique. Firstly, it fits a separate decision tree to each copy of the data, and then combine all of the trees in order to create a single predictive model (James et al., 2013). Boosting works in a similar way, except that the trees are grown sequentially i.e. each tree is grown using information from previously grown trees. Boosting does not involve bootstrap sampling, instead each tree is fit on a modified version of the original dataset. To apply a BRT model, the user should first select the BRT tuning parameters which include a loss function (distribution), number of iterations, number of trees, the depth of each tree, interaction depth, the learning rate parameter λ (shrinkage) and the subsampling rate p (bag fraction) (Suleiman et al., 2015).

BRT approaches have been applied in a number of environmental studies including air quality modelling. Carslaw and Taylor (2009) have applied BRT to air quality data collected near Heathrow Airport in London UK. They used hourly NOx data and modelled the complex interaction between various emission sources, especially they tried to separate air pollutant emissions from air crafts and road transports. In this study BRT model was applied to determine which variable had more effect on NOx concentrations. Air craft emissions of NOx were estimated and the effects of meteorological conditions and runway patterns were determined. Furthermore, the model results were compared with those from a more detailed independent field campaign, where a close association was observed between the model outcomes and field observations. Sayegh et al. (2016) applied BRT model to analyse how roadside NOx concentrations were affected by traffic density, background concentrations of NOx and meteorological conditions at three different monitoring sites, which were urban road, open motorway and motorway tunnel. BRT model showed a strong association between NOx concentrations and background NOx levels. Results showed that model prediction was strongly influenced by the quality and resolution of explanatory variables, especially background NOx levels.

Suleiman et al. (2015) compared the performance of artificial neural network (ANN) and BRT to model the concentrations of PM_{10}, $PM_{2.5}$ and particle number counts on Marylebone Road in London using air pollution, traffic and meteorological data as predictors in the model. The correlation coefficient (R) values of the ANN and BRT models were 0.96 and 0.95 for PM_{10}, 0.96 and 0.96 for $PM_{2.5}$ and 0.89 and 0.87 for PNC, respectively. Suleiman et al. (2015) recommended BRT for air quality modelling despite the fact that it showed slightly weaker performance because BRT model offered much better interpretation and permit feature selection. Furthermore, Suleiman (2015) assessed several statistical and machine learning approaches and concluded that machine learning models performed better than the ADMS-road model in spatiotemporal predictions of air pollution. Moreover, machine learning techniques performed significantly better in predicting the concentrations in street canyons.

2.2.2.5. Land-use regression model

Geographical Information System (GIS) based interpolation methods, such as Kriging, can be used to interpolate air pollutant concentrations between various air monitoring stations to provide better spatial coverage. However, when these approaches are applied at a local scale such as intra-city scale, these methodologies are known to produce considerable variations in air pollutant concentrations within a small area and are more effective at large scales, such as national or regional scale (Briggs, 2005). Land Use Regression (LUR) models provide an effective alternative on urban scale. LUR approach was introduced by Briggs et al. (1997) and since has been used in numerous studies around the world (e.g., Eeftens et al., 2012; Lee et al., 2013; Muttoo et al., 2017; Rehman et al., 2017). LUR is a regression modelling technique that regresses spatially-explicit variables onto monitored pollutant concentrations within a GIS system. More specifically, LUR models associate pollutant concentrations, such as NO_2 to site specific geographical characteristics, e.g., topography, land use, traffic, population density, altitude and meteorological parameters. The use of these variables in the regression model are known to capture small scale variability (Ryan and LeMasters, 2007). Rehman et al. (2017) developed a LUR model in Brisbane Australia to predict NO_2 and NOx during 2009 – 2012. The model was able to explain 64 % and 70 % variations in NO_2 and NOx, respectively. Distance to major roads and industrial areas were the common predictor variables for both NO_2 and NOx, suggesting an important role for road traffic and industrial emissions. Rehman et al. (2017) used the following independent variables in their model: distance to coast (km), distance to port (km), distance to airport (km), distance to nearest major road (km), distance to nearest minor road (km), major road length (km), minor road length (km), population density (person/km^2), land use by type (km^2) and elevation (m). Muttoo et al. (2017) used several geographic predictor variables to predict NOx concentrations in Durban South Africa employing a land use multivariate regression model. They used length of minor roads within a 1000 m radius, length of major roads within a 300 m radius and area of open space within a 1000 m radius in the model as independent variables. The LUR model was able to explain 73% variance in NOx concentrations, however cross validation resulted in R^2 value of 0.59.

Wang et al. (2014) developed a European and regional level LUR model for modelling NO_2 and PM. The LUR model used 17 PM and 23 NO_2 ESCAPE (European Study of Cohorts for Air Pollution Effects) study areas across 14 European countries. The LUR model explained 56% of the concentration variability across all sites for NO_2, 86% for $PM_{2.5}$, and 70% for $PM_{2.5}$ absorbance at European level. Molter et al. (2010, Part I & II) developed a LUR model in Great Manchester, UK to model NO_2 and PM_{10} and used the results of dispersion modelling (ADMS) for training the model instead of monitored data. This approach provides relative more number of sites as compared to monitored data. They have analysed both spatial and temporal variability in air pollution concentrations. Traffic intensity, emissions, land-use and physical geography were used as predictors in the models. He et al. (2018) have provided a review of the LUR models and listed the more commonly used predictors, viz. pollutant data, land use classification (e.g., residential, industrial, urban green space, street morphology, aspect ratio, traffic data (e.g., number and types of vehicles, railways network, road network), census data (e.g., houses density, population density), meteorology (e.g., wind speed, temperature, relative humidity), topography (e.g., altitude, slope angle), emission data, and remote sensing satellite data.

Hoek et al. (2008) has provided a detailed review on LUR models, identifying 25 studies on the subject. They have identified several significant predictors for LUR models, including various traffic characteristics (e.g., traffic flow, road length, distance to major and minor roads), population characteristics (e.g., number of houses, housing density), land use (e.g., urban, open space, industry, commercial), physical geography (e.g., altitude, distance to sea), and climatic conditions (e.g., wind speed, temperature, relative humidity). For details see Briggs et al. (1997), Stedman et al. (1997), Beelen et al. (2007), Ryan et al. (2007), and Hoek et al. (2008).

Gillespie et al. (2016) developed a LUR model to estimate exposure to NO_2 in Glasgow, Scotland. They used 135 NO_2 passive diffusion tubes, which were divided to four groups (32 – 35 sites per group) and models were developed using a combination of 1 to 3 groups as training sites to assess how the number of training sites affected the model performance. The explanatory variables used in the models were major road length, minor road length, all urban areas, building volume, distance to nearest major or minor road, green rural area, minor road length, and street configuration. The models were able to explain moderate to high variance in the data, where R^2 ranged from 0.62 to 0.89 for training dataset and 0.44 to 0.85 for hold-out dataset. Precision of estimated exposure was increased with increasing number of training sites. Gillespie et al. (2016) concluded that use of more than 60 training sites had a considerable beneficial effect on model performance.

Mostly, LUR approaches are applied in small or large urban areas (e.g., Beelen et al., 2013; Ryan and LeMasters, 2007), however, some researchers have also applied LUR to entire countries (e.g., Vienneau et al., 2010; Beelen et al., 2007; Stedman et al., 1997), who applied LUR models in Netherland and UK. Vienneau et al. (2010) developed land use regression models and compared their outputs in Great Britain and Netherland. The predictor variables included characteristics of traffic, population, land use and topography. They developed a common model for Great Britain and Netherland, which performed well with adjusted R^2 0.63 to predict NO_2 concentrations. However, the model developed specifically for each country improved the model predictability. Furthermore, the model based on common data from both countries showed slightly lower performance than the model using local data from within each country. Models developed in one country and transferred to the other country showed substantially worse performance than the country specific model. Therefore, they concluded that much care should be taken in transferring models from one area to another and in developing an LUR model for predicting air pollution in a large urban area.

Several authors have reported the potential of LUR modelling techniques for improving spatial and temporal coverage of air pollution in urban areas. However, further works is required to improve the LUR modelling performance. This is possible by: (a) Improving air quality monitoring network (AQMN) – historically AQMN were sparse which did not provide enough monitoring data for training LUR model. More recently, low-cost sensors have enabled researchers to design purpose-designed networks, which have much greater number of monitoring sites. Increasing the number of monitoring stations could provide better spatial coverage and hence improvement in LUR model performance. (b) Data fusion techniques – fusing measured and modelled data can provide better quality and better spatial coverage of air pollution in urban areas, which can lead to better LUR modelling results. (c) Advance modelling techniques – the LUR modelling reviewed above used multiple linear regression model, which is based on several assumptions not met by air quality and meteorology data. More advanced nonlinear regression approaches (e.g., GAM and BRT) can be applied to improve the performance of LUR modelling. The intention in this project is to set up a dense network of AQ sensors in Sheffield, which could provide a better platform for training the LUR model employing advanced regression and machine learning approaches.

2.3. Air quality data fusion techniques

Integrating and merging data and information from several sources is known as data fusion (Castanedo, 2013). In literature, data fusion is also referred to as decision fusion, data combination, data aggregation, multisensor data fusion and sensor fusion. Data fusion originated in 1970 in the US and has several benefits, e.g., it is cost-effective and convenient, it minimises the need for primary research to collect further data thus reducing cost, and it fits into existing analysis (Nielsen, 2007). Data fusion integrates data from numerous sensors which have different physical characteristics and thus provides a basis for decision making, planning and control of intelligent machines (Meti and Sangam, 2005). Furthermore, data that have been fused from several sources help achieve inferences that are unfeasible from an individual source. Also, data fusion helps achieve data from various information sources in such a way that it provides better performance than when each information source is used alone (Meti and Sangam, 2005). Data fusion techniques merge observed data with modelled data in a mathematical objective way, adding value to both observed and modelled data. The observed data are improved by filling spatiotemporal gaps, whereas the modelled data are improved by constraining it with observed data (Schneider et al., 2015). Therefore, data fusion of observed data with modelled data can improve urban scale air quality mapping.

Schneider et al. (2017) reported that the accuracy of fused data normally depends on a range of factors, which include (a) the total number of observations, (b) spatial distribution of the network, (c) uncertainty of the measured data, and (d) the ability of the model to accurately predict air pollutant levels with high spatial and temporal resolution. Son (2002) has classified data fusion methods into 3 main clusters: (1) Least square-based methods, e.g., optimal theory, Kalman filtering, uncertainty ellipsoids and regularization; (2) Probabilistic methods, e.g., evidence theory, Bayesian analysis of values of sensor, recursive operations and robust statistics; and (3) Intelligent aggregation methods, e.g., genetic algorithms, neural networks and fuzzy logic. Meti and Sangam (2005) reviewed data fusion approaches and focused on four approaches, viz., fuzzy logic, artificial neural networks, wavelets transform and image fusion.

Below some of the data fusion approaches are reviewed in light of the current literature.

Geostatistics is one of the most widely used techniques for data fusion in spatial analysis. Olea (1999) has defined geostatistics as a collection of numerical algorithms that characterises spatial attributes in a similar way that time series characterises temporal data. Basic components of geostatistics are (a) Variogram analysis (characterisation of spatial correlation), (b) Kriging (optimal interpolation which generates best linear unbiased estimate employing semivariogram model), and (c) Stochastic simulation (also employs semivariogram model to generate multiple equiprobable images of the variables) (Olea, 1999).

Schneider et al. (2017) employed geostatistics methodology to fuse data obtained from a network of low-cost sensors and EPISODE dispersion model. EPISODE is a 3-D Eulerian/Lagrangian dispersion model that provides atmospheric pollutants forecast at urban and regional scales. For more details on EPISODE model see Slordal et al. (2003). Schneider et al. (2017) evaluated the geostatistics methodology for using both measured and predicted data of NO_2 in Oslo, Norway during January 2016. The results showed that the fusing methods were able to produce realistic hourly NO_2 concentrations which inherit spatial trend of the pollutant from the EPISODE model. Furthermore, fused data were compared to measured data from a reference instrument and results showed reasonably good resemblance between measured and fused data with R^2 value of 0.89 and mean squared error of 14.3 $\mu g/m^3$.

One of the basic and probably most widely used approach for interpolation in geostatistics is Kriging. Kriging interpolates air pollutant levels (e.g., NO_2 concentrations) between two values by modelling them with Gaussian process. Schneider et al. (2017) used the universal kriging techniques to combine modelled and measured values of NO_2. Universal kriging is more advance and performs the data fusion by predicting NO_2 concentrations through the interpolation of the observed concentrations and uses the model data for spatial trend. Universal kriging through the usage of one or more independent variables makes the overall mean as non-constant. Universal kriging is mathematically equivalent to regression kriging or residual kriging (Hengl et al., 2007) but can perform both spatial interpolation and the linear regression in a single step.

Hsu et al., (2017) reviewed different methods used for interpolation and divided them into two categories: Geostatistical techniques and Non-geostatistical methods. The geostatistical techniques include Ordinary Kriging, Universal Kriging, Simple Kriging, Empirical Bayesian Kriging and Original CoKriging. The non-geostatistical techniques include Splines, Trend Surface Analysis, Inverse Distance Weighting and Natural Neighbor. According to the findings of Hsu et al. (2017) the geostatistical interpolations showed better prediction than the non-geostatistical interpolation techniques.

Liang et al. (2017) developed a data fusion method and compared the outputs with Kriging-with-external-drift (KED) and Chemistry module from WRF-Chem (Weather Research and Forecast Model with Chemistry Module). KED is a type Universal Kriging which takes into account the local trend of the variable (e.g. air pollutants concentrations) as well as external drift (a spatial trend) when minimising the variance of estimation (Hengl et al., 2003). Both KED and WRF-Chem were used to estimate daily $PM_{2.5}$ levels in 10 km grid cells in North

China during 2013. The estimated concentrations from both KED and WRF-Chem were then fused with measured observations. For fusion a simple linear regression model was applied between the observed and estimated concentrations for both KED and chemistry models in turns. The regression coefficients obtained from the regression model were used to adjust the predicted concentrations. The performance of the models was evaluated. KED and data fusion methods showed better performance in terms of R^2 value of 0.95 and 0.94, respectively as compared to 0.51 value for WARF-Chem model.

2.4. Various types of air quality sensors

Air quality monitoring is important to promote air quality awareness and to support abatement strategies (Borrego et al., 2016). Several techniques are used to monitor air quality (Penza et al., 2014), which include (a) Reference air quality monitoring stations; (b) Portable air quality monitors; (c) Passive diffusion tubes; and (d) Low-cost sensors. Reference air quality monitoring is used for air quality compliance purpose, studying exposure, supporting air quality management and developing policy for cutting emissions. Reference instruments are the most expensive to purchase and maintain, therefore their density is normally very low and not dense enough for detailed spatial mapping in urban areas. Portable monitors are either carried by individuals or installed on vehicles to be parked in places where fixed sensors are not possible to be installed. Portable instruments can be useful for monitoring air quality in certain cases and provide high-resolution temporal data for shorter period of time, but have limited application for spatial maps. Passive tubes are small tubes used for monitoring gaseous air pollutants such as NO_2 and provide long-term averaged data (e.g., annual average concentrations). Diffusion tubes are the cheapest technique and potentially can provide better spatial coverage, however, these can be used only for gaseous air pollutants and are unable to provide daily and hourly concentrations. Low-cost sensors are used to collect real-time air quality data providing high resolution spatial and temporal air quality data. This is the new trend in air quality monitoring and can support the conventional air quality monitoring stations (Heimann et al., 2015; Van den Bossche et al., 2015; Viana et al., 2015). The low-cost sensors use the latest microsensing technology and are considered the innovative tools for air quality monitoring (Castell et al., 2015; Snyder et al., 2013; Kumar et al., 2015; Stojanovic et al., 2015). Data collected by the low-cost sensors can be used for detailed spatial mapping of air pollution, especially over small areas such as an urban area or a part of it, for atmospheric model validation, and for assessing population exposure. However, the quality of the data produced by the low-cost sensors, their robustness to the environmental conditions, comparison with reference techniques, and comparison of low-cost sensors produced by difference manufacturers need further investigation over large spatial and temporal range.

Table 2.1. Comparison of various air quality monitoring sensors

Sensors type	Pollutant monitored	Power	Pro and cons
Reference sensors	Gases and particles	Mains	Expensive, most accurate, coarse spatial resolution due to high purchase price, large size and heavy
Diffusion tubes	Only gases	No need	Coarse temporal resolution, cheaper, high spatial resolution, low quality data
Low-cost sensors	Both gases and particles	Battery, solar or mains	Low quality data, cheaper, high temporal and spatial resolution.

2.4.1. Low-cost air quality sensors

Several authors have analysed the performance of the low-cost sensors in comparison to reference methods and to other low-cost sensors. Borrego et al. (2016) performed an assessment of the low-cost sensors in comparison to reference instruments in Aveiro (Portugal) from 13th to 27th October 2014. The instruments both low-cost sensors and reference instruments installed side by side were used to monitor the levels of gaseous pollutants (e.g., CO, NOx, O_3, SO_2), particulate matter (PM_{10}, $PM_{2.5}$) and meteorological parameters (e.g., temperature, wind speed and direction, relative humidity, solar radiation and precipitation) and their measurements were mutually compared. Different sensors showed significantly different performance in terms of the statistical metrics used for evaluating the sensors performance. The range of R^2 values for different air pollutants were: O_3 (0.12-0.77), CO (0.53-0.87), NO_2 (0.02-0.89), PM (0.07-0.36) and SO_2 (0.09-0.20), where lower R^2 value shows poor performance of the low-cost sensors. Borrego et al. (2016) concluded that low-cost sensors have great potentials for air quality monitoring if properly supported by post processing and data modelling tools.

Different sensor-systems use different principles to measure the concentrations of atmospheric pollutants (Borrego et al., 2016), which include optical particle counters (OPC), metal oxide semiconductor sensors (MOS), electrochemical sensors (EC), nondispersive infrared sensors (NDIR) and photo-ionisation detection sensors (PID). Aleixandre and Gerboles (2012) reported that low-cost air quality sensors (LCS) works through either measuring the electrochemical interaction between the sensing materials and the atmospheric chemicals or through absorption of visible light. Principle of light scattering or absorption is used for measuring the levels of PM.

Individual sensors are usually integrated into a platform of sensors known as sensor node. Sensor node contains a sensor board, the sensors, and a control board, which integrates all the necessary parts such as GPS, data storage, communication ports and signal conditioning. Some of the well-known LCS sensors and their manufacturers (developers) are provided in Table 2.2. For more details on these sensors see chapter 3.

Table 2.2. Various brands of low-cost sensors (LCS) and their specifications

Sensor name	Pollutant measured	Sensor type	Manufacturer
Cambridge university SNAQ[1]	Gaseous pollutants, particulate matter	Electrochemical, OPC	University of Cambridge, UK
AUTh-ISAG sensors[2]	Gaseous pollutants (e.g., NO_2 and O_3)	Metal oxides	Libellium
ECN – Airbox[3]	Particulate matter (e.g., UFP, PM_1, $PM_{2.5}$, and PM_{10}), gaseous (e.g., NO_2 and O)	OPC and electrochemical	Energy research Centre of the Netherlands
NanoEnvi platform[4]	Gaseous pollutants (e.g., SO_2, NO, NO_2, CO, CO_2, O_3, H_2S and VOCs), particulates (PM_{10} and $PM_{2.5}$)	OPC, metal oxides and electrochemical	Envira, Spain
AQMesh sensors[5]	Gaseous pollutants (NO, NO_2, O_3, CO) and particles (PM_{10}, $PM_{2.5}$, PM_1)	OPC and electrochemical	Environmental Instruments Ltd, UK
ENEA air sensor[6]	Gaseous pollutants and particles	OPC and electrochemical	ENEA (Energia Nucleare ed Energie Alternative), Italy
EveryAware sensor box[7]	CO, SO_2, NO_2, O_3, and PM_{10}	OPC and electrochemical	Vito, Belgium
Earthsense Zephyrs[8]	NO, NO_2, O_3 and PM_{10} and $PM_{2.5}$	OPC, electrochemical	EarthSense, UK
Envirowatch E-MOTEs[9]	Gaseous pollutants (NO, NO_2, CO)	electrochemical	Envirowatch Newcastle, UK

1. Mead et al. (2013); Popoola et al. (2013); Borrego et al. (2016)
2. Borrego et al. (2016)
3. Borrego et al. (2016); Hamm et al. (2016)
4. https://enviraiot.com/nanoenvi-and-its-applications-for-air-quality-control/
5. Borrego et al. (2016); Carruthers et al. (2016)
6. Suriano et al. (2015)
7. Borrego et al. (2016); Dongol (2015)
8. https://www.earthsense.co.uk/zephyr
9. http://www.envirowatch.ltd.uk/

2.4.2. Uncertainties in low-cost sensor measurements

Low-cost air quality sensors are cheaper, compact, user-friendly and provide high-resolution spatiotemporal maps of air pollutant concentrations. These sensors have the potential to enhance the existing air quality network run by the local authorities on local levels or by DEFRA at the national levels. In addition, these sensors can be installed independently by various research and government organisations to monitor public exposure to various air pollutants within a specific area. However, despite all these positive points, the quality of air pollution data collected by the low-cost sensors is questionable. There is a need for further investigation to quantify uncertainties in the low-cost sensor datasets. These uncertainties are related to the degree of harshness of the environmental conditions, especially extreme temperature and relative humidity and the time interval i.e. for how long these instruments work in the outdoor environment. Furthermore, uncertainties are also affected by the measuring principles of the sensors and the quality of the materials used by the manufacturers. Therefore, intercomparison of low-cost sensors made by different manufacturers and with reference instruments is required. Below a brief literature review of such studies is provided.

Castell et al. (2016) have evaluated the performance of the AQMesh pods measuring gaseous air pollutants (e.g., NOx, CO, and O_3) and particulate matter (PM_{10} and $PM_{2.5}$) in Oslo, Norway. They performed the evaluation of both outdoor and in laboratory conditions. They considered several types of emissions and environmental conditions such as roadside traffic conditions and urban background conditions over a 6 month's period (April to September, 2015). Castell et al. (2016) concluded that good performance of the low-cost sensors in the laboratory conditions does not imply such performance in the outdoor conditions. Therefore, to reduce uncertainties, sensors must be calibrated in outdoor field conditions. They also concluded that there is a lack of adequate outdoor testing of the sensors by the manufacturers before marketing such sensors, which can lead to poor performance and low-quality data, which is of great concern, especially when public use such instruments by themselves to collect and interpret air quality data.

Borrego et al. (2016) compared the performance of several low-cost sensors with reference instruments and reported that for measuring O_3 AQMesh and NanoEnvi sensors had the lowest errors and higher coefficient of determination ($R^2 > 0.70$), whereas ENEA Air-Sensors, ISAG and Cambridge SNAQ showed poor performance with $R^2 < 0.2$. To measure the levels of NO_2, Borrego et al. (2016) compared the performance of 6 senosrs, where the highest correlation and lowest errors were shown by AQMesh, ECN-Airbox and Cambridge University SNAQ with $R^2 > 0.80$ and mean biased error (MBE) close to zero. In contrast, ENEA Air-Sensors and AUTh-ISAG sensors demonstrated very poor correlation ($R^2 < 0.1$). For measuring the levels of CO, AQMesh and Cambridge University SNAQ had the highest correlation ($R^2 > 0.80$) with reference instruments, whereas the performance of the rest of the sensors was also satisfactory ($R^2 > 0.50$) (Borrego et al., 2016). For monitoring NO, AQMesh and Cambridge University SNAQ were compared, where AQMesh showed better correlation ($R^2 = 0.80$) than Cambridge University SNAQ ($R^2 = 0.30$). For measuring PM_{10}, all low-cost sensors showed poor correlation with reference instruments, where $R^2 = 0.36$ being the highest was observed for ECN Air-box (Borrego et al., 2016). ECN Air-box also showed the highest R-squared ($R^2 = 0.27$) with reference instruments for measuring $PM_{2.5}$, the other sensors showed even lower R^2-value.

Castell et al. (2017) compared the measurements from 24 commercial low-cost sensors AQMesh against reference instruments and reported that the quality of the data obtained from the low-cost sensors were questionable. The performance of the sensors varied both spatially and temporally and is dependent on the atmospheric composition and meteorological conditions, such as temperature and relative humidity. Furthermore, Castell et al. (2017) reported that the performance varied from unit to unit, therefore it is necessary to check the data quality of each unit separately before use. The sensors installed in the laboratory showed much stronger correlation ($R^2 > 0.95$ for all pollutants) with reference instruments, than those installed outdoor, where average R^2 values were 0.60, 0.86, 0.49, 0.54, 0.56 and 0.51 for CO, NO, NO_2, O_3, PM_{10} and $PM_{2.5}$, respectively. Air quality data collected by using low-cost sensors are suitable for promoting air quality awareness, general information and for highlighting air pollution hotspots, however, the data are not suitable for air quality compliance and research, especially for assessing health and environmental impacts of air pollution (Castell et al., 2017). Dongol (2015) has also concluded that air quality data collected by low-cost sensors cannot be used for air quality regulatory purposes and for other purposes where high accuracy of data is required. Therefore, there is a need for further legislation to regulate the usability of data obtained from low-cost sensors (Lewis and Edward, 2016).

Referring to the uncertainties in air quality data collected by low-cost sensors, Lewis and Edward (2016) have commented that the recent introduction of these sensors for monitoring public exposure to air pollution are generating a high volume of data, which remain mostly untested, and therefore their quality is questionable and will become a headache for air quality managers and planners in the future. Furthermore, Lewis and Edward (2016) mentioned that microsensors show stability and sensitivity issues and that the sensors readings are interfered by other long-lived air pollutants, e.g., CO_2 and H_2O and meteorological conditions like relative humidity, temperature and wind speed. The low-cost sensors performs better when air pollutant levels are high (Lewis and Edward, 2016).

2.4.3. Recommendations for future improvement of air quality sensors

The low-cost microsensors have potential to measure air pollutant levels in places where traditional monitoring is not possible. The low-cost microsensors are portable, cheaper, and can provide much better spatial and temporal coverage in real time providing more localized and timely warning to the public, however, the data obtained from these sensors are questionable and therefore further work is required to improve their quality.

The performance of the low-cost sensors vary from place to place, therefore, local outdoor calibration is needed. Frequent calibrations of sensors improve their correlation with reference instruments (Smith et al., 2017). Lewis et al. (2016) have shown that one potential solution to reduce the uncertainties of air quality data obtained by using low-cost sensors is the application of supervised machine learning techniques, for instance Boosted Regression Tree (BRT) model. Spinelle et al. (2017) applied three approaches for calibrating the concentration of NO_2, CO, and CO_2. The methods were simple linear regression, multiple linear regression and supervised machine learning technique (artificial neural network). In the simple linear regression only the reference concentration was used as an explanatory variable, whereas in the other models relative humidity and temperature were also used. Supervised learning technique showed better performance than the other two models.

Experimental design should be further improved to get better results. Studies (e.g., Castell et al., 2017) have shown that the performance of the low-cost sensors degraded with time. Therefore, shorter terms experiments should be designed which are likely to be less affected as compared to long-term experiments. Moreover, the sensors ideally should be used in locations and times where they are less likely to be exposed to extreme environmental conditions of temperature and relative humidity. Furthermore, using a cluster of sensors, in contrast to individual sensor, can significantly improve the data quality (Smith et al., 2017).

2.5. Research gaps

The aim of this chapter was to carry out a detailed literature review and identify research gaps related to urban air quality modelling to be addressed in this project. The research gaps identified are listed below:

- Since the emergence of low-cost sensors, several authors have deployed these sensors in urban areas. However, no formal approach has been used to identify the locations for sensors installation. Furthermore, the current air quality monitoring networks are structured using a single criterion (e.g., potential air pollution hotspots). To address this, in this project we aim to develop a formal approach for deploying a dense network of AQ sensors based on multiple criteria using sensors of different grades. This is related to **objective 1** of the project.
- From the literature review it was found that previously most of the sensor calibration studies whether in the laboratories or outdoor fields were carried out for a short period of time. Therefore, further research is required to compare the performance of low-cost sensors to each other and to reference sensor in outdoor fields for a longer period of time at least for a year, employing a nonlinear regression model. This is related to **objective 2.**
- The literature review showed that previously several approaches have been employed for modelling the spatial variability of air pollutants in urban area. However, little was done on comparing different modelling approaches for analysing the spatial variability of pollutants in urban areas. To address this gap, in this project several modelling techniques will be used and their performance will be compared. This is related to **objective 3** of the project.
- Literature review revealed that the application of data fusion techniques improve the quality of data, however, little is done to demonstrate how these techniques help in improving the spatial modelling of air quality in urban areas. This is related to **objective 4**, which aims to improve both model estimations and low-cost sensor measurements.
- How the concentrations of NO_2 measured by LCS and reference sensors vary on different temporal scales in urban areas, and what is the best time series model for modelling NO_2 concentrations. This is related to **objective 5** of the project.

2.6. Summary of the chapter

In this chapter a detailed review of different air quality modelling and monitoring was carried out. The aim was to find out as to what are the state of the art for AQ modelling and low-cost sensors in light of the current literature. Air quality modelling techniques are mainly divided into two main categories: Dispersion air quality modelling and statistical modelling. Dispersion modelling are further divided into: (a) Gaussian, (b) Lagrangian, (c) Eulerian, (d) Photochemical, and (e) Hybrid modelling. State of the art dispersion pollution modelling approaches are hybrid models, which incorporate more than one of these approaches to model air quality in urban areas. Statistical modelling techniques can be further divided into four main sub-groups which include time series, linear regression, non-linear regression and land-use regression. In time series modelling NARMAX is probably the more advance technique, which is able to handle nonlinearities in air pollutant concentrations and has the ability to utilise external parameters such as meteorological parameters and can be an effective tool for air quality modelling at urban scale. In regression models, LUR has the potential to model the spatiotemporal variability of air pollutants in urban areas using traffic, land-use and topographical characteristics. Literature review showed that in most of the previous studies linear LUR models had been developed. In this project a nonlinear LUR model will be developed that has the ability to handle non-normal distribution of air pollutant concentrations and address nonlinearities in the association of air pollutants with other controlling parameters like emission sources and land-use.

Low-cost sensors have the potential to improve traditional air quality monitoring programme as these sensors are cheap, compact, user-friendly and provide high-resolution spatiotemporal measurement of air pollutant concentrations. However, these sensors have their limitations and the data obtained from these sensors are questionable. The data from the low-cost sensors can be used for highlighting air pollutant hotspots, for public awareness and for complimenting traditional air quality monitoring programmes, however the data are not suitable for regulatory purpose and for assessing the air quality compliance with air quality standards. Therefore, further work is required to improve the performance of these sensors by developing their technology to make it more robust, by frequent calibration both in laboratory and outdoor and by improving the experimental designs.

The literature review was helpful in identifying research gaps to be addressed in this research project, which included structuring a multipurpose air quality monitoring network, outdoor calibration of low-cost sensors, comparing different approaches for modelling the spatial variability of air pollutants in urban areas, improving of spatial modelling estimates and low-cost sensor measurements by using data fusion techniques, and analysis of the temporal variability of air pollutant concentrations. These research gaps have been addressed in the coming chapters of the project. The next chapter addresses objective 2, which aim to present nonlinear calibration models for several gaseous pollutants measured by Envirowatch E-MOTEs in Sheffield.

2.7. References

1. Adhikari, R., Agrawal, A.K., 2013. An Introductory Study on Time Series Modeling and Forecasting. LAP Lambert Academic Publishing.
2. Aldrin, M., Haff, I.H., 2005. Generalised additive modelling of air pollution, traffic volume and meteorology. Atmospheric Environment 39, 2145–2155.
3. Aleixandre, M., Gerboles, M., 2012. Review of small commercial sensors for indicative monitoring of ambient gas. Chem. Eng. Trans. 30, 169–174.
4. Andersen, S. B., Weatherhead, E. C., Stevermer, A., Austin, J., Brühl, C., Fleming, E. L., de Grandpré, J., Grewe, V., Isaksen, I., Pitari, G., Portmann, R. W., Rognerud, B., Rosenfield, J. E., Smyshlyaev, S., Nagashima, T., Velders, G. J. M., Weisenstein, D. K., & Xia, J., 2006. Comparison of recent modelled and observed trends in total column ozone, Journal of Geophysical Research, 111, D02303. doi:10.1029/2005JD006091.
5. Arnold, S.R., M.P. Chipperfield, and M.A. Blitz, 2005. A three-dimensional model study of the effect of new temperature-dependent quantum yields for acetone photolysis, Geophysic Research, 110 (D22), D22305 doi:10.1029/2005JD005998.
6. Baur, D., Saisana, M. and Schulze, N., 2004. Modelling the effects of meteorological variables on ozone concentration-a quantile regression approach. Atmospheric Environment, 38 (28): 4689 – 4699.

7. Beelen, R., Hoek, G., Fischer, P., van den Brandt, P.A., Brunekreef, B., 2007. Estimated long-term outdoor air pollution concentrations in a cohort study. Atmos. Environ. 41, 1343–1358.
8. Beelen, R., Hoek, G., Vienneau, D., Eeftens., M., Dimakopoulou, K., Pedeli, X., Tsai, M.Y., 2013. Development of NO2 and NOx land use regression models for estimating air pollution exposure in 36 study areas in Europe e-The ESCAPE project. Atmospheric Environment 72 (2013) 10 - 23.
9. Berastegi, G.I., Madariaga, I., Elias, A., Agirre, E. and Uria, J., 2001. Long term changes of O3 and traffic in Bilbao. Atmospheric Environment, 35, (2001): 5581 – 5592.
10. Bluett, J., Gimson, Fisher, G., and Heydenrych, C. 2004. Ministry for the Environment. Good Practice Guide for Atmospheric Dispersion Modelling. National Institute of Water and Atmospheric Research, Aurora Pacific Limited and Earth Tech Incorporated for the Ministry for the Environment, New Zealand.
11. Borrego, C., Costa, A.M., Ginja, J., Amorim, M., Coutinho, M., Karatzas, K., Sioumis, Th., Katsifarakis, N., Konstantinidis, K., De Vito, S., Esposito, E., Smith, P., Andre, N., Gerard, Francis, P., Castell, L.A., Schneider, N., Viana, P., Minguillon, M., Reimringer, M.C., Otjes, W., von Sicard, R.P., Pohle, O., Elen, R., Suriano, B., Pfister, D., Prato, V., Dipinto, M., Penza, M., 2016. Assessment of air quality microsensors versus reference methods: The EuNetAir joint exercise. Atmospheric Environment 147: 246-263.
12. Brasseur, G. P., Hauglustaine, D. A., Walters, S., Rasch, P. J., Muller, J.-F., Granier, C., and Tie, X.-X.: MOZART, 1998. A global chemical transport model for ozone and related chemical tracers, Part 1: Model description, Journal of Geophysical Research, 103, 28265–28289.
13. Briggs, D.J, 2005. The Role of GIS: Coping With Space (And Time) in Air Pollution Exposure Assessment. J. Toxicol. Environ. Health (A) 2005, 68(13-14), 1243-61.
14. Briggs, D.J., Collins, S., Elliott, P., et al., 1997. Mapping urban air pollution using GIS: a regression-based approach. Int. J. Geogr. Inf. Sci. 11, 699–718.
15. Briggs, G.A., 1965. A plume rise model compared with observations *J. Air Poll. Control Association* 15:433.
16. Brunt, H., Barnes, J., Longhurst, J.W.S., Scally, G., Hayes, E., 2016. Local air quality management policy and practice in the UK: The case for greater public health integration and engagement. Environmental Science & Policy 58, 52–60.
17. Cade, B.S. and Noon, B.R., 2003. A Gentle Introduction to Quantile Regression for Ecologists. *Front. Ecol. Environ.* 1: 412–420.
18. Carruthers, D., Clarke, D., Dicks, K.J., Freshwater, R.A., Jackson, M., Jones, R.L., Lad, C., Leslie, I., Lewis, A. J., Lloyd, H., Popoola, O.A.M., Randle, A., Ulrich, S., 2016. Using a commercial low-cost sensor network (AQMesh) to quantify urban air quality: comparing measured and modelled (ADMS-urban) pollutant concentrations. Air Quality Monitoring: Evolving Issues and New Technologies Conference, December 13th, 2016, organised by Automation and Analytical Management Group - Royal Society of Chemistry.
19. Carslaw, D.C., Beevers, S.D., and Tate, J.E., 2007. Modelling and assessing trends in traffic-related emissions using a generalized additive modelling approach. Atmospheric Environment 41, 5289–5299.
20. Carslaw, D.C., Taylor, P.J., 2009. Analysis of air pollution data at a mixed source location using boosted regression trees. Atmos. Environ., 43 (22): 3563 - 3570.
21. Carslaw, D.C., Williams, M.L., Tate, J.E. and Beevers, S.D., 2013. The Importance of High Vehicle Power for Passenger Car Emissions. Atmos. Environ. 68: 8–16.

22. Castanedo, F., 2013. Article A Review of Data Fusion Techniques. The Scientific World Journal, 2013, Article ID 704504. http://dx.doi.org/10.1155/2013/704504.
23. Castell, N., Dauge, F.R., Dongol, R., Vogt, M., Schneider, P., 2016. Uncertainty in air quality observations using low-cost sensors. Geophysical Research Abstracts, 18: 5692. EGU General Assembly 2016.
24. Castell, N., Dauge, F.R., Schneider, P., Vogt, M., Lerner, U., Fishbain, B., Broday, D., Bartonova, A., 2017. Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? Environment International 99 (2017): 293 – 302.
25. Castell, N., Kobernus, M., Liu, H.Y., Schneider, P., Lahoz, W., Berre, A.J., Noll, J., 2015. Mobile technologies and services for environmental monitoring: the Citi-Sense-MOB approach. Urban Clim. 14 (3), 370-382.
26. CERC, 2017. Cambridge Environmental Research Consultancy Ltd. URL: http://www.cerc.co.uk/environmentalsoftware/ADMS-Urban-model.html.
27. Chen, S., Billings, S. A., 1989. Representation of non-linear systems: the NARMAX model. International Journal of Control, 49 (3), 1012-1032.
28. Cochrane, J.H., 1997. Time Series for Macroeconomics and Finance, Graduate School of Business, University of Chicago, spring 1997.
29. COMEAP, 2009. Long-term exposure to air pollution: effect on mortality. the committee on the medical effects of air pollutants.
30. COMEAP, 2010. The mortality effects of long-term exposure to particulate air pollution in the United Kingdom. The committee on the medical effects of air pollutants.
31. Daly, A., Zannetti, P., 2007. Air Pollution Modeling – An Overview. Chapter 2 of AMBIENT AIR POLLUTION (P. Zannetti, D. Al-Ajmi, and S. Al-Rashied, Editors). Published by The Arab School for Science and Technology (ASST) (http://www.arabschool.org.sy) and The EnviroComp Institute (http://www.envirocomp.org/).
32. Dacre, H. F., Grant, A. L. M., Hogan, R. J., Belcher, S. E., Thom-son, D. J., Devenish, B., Marenco, F., Hort, M. C., Haywoood, J. M., Ansmann, A., Mattis, I., and Clarisse, L., 2011. Evaluating the structure and magnitude of the ash plume during the initial phase of the 2010 Eyjafjallajokull eruption using lidar observa-tions and NAME simulations, J. Geophys. Res., 116, D00U03, doi:10.1029/2011JD015608.
33. DEFRA, 2017. Improving air quality in the UK: tackling nitrogen dioxide in our towns and cities, May 2017, Draft UK Air Quality Plan for tackling nitrogen dioxide. Available on: https://consult.defra.gov.uk/airquality/air-quality-plan-for-tackling-nitrogen-dioxide/supporting_documents/Draft%20Revised%20AQ%20Plan.pdf (accessed 05/10\2017).
34. Diaz-Robles, L.A., Ortega, J.C., Fu, J.S., Reed, G.D., Chow, J.C., Watson, J.G., Moncada-Herrera, J.A. A., 2008. Hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile. Atmos. Environ., 42, 8331–8340.
35. Dongol, R., 2015. Evaluation of the Usability of Low-cost Sensors for Public Air Quality Information. Master’s Thesis, Department of Informatics Programming and Networks, University of Oslo.
36. Duenas, C., Fernandez, M.C., Canete, S., Carretero, J., Liger, E., 2002. Assessment of ozone variations and meteorological effects in an urban area in the Mediterranean Coast. Science of The Total Environment, 299 (1-3):97 – 113.

37. Eeftens, M., Beelen, R., de Hoogh, K., et al., 2012. Development of land use regression models for PM(2.5), PM(2.5) absorbance, PM(10) and PM(coarse) in 20 European Study areas; results of the ESCAPE project. Environ Sci Technol.
38. El-Harbawi, M., 2013. Air quality modelling, simulation, and computational methods:a review. Environ. Rev. 21: 149–179. dx.doi.org/10.1139/er-2012-0056.
39. Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Animal Ecol. 77 (4), 802 - 813.
40. Gardner, M.W., and Dorling, S.R., 2000. Statistical surface ozone models: an improved methodology to account for non-linear behaviour. Atmospheric Environment, 34: 21 – 34.
41. Gillespie, J., Beverland, I.J., Hamilton, S., Padmanabhan, S., 2016. Development, Evaluation, and Comparison of Land Use Regression Modeling Methods to Estimate Residential Exposure to Nitrogen Dioxide in a Cohort Study. Environ. Sci. Technol. 2016, 50, 11085−11093.
42. Goyal, P. Chan, A.T., Jaiswal, N., 2006. Statistical models for the prediction of respirable suspended particulate matter in urban cities. Atmos. Environ., 40, 2068–2077.
43. Hamm, N.A.S., Van Lochem, M., Hoek, G., Otjes, R.P., Van der Sterren, S., Verhoeven, H., 2016. The Invisible Made Visible: Science and Technology. link. springer.com/book/10.1007/978-3-319-26940-5.
44. Hao, L. and Naiman, D.Q. (2007). Quantile Regression: Series-Quantitative Applications in the Social Sciences, Sage Publications, 7: 149.
45. Hastie, T.J., and Tibshirani, R.J., 1990. Generalised Additive Models. Chapman & Hall, London.
46. He, B., Heal, M.R., Reis, S., 2018. Land-Use Regression Modelling of Intra-Urban Air Pollution Variation in China: Current Status and Future Needs. *Atmosphere, 9*(4), 134; doi:10.3390/atmos9040134.
47. Heimann, I., Bright, V.B., McLeod, M.W., Mead, M.I., Popoola, O.A.M., Stewart, G.B., Jones, R.L., July 2015. Source attribution of air pollution by spatial scale separation using high spatial density networks of low cost air quality sensors. Atmos. Environ. 113, 10-19.
48. Hengl, T., et al., 2003. Comparison of kriging with external drift and regression-kriging. ITCpp. 51 Technical note.
49. Hipel, K.W., McLeod, A.I., 1994. Time Series Modelling of Water Resources and Environmental Systems. Amsterdam, Elsevier Science. ISBN: 9780080870366.
50. Hoek, G., Beelen, R., De Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., et al., 2008. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmos. Environ. 42, 7561-7578.
51. Hsu, S., Mavrogianni, A., Hamilton, I., 2017. Comparing spatial interpolation techniques of local urban temperature for heat-related health risk estimation in a subtropical city. Procedia Engineering, 198: 354 – 365.
52. Ivaskova, M, Kotes, P., Brodnan, M., 2015. Air pollution as an important factor in construction materials deterioration in Slovak Republic. Procedia Engineering 108, 131-138.
53. James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning: with Applications in R, Springer Texts in Statistics, DOI 10.1007/978-1-4614-7138-7 7.
54. Kadiyala, A., Kumar, A., 2012. Univariate time series prediction of air quality inside a public transportation bus using available software, Environmental Progess and Sustainable Energy, 31, 494–499.

55. Khallaf, M. (Ed.), 2011. The impact of air pollution on health, economy, environment and agricultural sources, InTech, DOI: 10.5772/17660. Available from: https://mts.intechopen.com/books/the-impact-of-air-pollution-on-health-economy-environment-and-agricultural-sources (accessed 29/03/2018).
56. Koenker, R., Bassett, G., 1978. Regression quantiles, Econometrica, 46 (1) (1978), pp. 33-50.
57. Kumar, P., Morawska, L., Martani, C., Biskos, G., Neophytou, M., Di Sabatino, S., Bell, M., Norford, L., Britter, R., 2015. The rise of microsensing for managing air pollution in cities. Environ. Int. 75, 199-205.
58. Lagzi, I., Meszaros, R., Gelybo, Leelossy, A., 2013. Atmospheric Chemistry. Eötvös Loránd University. Available online: http://elte.prompt.hu/sites/default/files/tananyagok/AtmosphericChemistry/ (accessed: 28/5/2018)
59. Landrigan, P.J., 2016. Air pollution and health. The Lancet Public Health, 2(1): 4 – 5. DOI: https://doi.org/10.1016/S2468-2667(16)30023-8.
60. Lee, J.-H.,Wu, C.-F., Hoek, G., et al., 2013. Land use regression models for estimating individual NOx and NO2 exposures in a metropolis with a high density of traffic roads and population. Sci. Total Environ. 472, 1163–1171.
61. Leontaritis, I., Billings, S., 1985. Input–out put parametric models for non-linear systems—part I deterministic non-linear systems. International Journal of Control 41,303–328.
62. Lewis A. C., Lee J. D., Edwards P. M., Shaw M. D., Evans M. J., Moller S. J., Smith K. R., Buckley J. W., Ellis M., Gillot S. R. and Whited A., 2016. Evaluating the performance of low cost chemical sensors for air pollution research. Faraday Discussions, 189, 85-103, 2016, DOI: 10.1039/C5FD00201J.
63. Lewis, A., Edwards, P., 2016. Validate personal air-pollution sensors. Nature 535 (7610), 29–31.
64. Li, L., Wu, J., Hudda, N., Sioutas, C., Fruin, S.A., Delfino, R.J., 2013. Mdeling the Concentrations of On-Road Air Pollutants in Southern California. Environ Sci Technol, 47(16): 9291–9299.
65. Liang, F., Gao, M., Xiao, Q., Carmichael, G.R., Pan, X., Liu, Y., 2017. Evaluation of a data fusion approach to estimate daily PM2.5 levels in North China. Environmental Research 158: 54–60.
66. Martins, F., Pires, J. and Sousa, S., 2009. Statistical Models for Predicting Ozone and PM_{10} Concentrations, In Modelling of Pollutants in Complex Environmental Systems, Hanrahan, G. (Eds.), p. 277–290.
67. Mead, M.I., Popoola, O. a. M., Stewart, G.B., Landshoff, P., Calleja, M., Hayes, M., Baldovi, J.J., McLeod, M.W., Hodgson, T.F., Dicks, J., Lewis, A., Cohen, J., Baron, R., Saffell, J.R., Jones, R.L., 2013. The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks. Atmos. Environ. 70, 186-203. http://dx.doi.org/10.1016/j.atmosenv.2012.11.060.
68. Meti, S.A., Sangam, V.G., 2005. Survey paper: multi sensor data fusion for sensor validation. International Journal of Advanced Computer Technology, 3(3): 21-28.
69. Modi, M., Ramachandra, V. P., Ahmed, L.S.K., Hussain, Z., 2013. A review on theoretical air pollutants dispersion models. International Journal of Pharmaceutical, Chemical and Biological Sciences (IJPCBS) 3(4), 1224-1230.
70. Molter, A., Lindley, S., de Vocht, F., Simpson, A., Agius, R., 2010. Modelling air pollution for epidemiologic research — Part I: A novel approach combining land use regression and air dispersion. Science of the Total Environment 408 (2010) 5862–5869.

71. Molter, A., Lindley, S., de Vocht, F., Simpson, A., Agius, R., 2010. Modelling air pollution for epidemiologic research – Part II: Predicting temporal variation through land use regression. Science of the Total Environment 409 (2010) 211–217.
72. Munir, S., Chen, H. and Ropkins, K. (2012). Modelling the Impact of Road Traffic on Ground Level Ozone Concentration using a Quantile Regression Approach. *Atmos. Environ.* 60: 283–291.
73. Munir, S., Chen, H., and K. Ropkins, 2014. Characterising the temporal variations of ground level ozone and its relationship with traffic-related air pollutants in the UK: a quantile regression approach. International Journal of Sustainable Development and Planning, 9 (1): 29 - 41.
74. Munir, S., Habeebullah, T.M., Seroji, A.R., Morsy, E.A., Mohammed, A.M.F., Saud, A.W., Abdou, A.E.A., Awad, A.H., 2013. Modelling Particulate Matter concentrations in Makkah, Applying a Statistical Modelling Approach. Aerosols and Air Quality Research, 13 (3): 901 – 910.
75. Muttoo, S., Ramsay, L., Brunekreef, B., Beelen, R., Meliefste, K., Naidoo, R.N., 2017. Land use regression modelling estimating nitrogen oxides exposure in industrial south Durban, South Africa. Science of the Total Environment 610–611 (2018) 1439–1447.
76. Nielsen, 2007. Introduction to data fusion, the Nielsen Company, 770 Broadway New York, NY 10003-9595. www.nielsen.com.
77. Nieto, P., J. C. Antón, J.C., 2014. Nonlinear air quality modeling using multivariate adaptive regression splines in Gijón urban area (Northern Spain) at local scale. Applied Mathematics and Computation, 235: 50 – 65.
78. Olea, R.A., 1999. Geostatistics for Engineers and Earth Scientists, Kluwer Academic Publishers. ISBN 978-1-4615-5001-3.
79. Parra, M., Santiago, J., Martín, F., Martilli, A., Santamaría, J., 2010. A methodology to urban air quality assessment during large time periods of winter using computational fluid dynamic models. Atmos. Environ., 44 (17), 2089-2097.
80. Pasquill, F., 1961. The estimation of the dispersion of windborne material. Met. Mag. 90: 33.
81. Pasquill, F., 1962. Some observed properties of medium-scale diffusion in atmosphere. Quart. J. R. Met. Soc. 88: 70. doi:10.1002/qj.49708837507.
82. Pasquill, F., 1974. Atmospheric Diffusion. 2nd ed. Ellis Horwood Ltd., Chichester. p. 228.
83. Penza, M., Suriano, D., Villani, M.G., Spinelle, L., Gerboles, M., 2014. Towards air quality indices in smart cities by calibrated low-cost sensors applied to networks. In: IEEE SENSORS 2014 Proc. 2012-2017.
84. Pires, J.C.M., Martins, F.G., Sousa, S. I. V. M., Alvim-Ferraz, C.M., Pereira, M. C., 2008. Prediction of the Daily Mean PM10 Concentrations Using Linear Models. American Journal of Environmental Sciences 4 (5): 445-453.
85. Pisoni, E., Farina, F., Carnevale, C., Piroddi, L., 2009. Forecasting peak air pollution levels using NARX models. Engineering Applications of Artificial Intelligence 22 (2009) 593–602.
86. Popoola, O., Mead, I., Bright, V., Baron, R., Saffell, J., Stewart, G., Kaye, P., Jones, R., 2013. A portable low-cost high density sensor network for air quality at London Heathrow airport. In: EGU General Assembly 2013, Held 7-12 April, 2013 in Vienna, Austria id. EGU2013e1907. http://wwwdev.snaq.org/posters/EGU_OAMP_2013.pdf.
87. Rehman, Md. M., Yeganeh, B., Clifford, S., Knibbs, L.D., Morawska, L., 2017. Development of a land use regression model for daily NO2 and NOx concentrations in the Brisbane metropolitan area, Australia. Environmental Modelling & Software 95 (2017) 168 - 179.

88. Reimann, C., Filzmoser, P., Garrett, R., Dutter, R., 2008. Statistical Data Analysis Explained: Applied Environmental Statistics, R. John Wiley and Sons, Ltd: London.
89. Ryan, P.H., LeMasters, G.K., 2007. A review of land-use regressionmodels for characterizing intraurban air pollution exposure. Inhal. Toxicol. 19 (Suppl. 1), 127–133.
90. Ryan, P.H., LeMasters, G.K., Biswas, P., Levin, L., Hu, S., Lindsey, M., Bernstein, D.I., Lockey, J.E., Villareal, M., Khurana Hershey, G.K., Grinshpun, S.A., 2007. A comparison of proximity and land use regression traffic exposure models and wheezing in infants. Environmental Health Perspect. 115, 278–284.
91. Salmond, J.A., Clarke, A.G., Tomlin, A.S., 2006. The atmosphere, chapter 2, pp: 8-76. In Harrison, R.M., an introduction to pollution science. The Royal Society of Chemistry.
92. Sanchez, B., Santiago, J.L., Martilli, A., Martin, F., Borge, R., Quaassdorff, C., de la Paz, D., 2017. Modelling NOX concentrations through CFD-RANS in an urban hotspot using high resolution traffic emissions and meteorology from a mesoscale model. Atmospheric Environment, 163: 155 - 165.
93. Santiago, J., Borge, R., Martin, F., de la Paz, D., Martilli, A., Lumbreras, J., Sanchez, B., 2017. Evaluation of a CFD-based approach to estimate pollutant distribution within a real urban canopy by means of passive samplers. Sci. Total Environ., 576, 46-58.
94. 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 (2016) 163 - 175.
95. Sayegh, A.S., Munir, S., Habeebullah, T.M., 2014. Comparing the Performance of Statistical Models for Predicting PM_{10} Concentrations. Aerosol and Air Quality Research, 14 (3): 653-665.
96. Schneider, P., Castell, N., Vogt, M., Dauge, F.R., Lahoz, W.A., Bartonova, A., 2017. Mapping urban air quality in near real-time using observations from lowcost sensors and model information. Environment International, 106: 234 – 247.
97. Singh KP, Gupta S, Kumar A, Shukla SP, 2012. Linear and nonlinear modeling approaches for urban air quality prediction. Sci Total Environ., 2012, 426:244-55.
98. Slordal, L.H., Walker, S.E., Solberg, S.S., 2003. The urban air dispersion model EPISODE applied in AirQUIS 2003 - technical description. In: Tech. rep. NILU – Norwegian Institute for Air Research, Kjeller, Norway.
99. Smith, K.R., Edwards P.M., Evans, M.J., Lee, J.D., Shaw, M.D., Squires, F., Wilde, S., Lewis, A.C., 2017. Clustering approaches to improve the performance of low cost air pollution sensors. Faraday Discuss., 200: 621-637.
100. Snyder, E., Watkins, T., Solomon, P., Thoma, E., Williams, R., Hagler, G., Shelow, D., Hindin, D., Kilaru, V., Preuss, P., 2013. The changing paradigm of air pollution monitoring. Environ. Sci. Technol. 47 (20), 11369-11377.
101. Son, C., 2002. Optimal Control Planning Strategies with Fuzzy Entropy and Sensor Fusion for Robotic Part Assembly Tasks. International Journal of Machine Tool & Manufacture, 42: 1335-1334.
102. Sousa, S.I.V., Pires, J.C.M., Mrtins, F.G., Pereira, M.C. and Alvim-Ferraz, M.C.M. (2008). Potentialities of Quantile Regression to Predict Ozone Concentrations. *Environmetrics* 20: 147–158.
103. Spinelle, L., Gerboles, M., Villani, M.G., Aleixandre, M., Bonavitacola, F., 2017. Field calibration of a cluster of low-cost commercially available sensors for air quality monitoring. Part B: NO, CO and CO2. Sensors and Actuators B 238 (2017) 706–715.

104. Srivastava, T., 2015. A Complete Tutorial on Time Series Modeling in R. Available online: https://www.scribd.com/document/343194193/A-Complete-Tutorial-on-Time-Series-Modeling-in-R (accessed: 19/06/2018)

105. Stedman, J., Vincent, K., Campbell, G., Goodwin, J., Downing, C., 1997. New high resolution maps of estimated background ambient NOx and NO2 concentrations in the U.K. Atmos. Environ. 31, 3591–3602.

106. Stojanovic, M., Bartonova, A., Topalovic, D., et al., 2015. On the use of small and cheaper sensors and devices for indicative citizen-based monitoring of respirable particulate matter. Environ. Pollut. 206, 696-704.

107. Suleiman, A., Tight, M.R., Quinn, A.D., 2015. Hybrid Neural Networks and Boosted Regression Tree Models for Predicting Roadside Particulate Matter. Environ Model Assess (2016) 21:731–750.

108. Suriano, D., Prato, M., Pfister, V., Cassano, G., Camporeale, G., Dipinto, S., Penza, M., 2015. Stationary and Mobile Low-Cost Gas Sensor-Systems for Air Quality Monitoring Applications. Fourth Scientific Meeting EuNetAir, 2015-06-03 - 2015-06-05, Linkoping University, Linkoping, Sweden. DOI: 10.5162/4EuNetAir2015/15.

109. Syafei, A.D., Fujiwara, A., Zhang, J., 2015. Prediction Model of Air Pollutant Levels Using Linear Model with Component Analysis. International Journal of Environmental Science and Development, 6 (7): 2015.

110. Van den Bossche, J., Peter, J., Verwaeren, J., Botteldooren, D., Theunis, J., De Baets, B., 2015. Mobile monitoring for mapping spatial variation in urban air quality: development and validation of a methodology based on an extensive dataset. Atmos. Environ. 105, 148-161.

111. Viana, M., Rivas, I., Reche, C., Fonseca, A.S., Perez, N., Querol, X., Alastuey, A., Alvarez-Pedrerol, M., Sunyer, J., 2015. Field comparison of portable and stationary instruments for outdoor urban air exposure assessments. Atmos. Environ. 123, 220-228.

112. Vienneau, D.; deHoogh, K.; Beelen, R.; Fischer, P.; Hoek, G.; Briggs, D. Comparison of land-use regression models between Great Britain and the Netherlands. *Atmos. Environ.* 2010, 44, 688-696.

113. Walters, S., Ayres, J., 2001. The health effects of air pollution. in pollution causes, effects and control. In Harrison, R.M. (Ed.), Chapter 11, pp. 275. Fourth Editions, Royal Society of Chemistry, Cambridge, UK. ISBN 0-85404-621-6.

114. Wang, M., Beelen, R., Bellander, T., Birk, M., et al. 2014. Performance of Multi-City Land Use Regression Models for Nitrogen Dioxide and Fine Particles. Environ Health Perspect, 8 (122). DOI:10.1289/ehp.1307271.

115. Westmoreland, E.M., Carslaw, N., Carslaw, D.C., Gillah, A. and Bates, E., 2007. Analysis of Air Quality within a Street Canyon Using Statistical and Dispersion Modelling Techniques. Atmos. Environ. 41: 9195–9205.

116. WHO, 2013. Health effects of particulate matter, policy implications for countries in eastern Europe, Caucasus and central Asia. Publications of WHO Regional Office for Europe UN City, Marmorvej 51 DK-2100 Copenhagen, Denmark.

117. Wilkening, I.H., Baraldi, D., 2007. CFD modelling of accidental hydrogen release from pipelines. International Journal of Hydrogen Energy, 32 (13): 2206-2215.

118. William, M.L., 2000. Atmospheric dispersal of pollutants and the modelling of air pollution, chapter 10, pp:246 – 266. In Harrison, R.M., Pollution: causes, effects and control, fourth edition. The Royal Society of Chemistry.

119. Williams, R., Vasu Kilaru, E. Snyder, A. Kaufman, T. Dye, A. Rutter, A. Russell, AND H. Hafner, 2014. Air Sensor Guidebook. U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-14/159 (NTIS PB2015-100610), 2014.

120. Wood, S.N., 2006. Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC.

121. Wood, S.N., 2017. *Generalized Additive Models: an introduction with R (2nd edition)*, CRC.

122. Wu, H., Reis, S., Lin, C., Heal, M.R., 2017. Effect of monitoring network design on land use regression models for estimating residential NO2 concentration. Atmospheric Environment 149, 24 - 33.

123. Yarwood, G., Emery, C., Baker, K., Dolwick, P., 2014. Resolving and Quantifying Ozone Contributions from Boundary Conditions Within Regional Models. In: Steyn D., Mathur R. (eds) Air Pollution Modelling and its Application XXIII. Springer Proceedings in Complexity. Springer, Cham.

124. Zito, G., Landau, I.D., 2005. A methodology for identification of NARMAX models applied to diesel engines. IFAC World Congress, Jul 2005, Prague, Czech Republic. Elsevier.

CHAPTER 3: ANALYSING THE PERFORMANCE OF LOW-COST AIR QUALITY SENSORS, THEIR DRIVERS, RELATIVE BENEFITS AND CALIBRATION IN CITIES - A CASE STUDY IN SHEFFIELD

[1]Said Munir*, [1]Martin Mayfield, [2]Daniel Coca, [1]Stephen A Jubb, [3]Ogo Osammor

[1]Department of Civil and Structural Engineering, the University of Sheffield, S1 3JD
[2]Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield, S1 3JD
[3]Air Quality Monitoring & Modelling, Sheffield City Council, Howden House, 1 Union Street, Sheffield, S1 2SH
*Corresponding author smunir2@sheffield.ac.uk

Published: Environmental Monitoring and Assessment (2019) 191:94
https://doi.org/10.1007/s10661-019-7231-8
Received: 23 September 2018 /Accepted: 10 January 2019

Abstract

Traditional real-time air quality monitoring instruments are expensive to install and maintain, therefore such existing air quality monitoring networks are sparsely deployed and lack the measurement density to develop high-resolution spatiotemporal air pollutant maps. More recently, low-cost sensors have been used to collect high-resolution spatial and temporal air pollution data in real-time. In this paper, for the first time Envirowatch E-MOTEs are employed for air quality monitoring as a case study in Sheffield. Ten E-MOTEs were deployed for a year (October 2016 to September 2017) monitoring several air pollutants (NO, NO_2, CO) and meteorological parameters. Their performance was compared to each other and to a reference instrument installed nearby. E-MOTEs were able to successfully capture the temporal variability such as diurnal, weekly and annual cycles in air pollutant concentrations and demonstrated significant similarity with reference instruments. NO_2 concentrations showed very strong positive correlation between various sensors. Mostly correlation coefficients (r-values) were greater than 0.92. CO from different sensors also had r-values mostly greater than 0.92, however, NO showed r-value less than 0.5. Furthermore, several Multiple Linear Regression Models (MLRM) and Generalised Additive Models (GAM) were developed to calibrate the E-MOTE data and reproduce NO and NO_2 concentrations measured by the reference instruments. GAMs demonstrated significantly better performance than linear models by capturing the nonlinear association between the response and explanatory variables. The best GAM developed for reproducing NO_2 concentrations returned values of 0.95, 3.91, 0.81, 0.005, and 0.61 for Factor of two (FAC2), Root Mean Square Error (RMSE), coefficient of determination (R^2), Normalised Mean Biased (NMB) and Coefficient of Efficiency (COE), respectively. The low-cost sensors offer a more affordable alternative for providing real time high-resolution spatiotemporal air quality and meteorological parameter data with acceptable performance.

Keywords: sensors cost, sensor networks, Envirowatch E-MOTEs, air pollution monitoring, generalised additive model.

3.1. Introduction

With an increasing trend towards urbanisation due to better job opportunities and greater access to amenities and facilities in cities, urban areas are expanding rapidly globally. Given this trend, air pollutant levels are increasing, especially in large urban agglomerations and at roadside locations, which adversely impact human health in a variety of ways. Air pollutants, especially high levels of nitrogen dioxide (NO_2) and particulate matter (PM_{10} and $PM_{2.5}$) are considered the most significant environmental risks to public health in urban areas in the UK (Department for Environment, Food and Rural Affairs (DEFRA) 2015; World Health Organisation (WHO) 2013). Atmospheric air pollutants were estimated to cause seven million premature deaths in 2012, worldwide (WHO 2014). Air pollutants (e.g., NO_2 and PM_{10}) emitted by various emission sources are risk factors and are reported to increase the risk of incidence of various diseases including heart disease, lung cancer, and both chronic and acute respiratory diseases, including asthma (WHO 2014).

Air quality monitoring is important to promote air quality awareness and to support abatement strategies (Borrego et al. 2016). Several techniques are used to monitor air quality (Penza et al. 2014), which include (a) Reference or conventional real-time air quality monitoring; (b) Portable air quality monitors; (c) Passive diffusion tubes; and (d) Digital sensors. Reference air quality monitoring instruments are the most accurate and are used for air quality compliance purposes, studying exposure, supporting air quality management and developing policies for reducing and controlling emissions. Reference instruments are expensive to purchase and maintain and therefore the spatial resolution of air quality measurement is low and insufficient for detailed spatiotemporal mapping. Portable or mobile monitors are either carried by individuals or installed in vehicles that can be stationed where fixed continuous monitors cannot be installed. Portable instruments can be useful for monitoring air quality in certain cases and can provide high-resolution temporal data for a short period of time, but have limited application for spatial mapping and long-term monitoring. Passive tubes are small collection devices used for monitoring gaseous air pollutants such as NO_2 and typically provide monthly average concentrations, which can be converted to annual averages. These diffusion tubes are the cheapest technique and provide better spatial coverage. However, these can be used only for gaseous air pollutants and for long-term monitoring (mainly monthly average). Low-cost sensors (LCS) are used to collect real-time air quality data providing high-resolution spatial and temporal air quality data. These type of sensors are the new trend in air quality monitoring and can support the conventional air quality monitoring stations to increase the density of the sensing network (Heimann et al. 2015; Van den Bossche et al. 2015; Viana et al. 2015). The low-cost sensors use the latest microsensing technology and are considered the innovative tools for air quality monitoring in the future (Castell et al. 2015; Snyder et al. 2013; Kumar et al. 2015; Stojanovic et al. 2015). Data collected by these sensors can be used for detailed spatial and temporal mapping of air pollution, especially over distinct areas such as city or an urban district, for atmospheric model validation and assessing population exposure, however, the data need to be handle with cautions and several corrections need to be applied first.

Several authors have analysed the performance of the LCS, comparing their performance with reference instruments and with each other. Borrego et al. (2016) performed such an assessment (sensors compared to reference instruments) in Aveiro, Portugal from 13th to 27th October 2014. LCS and reference instruments were colocated and monitored the levels of gaseous pollutants (e.g., CO, NOx, O_3, SO_2), particulate matter (PM_{10}, $PM_{2.5}$) and meteorological parameters (e.g., temperature, wind speed and direction, relative humidity, solar radiation and precipitation). The resultant measurements were mutually compared and different sensors showed significantly different performance in terms of the statistical metrics used for evaluating the sensors' performance. The range of R^2 (coefficient of determination) values for different air pollutants were: O_3 (0.12-0.77), CO (0.53-0.87), NO_2 (0.02-0.89), PM (0.07-0.36) and SO_2 (0.09-0.20), where a lower R^2 value shows poor measurement performance of the sensors. Borrego et al. (2016) concluded that LCS had great potential for air quality monitoring, if properly supported by post processing and data modelling tools.

Different sensor systems use different principles to measure the concentrations of atmospheric pollutants (Borrego et al. 2016). These include optical particle counters (OPC), metal oxide semiconductor sensors (MOS), electrochemical sensors (EC), nondispersive infrared sensors (NDIR) and photo-ionisation detection sensors (PID). Aleixandre and Gerboles (2012) reported that these air quality sensors work through either measuring the electrochemical interaction between the sensing materials and the atmospheric chemicals or through absorption of visible light. The principle of light scattering or absorption is used for measuring the levels of PM. Individual sensors are usually integrated into a platform of sensors known as a sensor node. Each sensor node contains a sensor board, the sensors and a control board which integrates all the elements of the hardware such as GPS, data storage, communication ports and signal conditioning. Examples of networks based on these types of sensors are: (a) Cambridge university Sensor Network for Air Quality (SNAQ) (Mead et al. 2013; Popoola et al. 2013; Borrego et al. 2016), (b) AUTh-ISAG AQ Microsensors (Borrego et al. 2016), (c) Energy Centre of Netherlands (ECN Airbox) (Borrego et al. 2016; Hamm et al. 2016); (d) NanoEnvi platform (Borrego et al. 2016), (e) AQMesh sensors (Borrego et al. 2016; Carruthers et al. 2016), (f) ENEA Air-Sensor (Suriano et al. 2015), (g) EveryAware Sensor Box (Borrego et al. 2016), and Envirowatch E-MOTE sensors (Reis et al. 2013). These sensors are briefly described below.

(a) Cambridge university SNAQ are microsensors for measuring the concentrations of multispecies including gases air pollutants, particulate matter and meteorological parameters. These are low cost sensors and can be powered by battery or mains. Mead et al. (2013) employed these microsensors for monitoring air quality in Cambridge. Mead et al. (2013) reported widely varying concentrations of air pollutants in the urban environment, which could not be characterised by sparse static conventional air quality network. Furthermore, Popoola et al. (2013) deployed these sensors in Heathrow airport in London for air quality monitoring. They reported considerable spatial and temporal variations in air pollutant concentrations across the air quality network. According to their findings high air pollutant levels were linked with stable weather conditions.

(b) AUTh-ISAG AQ Microsensors use the principle of Waspmote wireless network, developed by Libellium, which is an international IT and Engineering company. These sensors aim to reduce power consumption, reduce thermal noise, provide easy inspection and require low maintenance. Data are normally collected using an SD card and can be run using both battery and main power supply. These sensors were used by Borrego et al. (2016) in their study and their performance was compared to several other microsensors and reference instruments. These sensors can measure the concentrations of several air pollutants and meteorological parameters.

(c) ECN – Airbox were developed by the Energy research Centre of the Netherlands (ECN). Airbox sensors monitor particulate matter (e.g., Ultra-Fine Particles (UFP), PM_1, $PM_{2.5}$, and PM_{10}), gaseous (e.g., NO_2 and O_3) and meteorological parameters (e.g., temperature and relative humidity). Airbox sensors have been used for air quality monitoring in Netherland in the city of Eindhoven in 35 locations since 2013. These sensors are powered by battery and mains. Hamm et al. (2016) have provided a detailed review of these sensors, which could be read for further details.

(d) NanoEnvi sensors were manufactured by Envira. These analysers use several sensors with different technology. The sensors work is based on the changes in electrical properties that happen in the surface of sensors when pollutants are present. The air pollutants which can be measured by nanoEnvi are gaseous pollutants (e.g., SO_2, NO, NO_2, CO, CO_2, O_3, H_2S and VOCs), particulates (PM_{10} and $PM_{2.5}$), and meteorological parameters (e.g., wind characterisitcs, temperature, relative humidity).

(e) AQMesh sensors are manufactured by Environmental Instruments Ltd, UK. These are low-cost micro-scale sensors for effective environmental monitoring, which are developed for harsh outfield environmental conditions and are capable of working to high standards. AQMesh microsensors measure the concentrations of NO, NO_2, O_3, SO_2, and CO using the latest generation of electrochemical sensors. Particulate matter is measured using a light scattering optical particle counter. Using solid state sensors, they can also measure the levels of temperature, RH and atmospheric pressure. Carruthers et al. (2016) compared the performance of AQMesh in Cambridge with reference instruments where AQMesh showed considerably higher concentrations of NO_2, NO and PM_{10}, however overall they performed well and showed great potential for contributing to the air quality monitoring, especially improving the spatial coverage in the UK.

(f) ENEA Air-Sensor are manufactured by ENEA (Energia Nucleare ed Energie Alternative), which is an Italian agency for new technology, energy and environment. These sensors measure the levels of several air pollutants, such as CO, NO_2, O_3, SO_2, H_2S and PM_{10} and meteorological parameters such as relative humidity and temperature. These sensors can be operated via battery or mains. Suriano et al. (2015) evaluated the performance of these Air-sensors during a campaign of several months in Italian national projects for sustainable innovation in the smart cities. These sensors were used both as stationary and mobile air quality monitoring systems, and initial results indicated that these sensors potentially could improve air quality monitoring program.

(g) EveryAware sensors are manufactured by Vito (a leading independent research and technology organisation based in Belgium and works in the areas of cleantech and sustainable development) under the European Seven Framework Program (EU-FP7). The EveryAware sensors are used for air quality monitoring in Belgium, Italy and the UK. EveryAware is a low-cost, portable air quality monitor used for measuring personal exposure to traffic pollution. This device contains six low-cost gas sensors that react in the presence of traffic pollutants (e.g., CO, NOx). Borrego et al. (2016) used EveryAware sensors in Aveiro, Portugal to compare their performance with other microsensor and reference instruments.

Dongol (2015) has listed several sensor platforms which include DunavNet Platform, UrVamm, GeoTech, and ATEKNEA. In addition to these sensors, there are several other types of sensors available for air quality monitoring and the listing is growing with time. Sensors of this type are cheaper, compact, user-friendly and provide high resolution spatiotemporal air pollutant concentrations. They have the potential to enhance the existing air quality network run at local levels by local authorities and nationally by DEFRA. In addition, these sensors can be installed independently by various research and governmental organisations to monitor public exposure to various air pollutants within a specific area. Despite all these positive points, the quality of air pollution data collected by these sensors is unproven and cannot be used for regulatory and compliance purposes, however, the data can be used for highlighting air pollution hotpots, for public awareness and for complementing traditional air quality monitoring programmes. There is a need for further investigation to quantify uncertainties in the datasets these types of sensors produce. These uncertainties are related to exposure to harsh environmental conditions, especially extreme temperature and relative humidity and the associated time interval (i.e. the length of time the instruments are operated in such a harsh environment). Furthermore, uncertainties are also affected by the measuring principles of the sensors and the quality of the materials used by the manufacturers. Therefore, inter-comparison of LCS made by different manufacturers and with reference instruments is required. Further work is, also required to improve the performance of these sensors by (a) improving their technology further to make it more robust, (b) frequent calibration both in laboratory and outdoor and (c) improving the experimental designs.

In this project the aim is to install LCS in the City of Sheffield to provide high resolution spatiotemporal maps of various air pollutants, especially NO_2 which is a pollutant of particularly concern in Sheffield as well as the rest of the UK. In this paper the aim is to evaluate the monitoring capability of Envirowatch E-MOTEs for air quality monitoring. This is the first paper comparing the performance of Envirowatch E-MOTEs with each other and with reference instruments, which are recommended by the European Union and UK DEFRA for air quality monitoring. The paper analyses a year’s worth of data and provides a more detailed assessment in comparison to previous studies (which have generally analysed sensor data for a limited time ranging from a week to a couple of months). Furthermore, supervised machine learning approaches including multiple linear regression and generalised additive modelling approaches are employed to calibrate the sensors by comparing their measurements with the reference instruments and setting up the slope and intercept.

3.2. Methodology

In this project the aim is to analyse CO (ppm), NO and NO_2 (ppb) data measured by LCS (Envirowatch E-MOTEs) and NO and NO_2 (ppb) measured by reference sensors, along with meteorological data such as wind speed, temperature and relative humidity, to assess the performance of LCS. All these data were available for the period October 2016 to September 2017. In this section, firstly we describe Envirowatch E-MOTEs, their operating principle and the air quality monitoring network in Sheffield. This is followed by a statistical analysis which includes model selection, development and assessment.

3.2.1. Envirowatch E-MOTEs

In this project E-MOTEs developed by Envirowatch Newcastle, UK were employed. The E-MOTE was launched by Envirowatch in 2010. Precision or reference instruments used for air quality monitoring are large and expensive to both purchase and maintain, in contrast these sensors are cheaper, small and suitable for a high density air quality monitoring network. E-

MOTEs work on a similar principle as the AQMesh pods, which use the latest generation of electrochemical sensors made by Alphasense. E-MOTEs were used to measure the levels of three gaseous pollutants: carbon monoxide (CO), nitric oxide (NO) and nitrogen dioxide (NO_2).

The E-MOTEs use wireless technology to communicate their sensor reading and can be deployed on lamp posts or other street furniture (Fig. 3.1). E-MOTEs in a cluster communicate with a gateway by means of the Zigbee protocol within a specific area for high-resolution monitoring. The use of this protocol allows the individual units to communicate with each other and pass data from sensors that are not in range or without line-of-sight of the gateway. Using GPRS, the gateway device communicates the collected data over an internet connection to a cloud server operated by Envirowatch. The data are post-processed and presented for access by users via the Enviroview web interface as well made available for download via an application programming interface (API).

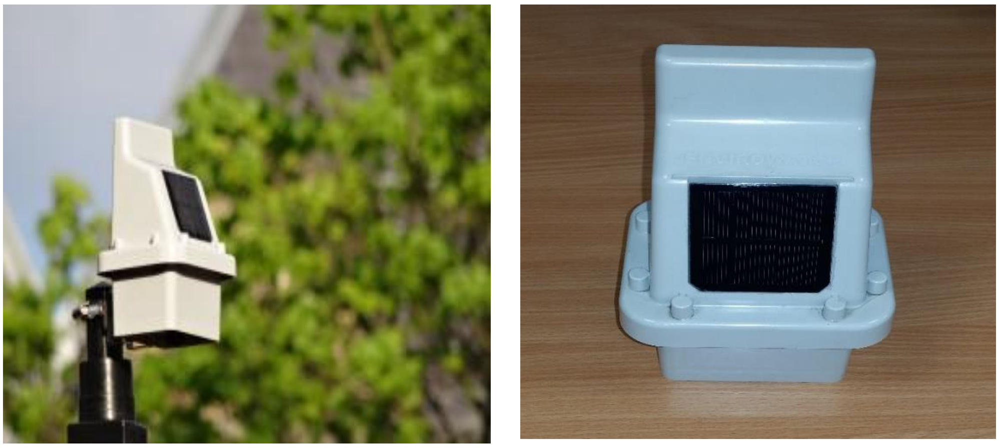


Fig. 3.1: Envirowatch E-MOTEs post-mountedt (left) and showing the solar panel used for battery charging (right). Ten of these E-MOTEs were used for collecting data used in this study.

LCS are more compact, portable and use less power as compared to reference instruments. E-MOTEs use electrochemical technology for measuring gaseous air pollutants, including NO_x, CO and O_3. Electrochemical sensors work by reacting to the target gas, generating an electrical output, which varies with the concentration of target gases present in air. Independent Envirowatch E-MOTEs transmit raw measurement data to a cloud server. These data are not concentration readings as such and require post-processing. Once readings are received mathematical processing is applied to correct cross-gas effects and prevailing environmental factors.

An electrochemical sensor contains a cell where three electrodes are present. These electrodes are known as the working or sensing electrode, counter electrode and reference electrode. The electrodes are separated by wetting filters, which are hydrophobic separators enabling ionic (cation and anion) contact between the electrodes, allowing transport of the electrolyte via capillary action. The sensed gas is either reduced or oxidised at the working electrode. These reactions are catalysed by the electrode materials specifically developed for the gas in question. Normally, the rate of diffusion of the sensed gas to the sensor electrode is slower than the rate of reaction of the gas at the electrode. Therefore, the concentration of the sensed gas determines the electrical current output by the sensor (Mead et al. 2013). The potential difference between the working and counter electrodes then generates an electric current which is the output signal of the sensor. With a resistor connected across the electrodes, a current proportional to the gas concentration flows between the anode and the cathode. Thus the current can be measured to determine the gas concentration. The current generated by these types of electrochemical sensors is measured using suitable electronics and, following further processing, displayed as a concentration measurement in ppm (for CO) or ppb (for NOx, and O_3).

3.2.2. Air Quality Monitoring Network (AQMN)

Air quality data analysed in this paper are mainly from two sources: LCS and reference instruments, which are described below:

LCS network

LCS used for air quality monitoring were Envirowatch E-MOTEs. Ten E-MOTEs were deployed at the University of Sheffield Campus (Fig. 3.2) for a year (October 2016 to September 2017). This area is bounded by Mappin Street, Rockingham Street, Portobello Street and Broad Lane and can be classified as urban background area. This area is part of the University of Sheffield and is mainly comprised of offices, lecture theatres and student accommodation. E-MOTEs provide minute-by-minute air pollutant measurements, which were converted to hourly averages to make them comparable to the data collected by reference instruments. Sensor identities and coordinates of their locations are shown in Table 1 along with the average annual concentration of each pollutant measured.

Reference instruments network

Several reference instruments are installed to monitor various air pollutant concentrations in Sheffield. These total nine (9) continuous air quality monitoring stations (AQMS) and provide hourly concentrations of air pollutants, including NOx, CO, SO_2, O_3 and particulate matter mainly PM_{10} and $PM_{2.5}$. Out of these, three (3) of the monitoring stations are part of the Automatic Urban and Rural Network (AURN) run by the UK government's DEFRA, whereas the remaining six sites are installed and managed by Sheffield City Council (Fig. 3.3). Devonshire Green (AURN), Waingate (RM1) and Wicker (GH4) are the nearest to the E-MOTEs network. However, data from October 2016 to September 2017 were available only from Devonshire Green (DG) monitoring station, which are compared with data from the installed sensors. Fig. 3.4 shows box plots comparing NO (lower-panel) and NO_2 concentrations (middle-panel) measured by each of the E-MOTEs and with reference sensors (upper-panel). The box plots show the distribution of the concentrations with some descriptive statistics including median (middle line of the box), lower or first quartile (lower end of the box), upper or third quartile (upper end of the box), inter-quartile range (representing middle 50 % of the data points), upper and lower whiskers representing concentrations outside the middle 50%, and outliers (point lying beyond the whiskers. Box plots compare both central tendency and variability or distribution of the concentrations. NO_2 concentrations measured by the various sensors exhibit a similar pattern, in contrast NO concentrations show much more variability.

Table 3.1: Coordinates of the sensors and data summary showing the mean concentrations (annual mean) of various air pollutants from October 2016 to September 2017.

Sensors ID	Northing (m)	Easting (m)	CO (ppm)	NO (ppb)	NO_2 (ppb)
S701	392846	631411	0.33	2.39	58.00
S702	392846	631425	0.46	20.31	16.60
S703	392845	631437	0.33	10.95	13.30
S704	392878	631425	0.33	3.70	18.85
S705	392878	631409	0.35	11.55	19.03
S706	392883	631390	0.43	15.98	18.60
S707	392878	631418	0.33	7.87	17.59
S708	392900	631429	0.33	9.60	17.93
S709	392837	631400	0.33	9.49	18.70
S710	392837	631418	0.32	5.66	17.07

ATTERCLIFFE
HILLFOOT
NEEPSEND
BURNGREAVE
Kelham Island Museum
CROOKES
CROOKESMOOR
WEST BAR
The University of Sheffield
PARK HILL
Sheffield
Royal Hallamshire Hospital
Crosspool
Sheffield Botanical Gardens
ENDCLIFFE
SHARROW
Norfolk Heritage Park

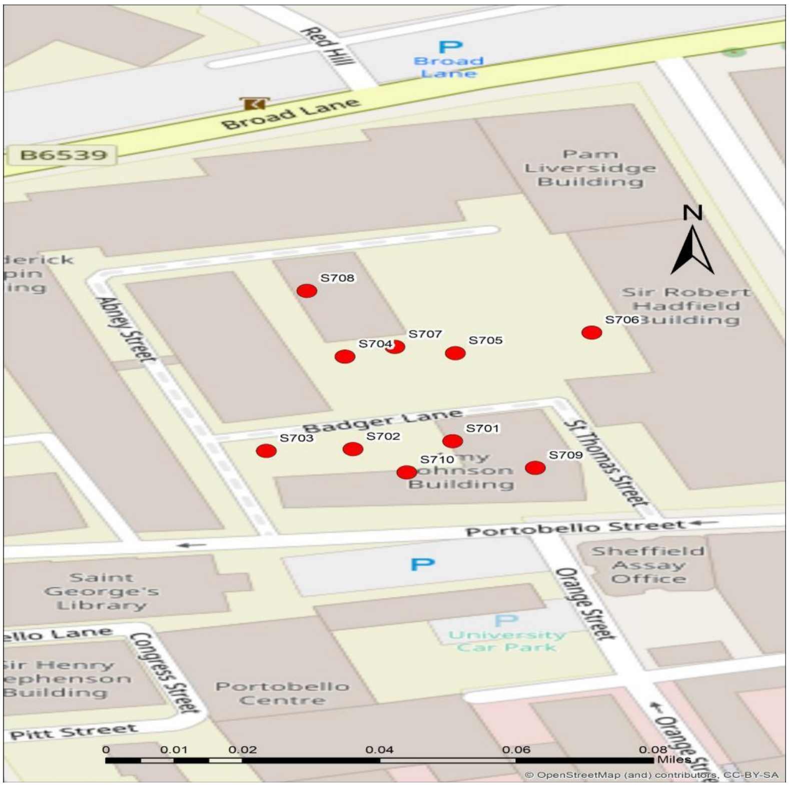


Fig. 3.2: Map of the locations of the Envirowatch E-MOTEs included in this study, where the red rectangle in the upper panel shows the location where sensors were deployed and the lower panel shows their localisation sites (the map was developed in ArcMap 10.4.1 using basemaps of OpenStreetMap).

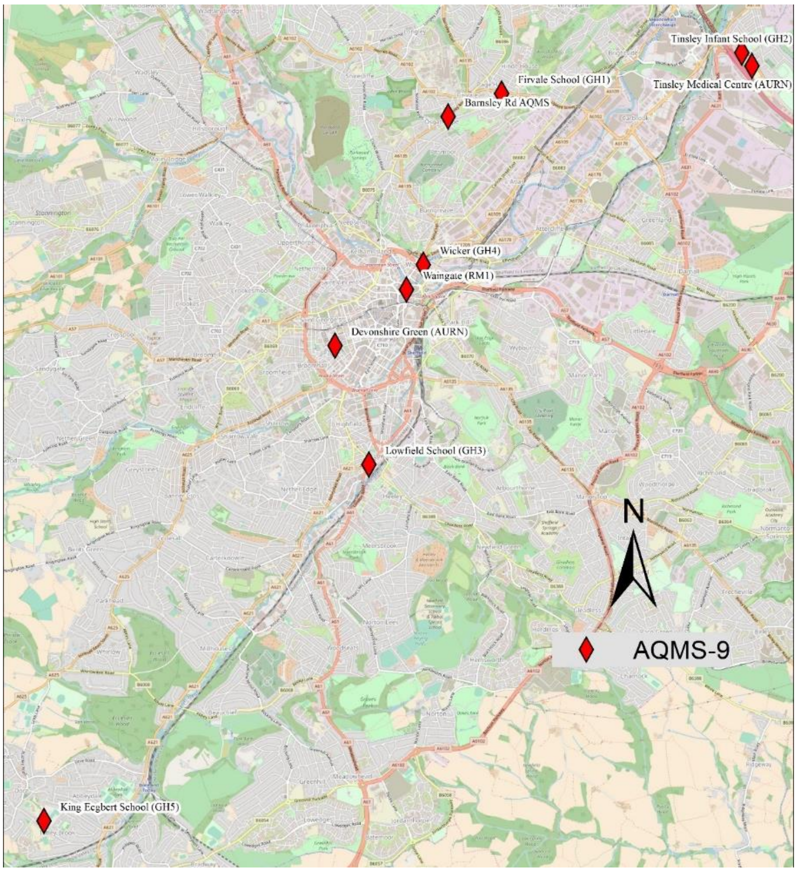
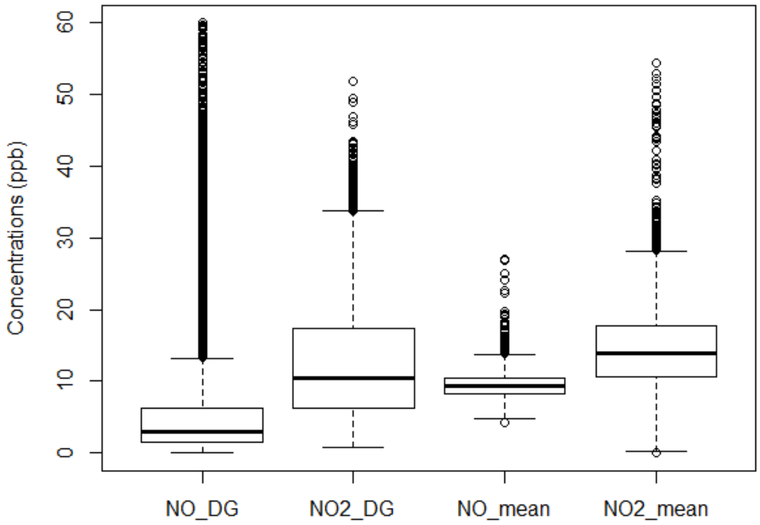
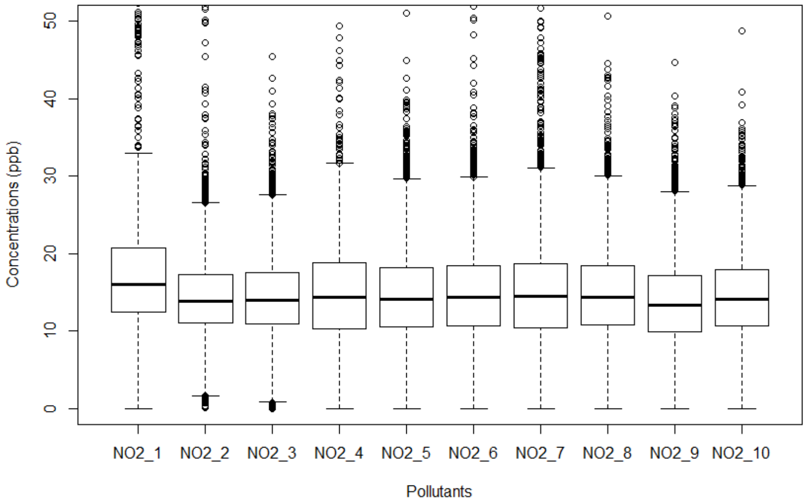


Fig. 3.3: Air quality monitoring network of continuous monitoring stations in Sheffield comprised of AURN sites run by DEFRA and Sheffield City Council sites (the map was developed in ArcMap 10.4.1 using basemap of OpenStreetMap).





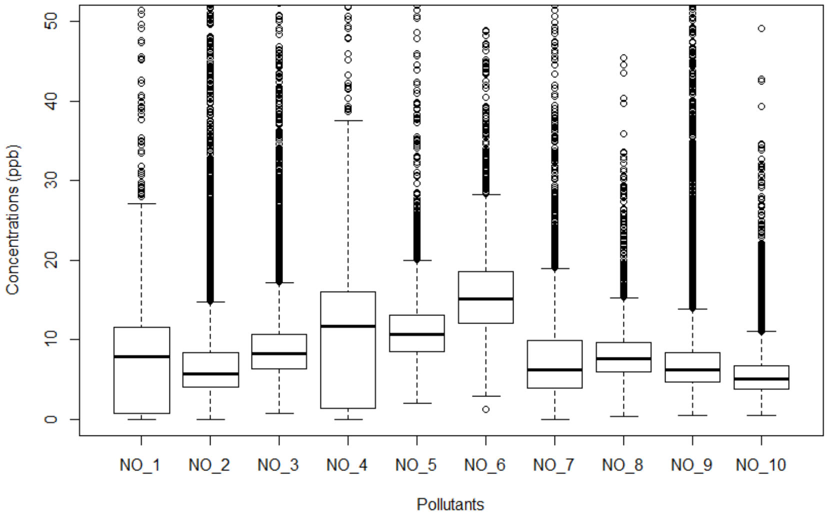


Fig. 3.4: Box plots of hourly concentrations (ppb) NO (lower panel), NO_2 (centre panel) measured by E-MOTEs and their mean compared with reference measurements from Devonshire Green monitoring station (upper panel).

3.2.3. Statistical analysis

Statistical analyses were carried out, comprising correlation analysis, regression analysis and graphical presentations, in the base packages of the R programming language (R Core Team 2017) and two of its additional packages known as ‘openair’ (Carslaw 2016) and ‘mgcv’ (Wood 2017).

In this paper supervised machine learning approaches are suggested for calibrating E-MOTEs outputs in comparison with measurements gathered from the reference instruments. Although these sensors are pre-calibrated by the manufacturers, they require local outfield calibration to account for cross interference of other pollutants and meteorological parameters, e.g., temperature and relative humidity. Two modelling approaches are employed in this study: (a) Linear Regression Models (LRM); and (b) Generalised Additive Models (GAMs). For details on these models see Hastie and Tibshirani (1990), Wood (2006), Munir et al. (2013) and Sayegh et al. (2014).

3.2.3.1. Model selection - choosing the best set of predictors

Air pollutant data were obtained from ten E-MOTEs and a reference AQMS each measuring NO and NO_2. Meteorological data of wind speed, relative humidity and temperature were also available from a weather station collocated with reference station. Firstly, NO and NO_2 from all ten E-MOTEs (making twenty variables) along with relative humidity, wind speed and temperature were considered as predictors (independent variables) for predicting the concentration of NO and NO_2 measured by the reference instrument (Fig. 3.5, upper-panel). Various other combinations of predictors were also tested to find the best set of predictors using Best Subset Regression (BSR). After testing a combination of various predictors, six predictors were chosen and were used in the model development to model the concentrations of NO_2 and NO measured by reference instrument. It can be seen in (Fig. 3.5 upper-panel) that the value of R^2 increases with an increase in the number of independent variables, however, after adding a certain number of covariates the line becomes horizontal showing little improvement in the R^2 value. Considering the results of BSR and the outputs of the actual LRM and GAM (discussed in coming sections), the final number of covariates were decided. The whole dataset was divided into two subsets: a training dataset (75%) and a testing dataset (25%) both selected randomly. The raining dataset was used to train the model, whereas the testing dataset was used to assess the model’s performance and check its validity.

The model selection process examines all possible sets of predictors in ordinary least square (OLS) regressions and leads to choosing one that fits best according to some criterion. The criterion could be based on p-value as in the standard stepwise methods (e.g., backwards stepwise regression), which take one variable away and then re-examine the model. Alternatively, the criterion could be based on R^2 or adj-R^2. This is called BSR or leaps-and-bounds approach. Criterion based on R^2 and adjR^2 is technically much stronger than on the p-value, therefore, in this paper the leaps-and-bounds method is adopted. To apply the leaps-and-bounds method, we employed one of the package of R programming language known as ‘Leaps’ to select the best set of predictors.

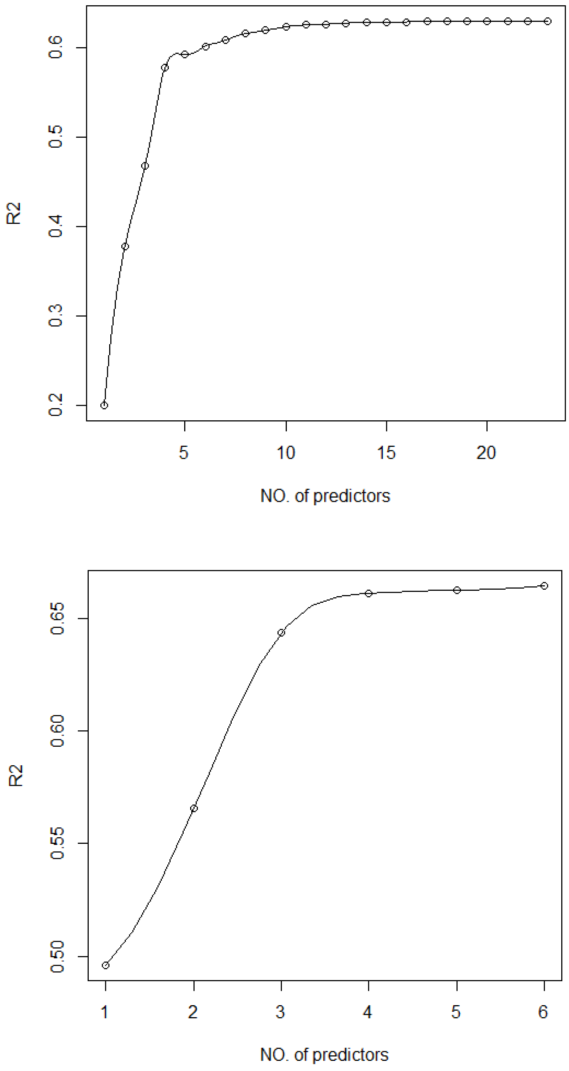


Fig. 3.5: Best Subset Regression (BSR) using 23 predictors (NO_1 to NO_10, NO_2_1 to NO_2_10, Wind Speed (WS), Temperature (Temp), and Relative Humidity (RH)) for predicting NO_2_DG (upper panel) and 6 predictors (NO_mean, NO_2_mean, NO_DG, WS, Temp and RH) for predicting NO_2_DG (lower panel).

3.2.3.2. Model development

In this paper two modelling approaches are employed: Linear Regression Model (LRM) and Generalised Additive Model (GAM).

(a) LRM

Two types of linear models were developed: Simple linear regression and multiple linear regression model. In simple linear regression model only one dependent variable (predictor) was used. This helps correct slopes and offsets (intercepts) values of the low-cost sensors to improve the accuracy of results. During calibration the measurements are regressed versus reference measurements, where readings from the E-MOTEs (NO_mean or NO_2_mean) are taken as independent (x-axis) and reference readings (NO_DG or NO_2_DG) as the dependent (y-axis) variable. The regression model is run and values of slopes and intercepts are calculated as shown in equation 3.1 and 3.2, here DG stands for Devonshire Green, which is the location of a reference air quality monitoring station and NO_mean is the average of the readings from all the low-cost sensors.

$$\text{NO\_DG} = \beta_o + \beta_1(\text{NO\_mean}) + \varepsilon \qquad (3.1)$$

$$\text{NO}_2\text{\_DG} = \beta_o + \beta_1(\text{NO}_2\text{\_mean}) + \varepsilon \qquad (3.2)$$

The values of slopes and intercepts are then applied to the whole dataset of E-MOTEs. β_o is the intercept, β1 is the coefficient or slope, Ɛ is the error term (the difference between observed and modelled concentrations).

To account for cross interference and for the effect of meteorological parameters, a multiple linear regression model was developed for each NO and NO_2 value as given in equations 3.3 and 3.4 using the predictors selected in the model selection section (3.2.1).

$$\text{NO\_DG} = \beta_o + \beta_1 (\text{NO\_mean}) + \beta_2 (\text{NO}_2\text{\_DG}) + \beta_3 (\text{NO}_2\text{\_mean}) + \beta_4 (\text{WS}) + \beta_5 (\text{RH}) + \beta_6 (\text{Temp}) + \varepsilon \qquad (3.3)$$

$$\text{NO}_2\text{\_DG} = \beta_o + \beta_1 (\text{NO\_DG}) + \beta_2 (\text{NO}_2\text{\_mean}) + \beta_3 (\text{NO\_mean}) + \beta_4 (\text{WS}) + \beta_5 (\text{RH}) + \beta_6 (\text{Temp}) + \varepsilon \qquad (3.4)$$

In the above equations β_o is the intercept, β1 to β6 are the coefficients or slopes and Ɛ is the error term. Furthermore, NO_mean and NO_2_mean are average concentrations of NO and NO_2 from the low-cost sensors, NO_DG and NO_2_DG are the concentrations from the Devonshire Green monitoring station, WS is wind speed (m/s), RH is relative humidity (%) and Temp is the air temperature (^{o}C).

(b) GAMs

GAMs are advanced modelling techniques which are applicable to both normal and non-normal data distribution and do not assume the relationship between response and explanatory variables to be linear. GAMs rather permit the response probability distribution to be any member of the exponential family (e.g., normal, exponential, gamma and poisson distribution). In contrast, a linear model assumes the response distribution to be normal and the relationship between response and explanatory variables to be linear. The GAM models developed in this study are shown in equations 3.5 to 3.8 below, using the same predictors used by LRM shown in equations 3.1 to 3.4.

$$\text{NO\_DG} = s1\ (\text{NO\_mean}) + \varepsilon \qquad (3.5)$$

$$\text{NO}_2\text{\_DG} = s1\ (\text{NO}_2\text{\_mean}\ ) + \varepsilon \qquad (3.6)$$

$$\text{NO\_DG} = s1\ (\text{NO\_mean}) + s2\ (\text{NO}_2\text{\_DG}) + s3\ (\text{NO}_2\text{\_mean}) + s4\ (\text{WS}) + s5\ (\text{RH}) + s6\ (\text{Temp}) + \varepsilon \qquad (3.7)$$

$$\text{NO}_2\text{\_DG} = s1\ (\text{NO\_DG}) + s2\ (\text{NO}_2\text{\_mean}) + s3\ (\text{NO\_mean}) + s4\ (\text{WS}) + s5\ (\text{RH}) + s6\ (\text{Temp}) + \varepsilon \qquad (3.8)$$

In the above models (3.5 to 3.8), s1 to s6 are the smoothing terms (Wood 2006), each one of these is associated with the adjacent explanatory variable. Response or modelled variables are given on the left and the explanatory variables of each model are given on the right of the equations.

3.2.3.3. Models' assessment

To evaluate the models' performance predicted and measured (observed) concentrations were compared. For this purpose, several statistical metrics were calculated including correlation coefficient (r), coefficient of determination (R^2), Root Mean Square Error (RMSE), Normalised Mean Biased (NMB), Factor of two (FAC2) and Coefficient of Efficiency (COE), which are defined by Carslaw (2016) and Sayegh et al. (2014). RMSE provides a good measure of the model error by calculating how close or far the predicted values are to the observed values. NMB estimates average over or under prediction, whereas 'r' is the strength of the linear relationship between two variables (here modelled and observed concentrations). NMB value between +0.02 and -0.02 shows acceptable model performance. We would like 'r' to have a value as close to one (±1) as possible, however generally a value ranging from ±0.5 to ±0.99 indicates reasonably good performance. FAC2 is the fraction of modelled values within a factor of 2 of the observed values. FAC2 should satisfy the condition that $0.5 \leq Mi/Oi \leq 2$, where Mi represents the modelled values and Oi represents the observed values. A highly efficient or perfect model should have COE value of 1, however when analysing real data, a model should have a COE value of less than 1. COE having a zero value (COE = 0) means the model prediction is not better than the mean of the observed value, which in other words means its prediction power is zero; it has no predictive advantage.

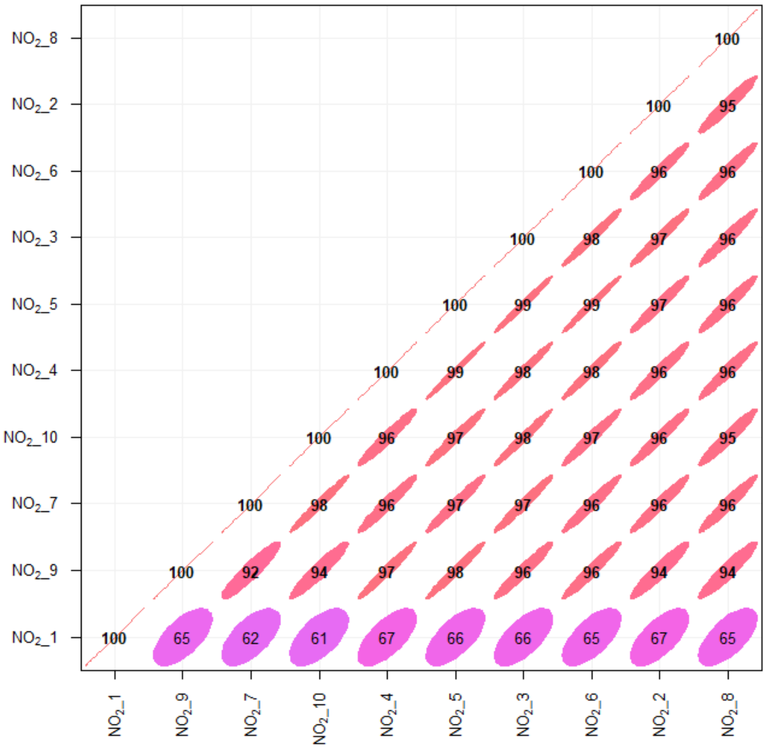
3.3. Results and discussion

3.3.1. Temporal variability and correlation analysis

Hourly average NO_2 (ppb), NO (ppb) and CO concentrations (ppm) measured by ten E-MOTEs seemed reasonable and had an overall mean of about 22 ppb, 10 ppb and 0.35 ppm, respectively. Overall, various air pollutant concentrations showed a similar pattern at different monitoring sites during different seasons, for instance, NO_2 concentration was higher in winter months and lower in summer (time plots not shown for brevity). These seasonal trends are further analysed in coming sections. NO_2 and NO concentrations measured at the Devonshire Green monitoring site also showed higher concentrations in colder months and lower concentrations in warmer months. Obara et al. (2011) and Cai et al. (2016) have reported that air pollutants levels are strongly associated with stable weather conditions, atmospheric inversion, low wind speed and shallow boundary layer, which are generally found in winter seasons in the UK. In such meteorological conditions air pollutants emitted by various sources do not disperse and stay near the emission sources due to poor horizontal and vertical dispersion.

Fig. 3.6 shows correlation plots of hourly average NO_2 (upper-panel), NO (centre-panel) and CO (lower-panel) concentrations collected by the ten E-MOTEs. The correlation coefficient values, ranging from -1 to +1, are normally represented as a decimal number (e.g., 0.xx). However, here to facilitate presentation both zero and decimal points are avoided, following the default format of 'openair' suggested by Carslaw (2016). NO_2 concentrations show very strong positive correlation between various sensors. Mostly correlation coefficients are greater than 0.92 ($r > 0.92$), except sensor-1 (NO_2_1), which shows relatively weaker correlation, with r - values ranging from 0.60 to 0.67. The cause of this weaker correlation is likely due to erroneous data caused by bad communication between the sensor and the gateway. Taking this into account this shows all the E-MOTE measurements of NO_2 are consistent with each other and show strong similarity with each other. This strong similarity puts confidence in the consistency of these sensors. This is the first study reporting the performance of E-MOTEs, therefore no comparison was possible with previous studies. However, several researchers have assessed the performance of other LCS, such as AQMesh pods both in the UK and Europe and reported that their performance varied both spatially and temporally from sensor to sensor (Castell et al. 2017).

In contrast, NO concentrations (Fig. 3.6, middle-panel) showed weaker correlation. NO_5 vs NO_6 and NO_5 vs NO_7 showed strongest correlation with r - value of 0.48 each. NO_6 vs NO_9 show zero r - value, whereas NO_2 vs NO_3 showed negative correlation. Fig. 3.6 (lower-panel) presents correlation plots of CO concentrations showing much stronger correlation than NO concentrations. Except for CO_2 and CO_6, the remaining sensors compared against each other showed r - values greater than 0.90. CO_2 and CO_6 have r - values ranging from 0.35 to 0.64, which are those for CO_2 vs CO_6 and CO_1 vs CO_6, respectively. This confirms that E-MOTESs produce consistent measurements of CO concentration. For further analysis, time variation plots are constructed in the next section to see how the pollutant concentrations vary at various time scales, such as diurnal, weekly and annually.



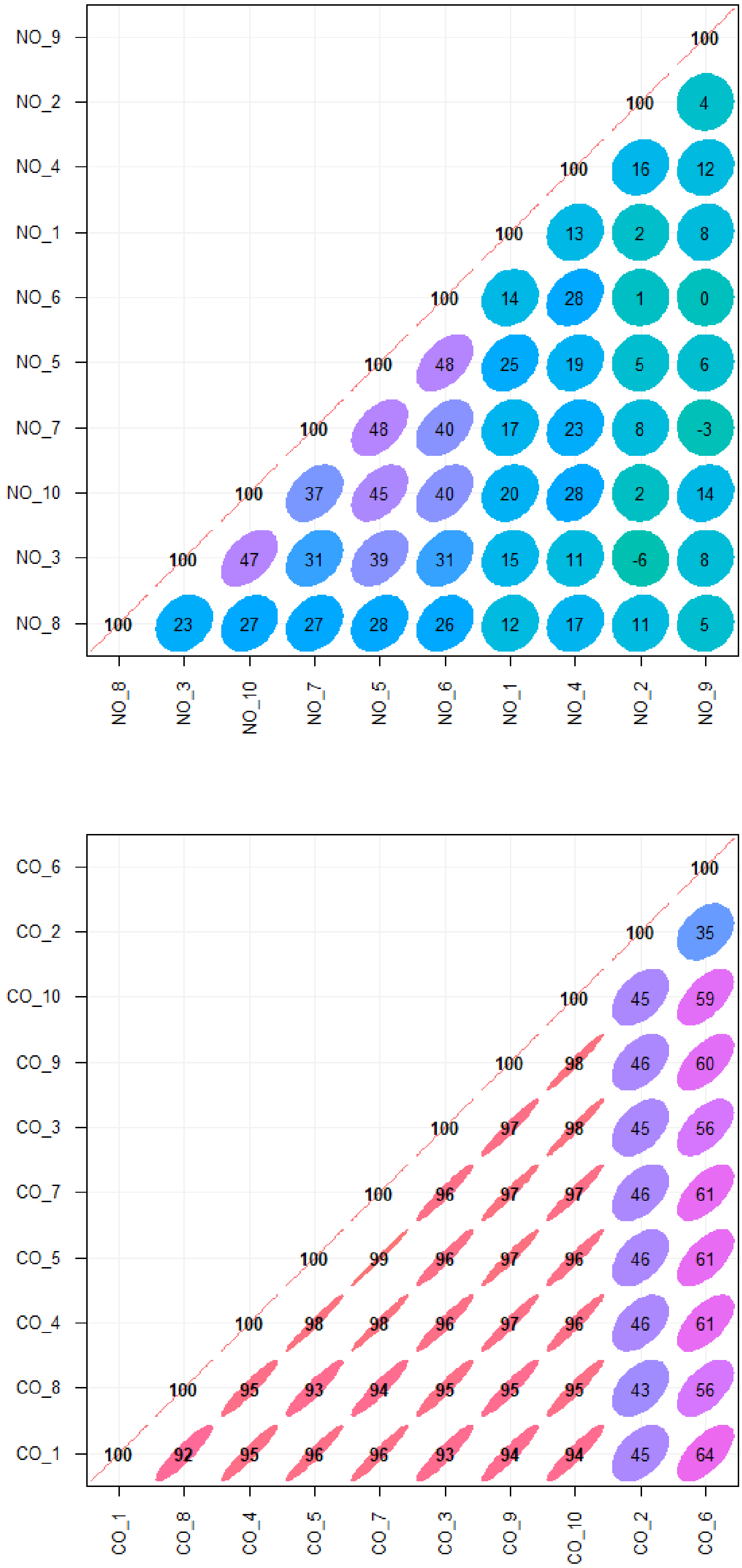


Fig. 3.6: Correlation plots of NO_2 ppb (upper-panel), NO ppb (centre-panel) and CO concentrations (ppm) (lower-panel) from ten E-MOTEs during Oct 2016 to Sept 2017 in

Sheffield. All r – values should have been presented as decimal number, however here both zero and decimal points are avoided to facilitate presentation.

Fig. 3.7 shows time variation plots of NO_2 concentrations (ppb) collected by nine of the E-MOTEs. NO_2_1 was removed due to missing and likely incorrect measurements. These plots show strong similarities among the nine sensors on all time scales i.e. diurnal, weekly and annual cycles. During the diurnal cycle (Fig. 3.7, lower-left-panel) NO_2 concentrations (ppb) start decreasing after mid-night and continue to do so until about 05:00 hours, then slightly increase at about 06:00 - 08:00 hours probably due to morning traffic peak hours. Afterwards, NO_2 levels gradually decrease and reach a minimum level around midday (12:00 hours), most probably due to low traffic activities and atmospheric conditions which help disperse air pollutants quickly. Relatively high temperature, high wind speed and wider atmospheric boundary layer during the afternoon improve both horizontal and vertical air pollutant dispersion. Diurnal cycles of temperature (^oC) and wind speed (m/s) during 2017 at the Devonshire Green monitoring stations are shown in Fig. 3.8, which clearly shows that wind speed and temperature reach the highest levels during the afternoon, which leads to a widening of the atmospheric boundary layer and help disperse locally emitted pollutants. After 14:00 hours NO_2 levels begin increasing and reach their highest levels in response to the evening's busiest traffic hours (about 18:00 - 20:00 hours), when this activity cause pollutant emissions to increase. Furthermore, in the evening the atmosphere is colder and more stable which discourages air pollutants dispersion. The stable atmosphere continues as the night progresses, although, traffic levels decline. This reduction in traffic levels results in a slight decrease in NO_2 levels. It is worth noting that all the sensors produce almost the same temporal pattern on daily basis. Diurnal cycles on individual days (Monday to Sunday) are shown in Fig. 3.7 (upper-panel). Weekly cycles of NO_2 concentrations (ppb) are shown in Fig. 3.7 (lower-right-panel), where a uniform pattern of various sensors can be observed. As expected different traffic patterns during the weekend result in lower levels of NO_2 on Saturday and Sunday.

Annual cycles of NO_2 (Fig. 3.7, lower-middle-panel) are somewhat confusing showing much higher levels of NO_2 during October. It was expected that NO_2 levels would have been higher during the colder months (i.e., November, December and January) and lower during the hotter months (i.e., May, June and July). This is seen in Fig. 3.9, which depicts NO_2 levels measured at the Devonshire Green monitoring station during the same period as shown in Fig. 3.7. Concentrations measured at this location are shown as NO_DG and NO_2_DG, and average concentrations of the E-MOTEs are shown as NO_mean and NO_2_mean. CO is not monitored at this site and therefore comparison with the E-MOTEs was not possible. All E-MOTE sensors have a strong correlation with each other and have the same temporal pattern therefore it is convenient to average their measurements to facilitate comparison with the measurements from the Devonshire Green site. NO_2_mean and NO_mean are closely related with NO_2_DG and NO_DG at diurnal, weekly and annual cycles, however, some differences can be observed at various temporal intervals. To summarise, it can be said that generally E-MOTEs show close similarities with the reference instrument, however there are some dissimilarities at various temporal scales. NO_2 and NO concentrations (ppb) at Devonshire Green produced a smooth annual cycle going down from January to June-July and then going up until December. Such a smooth annual cycle does not exist when mean NO and NO_2 concentrations measured by E-MOTEs were plotted. NO_2_mean showed lowest level in September and highest in October and the clear summer and winter difference demonstrated by Devonshire Green has disappeared here. Overall, the results discussed above are encouraging as they successfully capture the temporal trends of air pollutants and show a consistent performance by showing strong correlation with each other.

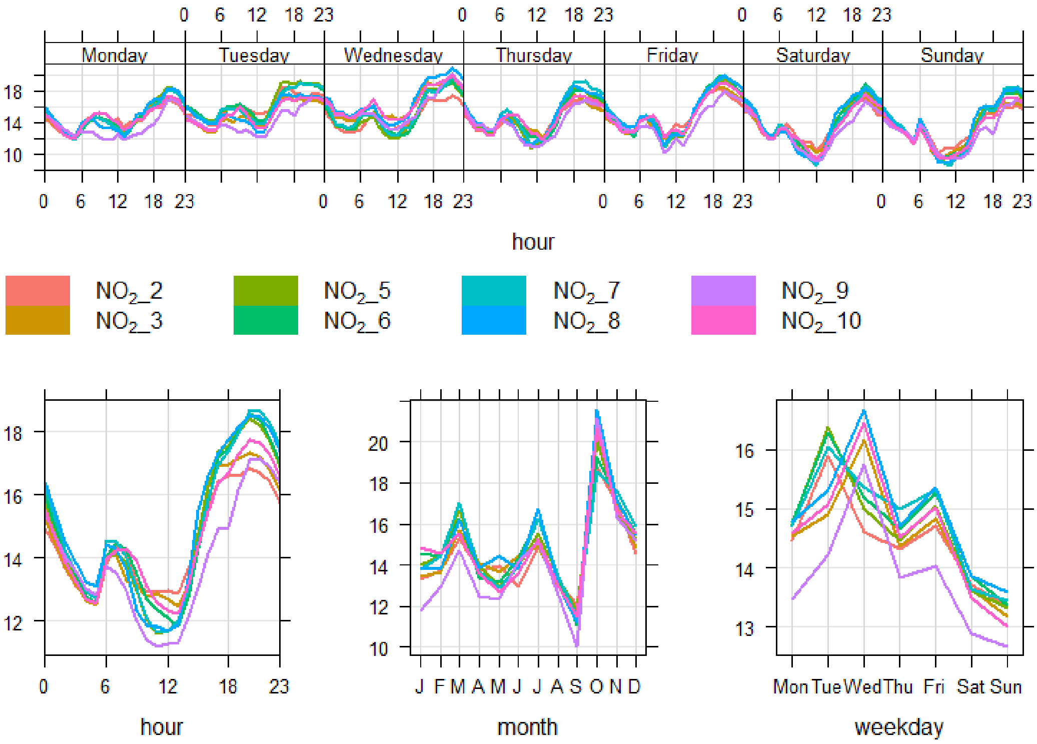


Fig. 3.7: Time variation plots of NO_2 concentrations (ppb) from nine sensors from October 2016 to September 2017 (Readings from one sensor, NO_2_1 were excluded due to missing and erroneous data).

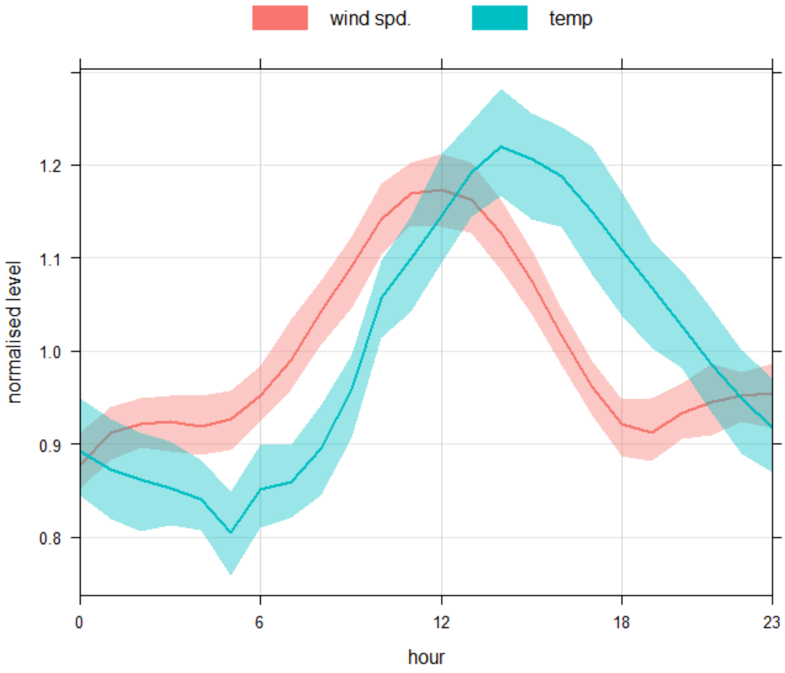


Fig. 3.8: Diurnal cycles of wind speed (m/s) and temperature (°C) at the Devonshire Green monitoring station during 2017, showing highest wind speed and temperature during the afternoon.

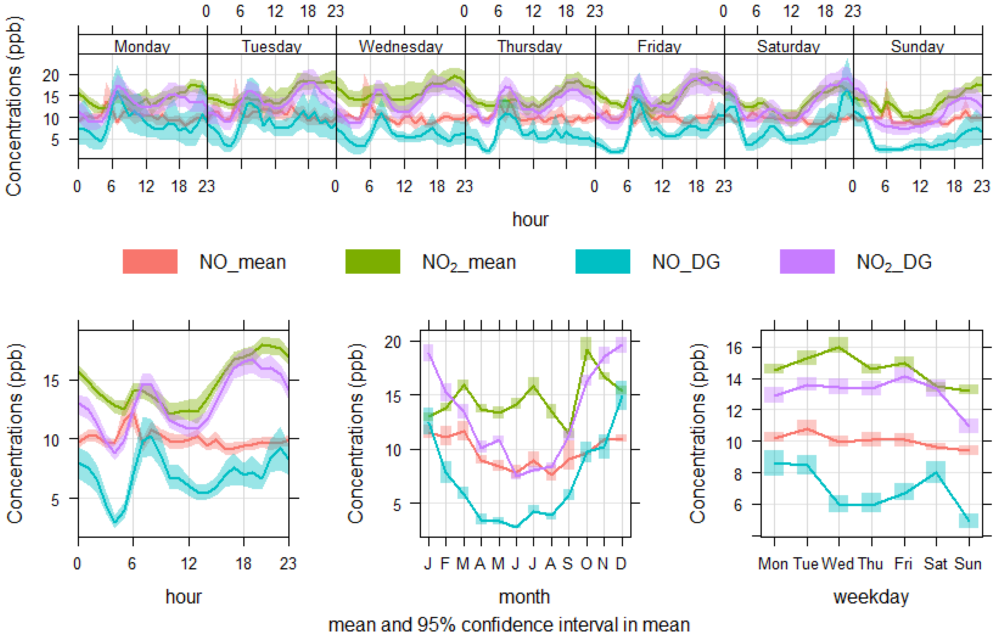


Fig. 3.9: Time variation plots comparing diurnal, weekly and annual cycles of NO_2 and NO at Devonshire Green and the mean of all 10 E-MOTE sensors during Oct 2016 to Sept 2017 in Sheffield.

3.3.2. Modelling

In this section both linear and nonlinear regression modelling approaches are employed and their performances are compared using several statistical metrics.

3.3.2.1. Linear regression models

The outputs of model 3.1 to 3.4 are presented in Table 3.2, showing the values of various statistical metrics. Table 3.2 shows that the multiple linear regression model (MLRM) demonstrated much better performance than the simple linear regression model (SLRM). This was expected as MLRMs used several extra explanatory variables including temperature, wind speed and relative humidity. The values of FAC2, RMSE, R^2, NMB and COE are shown in Table 3.2. The values of NMB demonstrate acceptable model performance since they lie within the range of +0.02 to -0.02 (Table 3.2). The other metrics also signify a small degree of error in the model and good predictability. Fig. 3.10 shows a scatter plot with model lines and shows that most of the points lie between the FAC2 region, which again demonstrates acceptable model performance. It should be noted that these metrics were calculated using the testing data (25% randomly selected) and for the training dataset the values returned for these metrics displayed even better performance (not shown for brevity). This shows that using air quality data measured by LCS and meteorological data as explanatory variables, we can successfully predict (reproduce) NO_2 concentrations measured by reference instruments. Further details of model 3.4 are given in Table 3.3, which shows that all explanatory parameters in the model had highly significant effects (p-value < 0.01) on the response variable. Explanatory variables with positive coefficients (i.e., NO_mean and NO_2_mean), show positive effect on the response variable, whereas the variables with negative coefficients (e.g., temperature and wind speed) show negative effect on the response variable. The negative effect of temperature and wind speed suggests that warmer and windier conditions help disperse locally emitted pollutants and hence decrease NO_2 concentrations. The negative correlation between relative humidity and temperature is well known, therefore relative humidity is showing positive associations with NO_2. Positive association between different NOx species is expected as they have the same emission source and therefore show positive coefficients in Table 3.3. Linear regression is unable to address the nonlinear relationship between response and explanatory variables therefore a nonlinear regression model is employed in the next section to test how it performs in comparison to its linear counterpart.

Table 3.2: Showing the outputs of simple linear regression model (SLRM) and multiple linear regression models (MLRM)

Model	**Response variable**	**Explanatory Variable(s)**	**FAC2**	**RMSE**	**R^2**	**NMB**	**COE**
SLRM	NO_DG	NO_mean	0.98	2.84	0.25	0.002	0.10
SLRM	NO_2_DG	NO_2_mean	0.78	10.15	0.15	0.013	0.05
MLRM	NO_DG	NO_mean, WS, NO_2_mean, RH, NO_2_DG, Temp	0.30	12.79	0.51	0.012	0.12
MLRM	NO_2_DG	NO_mean, WS, NO_2_mean, RH, Temp, NO_DG	0.83	5.76	0.64	0.001	0.41

Table 3.3: Showing various parameters of model 3.4 along with their slopes and p-values

Explanatory Variable	**Coefficient (slopes)**	**Significance value (p-value)**
Intercept	14.74	0.000***
NO_mean	0.125	0.000 ***
NO_DG	0.250	0.000***
NO2_mean	0.168	0.000***
Temp	-0.412	0.000***
WS	-1.219	0.000***
RH	0.026	0.001 **

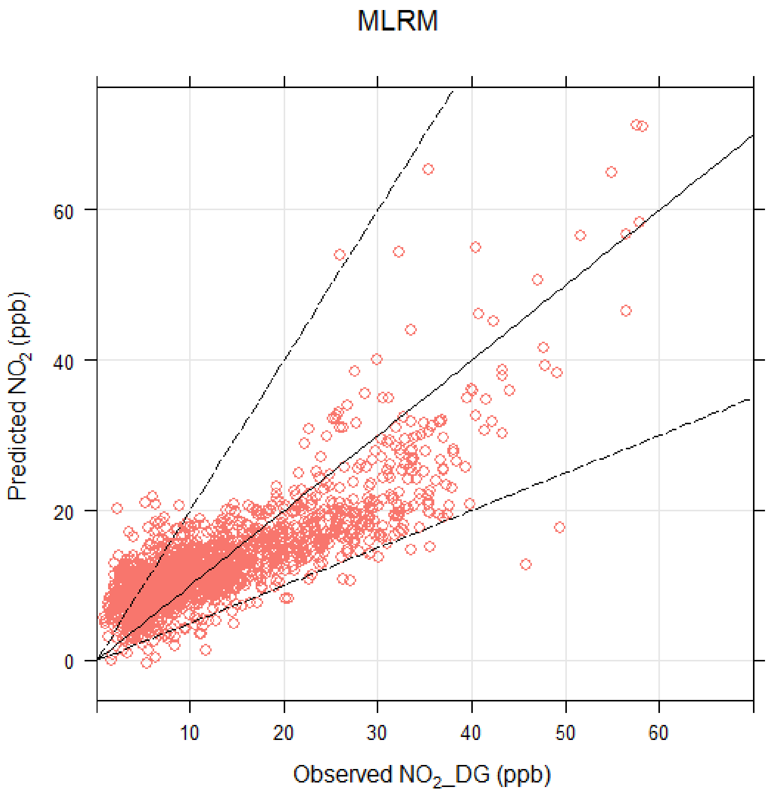


Fig. 3.10: Scatter plot comparing observed and MLRM predicted concentrations of NO_2_DG (ppb) based on the testing data (25% randomly selected), where the solid middle line is the 1:1 line, whereas the upper and lower lines represent 2:1 and 0.5:1, respectively. Most of the points lie within these lines demonstrating acceptable model performance.

3.3.2.2. Generalised Additive Model (GAM)

GAM models are shown in Equations 3.5 to 3.8. After running these models predicted and observed concentrations were compared and several metrics calculated to assess their performance, which are presented in Table 3.4. Comparing Tables 3.2 and 3.4, it can be observed that using the same explanatory variables GAM model performs better and display greater predictability. Comparing these models, model 3.8 showed best performance. Its outputs are shown in Fig. 3.11, which shows how the response variable (NO_2_DG) changes with each explanatory variable. This figure also shows that the association between explanatory variables and response variable (NO_2_DG) is not linear and changes for different values of the explanatory variables. It is interesting to see that the effect of temperature on NO_2 is negative (the curve is downward) until around 20^{o}C is reached, afterwards as temperature increases further the curve turns upward, showing a positive effect, most probably due to the formation of secondary NO_2 in the atmosphere. In contrast, the effect of wind speed results in a downward curve regardless of wind speed, which is probably due to the fact that high wind speed disperses locally emitted pollutants more effectively. GAM models successfully address the non-linear relationship between response and explanatory variables and probably this is the reason that the GAM model performs significantly better than the MLRM, using the same explanatory variables. As an example, let us compare the GAM and MLRM models based on NO_2_DG. GAM has resulted in a high R^2 - value (0.83) and lower RMSE (3.91) than MLRM where the R^2 - value was 0.64 and RMSE was 5.76. This shows that GAM has predicted NO_2_DG more accurately. Fig. 3.12 compares observed and predicted NO_2 and the plot shows a linear association between observed and predicted concentrations with most of the points lying within FAC2 region. All independent variables have highly significant effects ($P < 0.001$) on NO_2_DG. Although GAM shows better performance than MLRM, MLRM are used more often by researchers due to the ease with which it can be applied and interpreted. MLRM models provide a slope for each explanatory variable as it assumes a linear relationship, whereas in the case of GAM the slope changes almost at every point (Fig. 3.11). In real-life situations especially in the case of air quality data, relationships are not always linear, therefore GAM models provide a better option for air quality modelling and display greater predictability as shown in this study. To explain this further, several plots are shown in Fig. 3.13 showing that the association between various air pollutants is not linear. To address the non-linear association, we need a non-linear model. GAM successfully addresses the non-linear association between various air pollutants and so performs better than a linear model. A demonstrative is shown in Fig. 3.13 (lower-right panel), where the value of R^2 is 0.79 for GAM and 0.5 for LRM showing considerable difference in performance of the two models.

Table 3.4: Showing different statistical metrics for the GAM models

Response variable	**Explanatory variable(s)**	**FAC2**	**RMSE**	**R^2**	**NMB**	**COE**
NO_DG	NO_mean	0.98	2.80	0.17	0.014	0.101
NO_2_DG	NO_2_mean	0.80	10.06	0.16	0.012	0.048
NO_DG	NO_mean, NO_2_mean, NO_2_DG, WS, RH, Temp	0.53	9.89	0.70	0.008	0.50
NO_2_DG	NO_mean, NO_2_mean, NO_DG, WS, RH, Temp	0.95	3.91	0.83	0.005	0.614

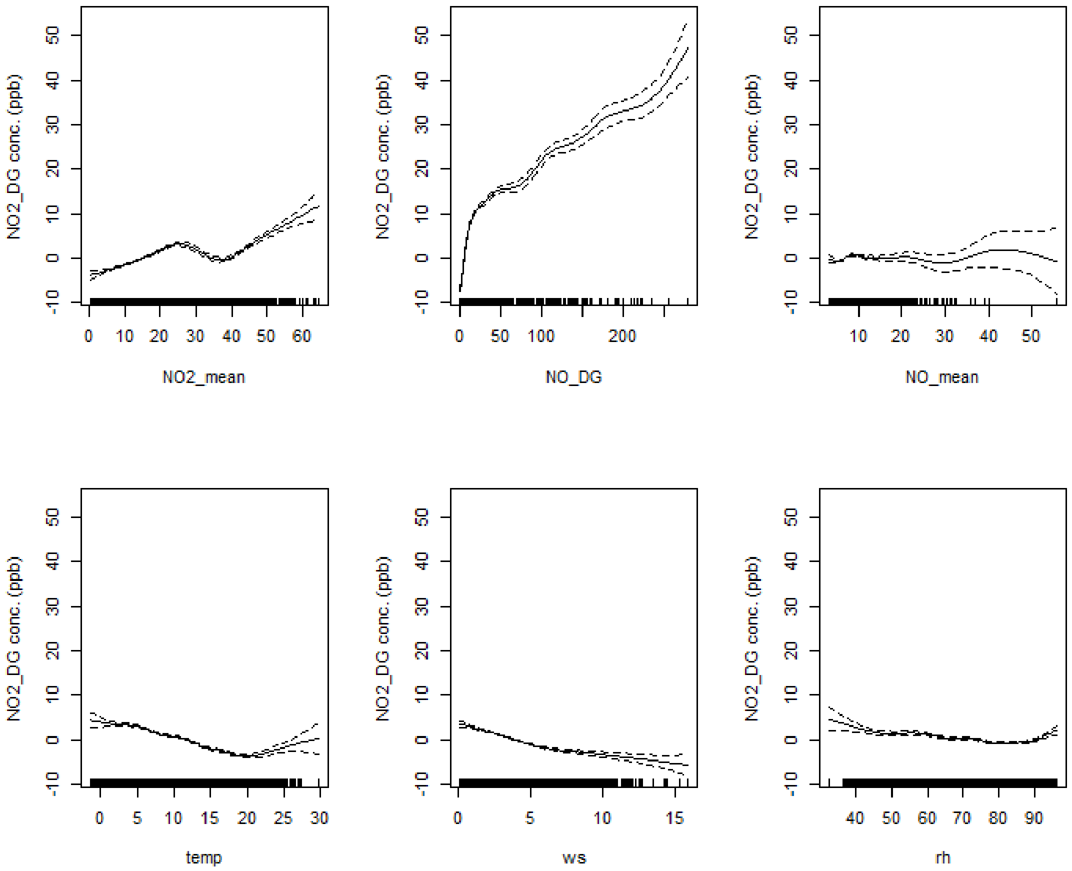


Fig. 3.11: Outputs of GAM model (equation 3.8), in which NO_2_DG (ppb) was used as the response variable and NO_2_mean (ppb), NO_DG (ppb), NO_mean (ppb), Temperature (temp ^{o}C), wind speed (ws m/s) and relative humidity (rh %) were used as explanatory variables. The dashed lines are the estimated 95 % confidence interval, whereas the vertical short lines on the x-axis show the data presence.

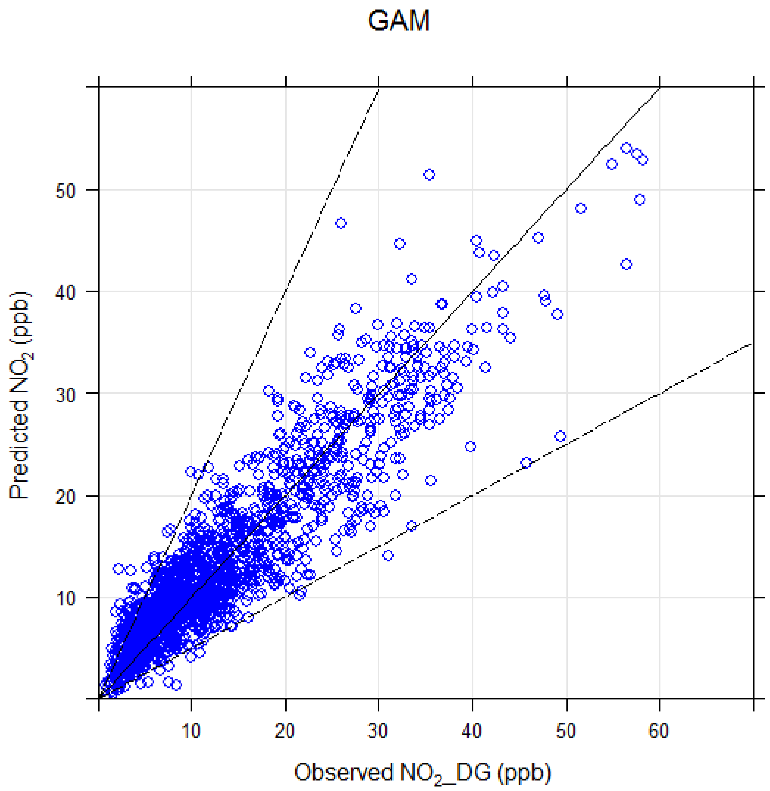
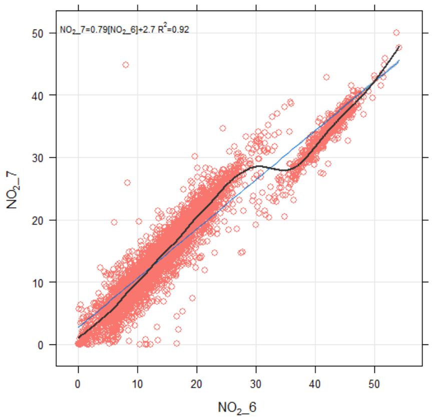
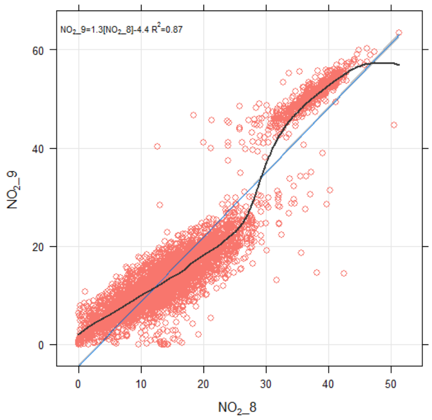


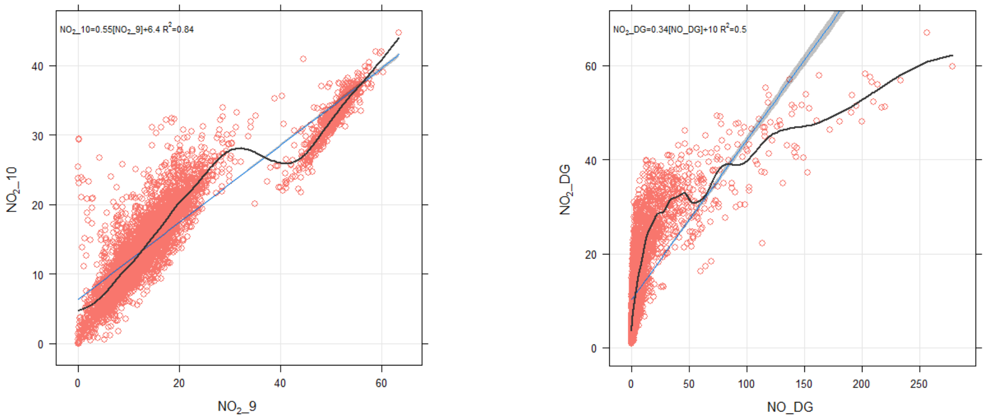
Fig. 3.12: Scatter plot comparing observed and GAM predicted concentrations of NO_2_DG (ppb) based on the testing data (25% randomly selected), where the solid middle line is the 1:1 line, whereas the upper and lower lines are 2:1 and 0.5:1 lines respectively. The dashed lines show within the factor of two regions. Most of the points lie within these lines showing an acceptable model performance.



R^2 for GAM = 0.95 and LRM = 0.92



R^2 for GAM 0.9 and LRM 0.87



R^2 for GAM 0.91 and LRM 0.84 R^2 for GAM 0.79 and LRM 0.5

Fig. 3.13. Comparing the performance of linear (LRM) and non-linear (GAM) models.

3.3.3. Further discussion of low-cost sensors

Castell et al. (2016) have evaluated the performance of the AQMesh sensors measuring gaseous air pollutants (e.g., NOx, CO, and O_3) and particulate matter (PM_{10} and $PM_{2.5}$) in Oslo, Norway. They performed the evaluation of both outdoors and under indoor laboratory conditions. They considered several types of emissions and environmental conditions such as roadside traffic and urban background over a 6-month period (April to September, 2015). Castell et al. (2016) concluded that good performance of the low-cost sensors in the laboratory does not imply similar performance when sited outdoors. Therefore, to reduce uncertainties, sensors must be calibrated in outdoor field locations. They also concluded that there is a lack of adequate outdoor testing of the sensors by the manufacturers before marketing such sensors, which can lead to poor performance and misleading data, which is of great concern, especially when members of the public use such instruments without scientific supervision to collect and interpret air quality data.

Borrego et al. (2016) compared the performance of several LCS with reference instruments from the 13th to 27th October 2014 and reported that for measuring O_3 AQMesh and NanoEnvi sensors had the lowest errors and higher coefficient of determination ($R^2 > 0.70$), whereas ENEA Air-Sensors, ISAG and Cambridge SNAQ showed poor performance with $R^2 < 0.2$. To measure the levels of NO_2, Borrego et al. (2016) compared the performance of six platforms, where the highest correlation and lowest errors were shown by AQMesh, ECN-Airbox and Cambridge University SNAQ with $R^2 > 0.80$ and mean biased error (MBE) close to zero. In contrast, ENEA Air-Sensors and AUTh-ISAG AQ Microsensors demonstrated very poor correlation ($R^2 < 0.1$). For measuring the levels of CO, AQMesh and Cambridge University SNAQ had the highest correlation ($R^2 > 0.80$) with reference instruments, whereas the performance of the rest of the sensors was also satisfactory ($R^2 > 0.50$) (Borrego et al. 2016). For monitoring NO, AQMesh and Cambridge University SNAQ were compared, where AQMesh showed better correlation ($R^2 = 0.80$) than Cambridge University SNAQ ($R^2 = 0.30$). For measuring PM_{10}, all sensors showed poor correlation with reference instruments, with $R^2 = 0.36$ being the highest which was observed with the ECN Air-box (Borrego et al. 2016). The ECN Air-box also showed the highest correlation ($R^2 = 0.27$) with reference instruments for measuring $PM_{2.5}$, the other sensors had lower R^2-values.

Castell et al. (2017) compared the measurements from 24 AQMesh sensors against reference instruments and reported that the quality of the data obtained from the LCS were questionable. The performance of the sensors varied both spatially and temporally and was dependent on the atmospheric composition and meteorological conditions, such as temperature and relative humidity. Furthermore, Castell et al. (2017) reported that the performance varied from unit to unit therefore it is necessary to check the data quality of each pod separately before use. The sensors installed in the laboratory showed much stronger correlation ($R^2 > 0.95$ for all pollutants) with reference instruments than those installed outdoors, where the average R^2 values were 0.60, 0.86, 0.49, 0.54, 0.56 and 0.51 for CO, NO, NO_2, O_3, PM_{10} and $PM_{2.5}$, respectively. Air quality data collected by means of LCS are suitable for promoting air quality awareness, general information and for highlighting air pollution hotpots however the data are not suitable for air quality compliance and research, especially for assessing health and environmental impacts of air pollution (Castell et al. 2017). Dongol (2015) has also concluded that air quality data collected by LCS cannot be used for air quality regulatory purposes and for other purposes where highly accurate data are required. Therefore, Lewis and Edward (2016) state there is a need for further legislation to regulate the usability of data obtained from low-cost sensors.

Referring to the uncertainties in air quality data collected by LCS, Lewis and Edward (2016) have commented that the recent introduction of these sensors for monitoring public exposure to air pollution are generating a large volume of data, which remain mostly untested, and therefore their quality is questionable and will create difficulty for air quality managers and planners in the future. Furthermore, Lewis and Edward (2016) mentioned that these sensors show stability and sensitivity issues and that the sensors' readings are subject to interference from other long-lived air pollutants, e.g., CO_2 and H_2 and prevailing meteorological conditions like relative humidity, temperature and wind speed. The lower-cost sensors perform better when air pollutant levels are high (Lewis and Edward 2016). The lower-cost sensors have potential to measure air pollutant levels in places where traditional monitoring was not previously possible. They are portable, cheaper, and can provide much better spatial and temporal coverage in real-time, providing more localized and timely warnings to the public.

Lewis et al. (2016) have shown that one potential solution to reduce the uncertainties of air quality data obtained by using this class of sensors is by applying supervised machine learning techniques, such as the Boosted Regression Tree (BRT) model. Spinelle et al. (2017) applied three approaches for calibrating the concentration of NO_2, CO, and CO_2. The methods were linear regression, multiple linear regression and a supervised machine learning technique (artificial neural network). Using simple linear regression only the reference concentration was used as an explanatory variable, whereas in the other models relative humidity and temperature were also used. Supervised learning technique showed better performance than the other two models. The finding of this current study agrees with the above previous studies and show that the quality of NO_2 concentrations measured by LCS can be much improved by applying supervised machine learning techniques based on GAM.

3.4. Conclusions

LCS have the potential to contribute to real-time air quality monitoring networks installed to-date as this type of sensors are cheap, compact, user-friendly and provide high-resolution spatiotemporal measurements of air pollutant concentrations. However, these sensors have limitations, therefore the sensors require outdoor calibration and the data obtained from these sensors require further processing employing advanced statistical modelling approaches, such as GAM. In this paper, air pollutant data from ten Envirowatch E-MOTEs were compared with each other and with reference instruments. The sensors were able to capture the diurnal, weekly and annual cycles of air pollutant concentrations with some discrepancies. NO_2 and CO showed stronger correlation between various sensors, where most of the correlation coefficients were greater than 0.9, however NO showed relatively weaker correlation between the various sensor locations. NO_2 concentrations showed very strong positive correlation between various sensors. Mostly correlation coefficients (r-values) were greater than 0.92. CO from different sensors also had r-values mostly greater than 0.92, however, NO showed r-value less than 0.5. Several linear and non-linear models were developed for sensor calibration and for predicting NO_2_DG and NO_DG concentrations using NO_mean and NO_2_mean and meteorological parameters as explanatory variables. GAM models demonstrated better performance by exhibiting stronger similarity (e.g., greater correlation coefficient and FAC2 values) and lower error (e.g., weaker RMSE and NMB) between observed and modelled concentrations of NO and NO_2. GAM models were able to capture the non-linear association between various air pollutants and performed better than linear models. The best GAM developed for reproducing NO_2 concentrations returned values of 0.95, 3.91, 0.81, 0.005, and 0.61 for Factor of two (FAC2), Root Mean Square Error (RMSE), coefficient of determination (R^2), Normalised Mean Biased (NMB) and Coefficient of Efficiency (COE), respectively. Therefore, GAM models are recommended for LCS calibration and for reproducing measured NO_2. In the coming projects, we intend to deploy a more dense network of LCS in the whole city of Sheffield to collect high resolution spatial and temporal air quality data. We also aim to improve experimental designs of the sensors network, test other sensor technologies and identify new calibration approaches for better performance in the future.

Acknowledgement

This project is funded by EPSRC and Siemens Ltd as a part of the first authors' PhD study. We acknowledge their financial support.

3.5. References

Aleixandre, M., & Gerboles, M. (2012). Review of small commercial sensors for indicative monitoring of ambient gas. Chemical Engineering Transactions, 30, 169–174.

Borrego, C., Costa, A.M., Ginja, J., Amorim, M., Coutinho, M., Karatzas, K., Sioumis, Th., Katsifarakis, N., Konstantinidis, K., De Vito, S., Esposito, E., Smith, P., Andre, N., Gerard, Francis, P., Castell, L.A., Schneider, N., Viana, P., Minguillon, M., Reimringer, M.C., Otjes, W., von Sicard, R.P., Pohle, O., Elen, R., Suriano, B., Pfister, D., Prato, V., Dipinto, M., & Penza, M. (2016). Assessment of air quality microsensors versus reference methods: The EuNetAir joint exercise. Atmospheric Environment, 147, 246-263.

Cai, Z., Jiang, F., Chen, J., Jiang, Z., & Wang, X. (2016). Weather Condition dominates Regional PM2.5 Pollutions in the Eastern Coastal Provinces of China during winter. Aersols and air quality research, 18: 969–980.

Carruthers, D., Clarke, D., Dicks, K.J., Freshwater, R.A., Jackson, M., Jones, R.L., Lad, C., Leslie, I., Lewis, A. J., Lloyd, H., Popoola, O.A.M., Randle, A., & Ulrich, S. (2016). Using a commercial low-cost sensor network (AQMesh) to quantify urban air quality:

comparing measured and modelled (ADMS-urban) pollutant concentrations. Air Quality Monitoring: Evolving Issues and New Technologies Conference, December 13th, 2016, Automation and Analytical Management Group - Royal Society of Chemistry.

Carslaw, D.C. (2016). The openair manual — open-source tools for analysing air pollution data. Manual for version 1.1-4, King's College London.

Castell, N., Dauge, F.R., Dongol, R., Vogt, M., & Schneider, P. (2016). Uncertainty in air quality observations using low-cost sensors. Geophysical Research Abstracts, 18: 5692. EGU General Assembly 2016.

Castell, N., Dauge, F.R., Schneider, P., Vogt, M., Lerner, U., Fishbain, B., Broday, D., & Bartonova, A. (2017). Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? Environment International, 99, 293 – 302.

Castell, N., Kobernus, M., Liu, H.Y., Schneider, P., Lahoz, W., Berre, A.J., & Noll, J. (2015). Mobile technologies and services for environmental monitoring: the Citi-Sense-MOB approach. Urban Climate, 14 (3), 370-382.

DEFRA, 2015. Improving air quality in the UK: Tackling nitrogen dioxide in our towns and cities, UK overview document, December 2015.

Dongol, R. (2015). Evaluation of the Usability of Low-cost Sensors for Public Air Quality Information. Master's Thesis, Department of Informatics Programming and Networks, University of Oslo.

Hamm, N.A.S., Van Lochem, M., Hoek, G., Otjes, R.P., Van der Sterren, S., & Verhoeven, H. (2016). The Invisible Made Visible: Science and Technology. link. springer.com/book/10.1007/978-3-319-26940-5.

Hastie, T.J., & Tibshirani, R.J. (1990). Generalised Additive Models, Chapman & Hall, London.

Heimann, I., Bright, V.B., McLeod, M.W., Mead, M.I., Popoola, O.A.M., Stewart, G.B., & Jones, R.L. (2015). Source attribution of air pollution by spatial scale separation using high spatial density networks of low cost air quality sensors. Atmospheric Environment, 113, 10-19.

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

Lewis A. C., Lee J. D., Edwards P. M., Shaw M. D., Evans M. J., Moller S. J., Smith K. R., Buckley J. W., Ellis M., Gillot S. R., & Whited A. (2016). Evaluating the performance of low cost chemical sensors for air pollution research. Faraday Discussions, 189, 85-103.

Lewis, A., & Edwards, P. (2016). Validate personal air-pollution sensors. Nature 535 (7610), 29–31.

Mead, M.I., Popoola, O. a. M., Stewart, G.B., Landshoff, P., Calleja, M., Hayes, M., Baldovi, J.J., McLeod, M.W., Hodgson, T.F., Dicks, J., Lewis, A., Cohen, J., Baron, R., Saffell, J.R., & Jones, R.L. (2013). The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks. Atmospheric Environment, 70, 186-203.

Munir, S., Habeebullah, T.M., Seroji, A.R., Morsy, E.A., Mohammed, A.M.F., Saud, W.A., Abdou, A., & Awad, A.H. (2013). Modelling Particulate Matter concentrations in Makkah, Applying a Statistical Modelling Approach. Aerosols and Air Quality Research, 13 (3), 901 – 910.

Obara, P.G., Roberts, C.L., Young, C.H., & Williams, C.D. (2011). Validating the correlation of traffic-associated hydrocarbon and nitrogen dioxide with distance from a trunk road within a rural environment in UK. Microchemical Journal, 99, 138–144.

Penza, M., Suriano, D., Villani, M.G., Spinelle, L., & Gerboles, M. (2014). Towards air quality indices in smart cities by calibrated low-cost sensors applied to networks. In: IEEE SENSORS 2014.

Popoola, O., Mead, I., Bright, V., Baron, R., Saffell, J., Stewart, G., Kaye, P., & Jones, R. (2013). A portable low-cost high density sensor network for air quality at London Heathrow airport. In: EGU General Assembly 2013, Held 7-12 April, 2013 in Vienna, Austria id. EGU2013-1907. http://wwwdev.snaq.org/posters/EGU_OAMP_2013.pdf.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reis, S., Cowie, H., Riddell, K., Semple, S., Steinle, S., Apsley, A., & Roy, H. (2013). Urban air quality citizen science. Phase 1: review of methods and projects. Publishers Scottish Environment Protection Agency.

Sayegh, A.S., Munir, S., & Habeebullah, T.M. (2014). Comparing the Performance of Statistical Models for Predicting PM_{10} Concentrations. Aerosol and Air Quality Research, 14 (3), 653-665.

Snyder, E., Watkins, T., Solomon, P., Thoma, E., Williams, R., Hagler, G., Shelow, D., Hindin, D., Kilaru, V., & Preuss, P. (2013). The changing paradigm of air pollution monitoring. Environmental Science Technology, 47 (20), 11369-11377.

Spinelle, L., Gerboles, M., Villani, M.G., Aleixandre, M., & Bonavitacola, F. (2017). Field calibration of a cluster of low-cost commercially available sensors for air quality monitoring. Part B: NO, CO and CO2. Sensors and Actuators B, 238, 706–715.

Stojanovic, M., Bartonova, A., & Topalovic, D. (2015). On the use of small and cheaper sensors and devices for indicative citizen-based monitoring of respirable particulate matter. Environmental Pollution, 206, 696-704.

Suriano, D., Prato, M., Pfister, V., Cassano, G., Camporeale, G., Dipinto, S., & Penza, M. (2015). Stationary and Mobile Low-Cost Gas Sensor-Systems for Air Quality Monitoring Applications. Fourth Scientific Meeting EuNetAir, Linkoping University, Linkoping, Sweden. DOI:10.5162/4EuNetAir2015/15.

Van den Bossche, J., Peter, J., Verwaeren, J., Botteldooren, D., Theunis, J., & De Baets, B. (2015). Mobile monitoring for mapping spatial variation in urban air quality: development and validation of a methodology based on an extensive dataset. Atmospheric Environment,105, 148-161.

Viana, M., Rivas, I., Reche, C., Fonseca, A.S., Perez, N., Querol, X., Alastuey, A., Alvarez-Pedrerol, M., & Sunyer, J. (2015). Field comparison of portable and stationary instruments for outdoor urban air exposure assessments. Atmospheric Environment, 123, 220-228.

WHO (2013). Health effects of particulate matter, policy implications for countries in eastern Europe, Caucasus and central Asia. Publications of WHO Regional Office for Europe UN City, Marmorvej 51 DK-2100 Copenhagen, Denmark.

WHO (2014). WHO media centre. 7 million premature deaths annually linked to air pollution: http://www.who.int/mediacentre/news/releases/2014/air-pollution/en/

Wood, S.N. (2006). Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC.

Wood, S.N. (2017). *Generalized Additive Models: an introduction with R (2nd edition)*, CRC.

CHAPTER 4: STRUCTURING AN INTEGRATED AIR QUALITY MONITORING NETWORK IN LARGE URBAN AREAS – DISCUSSING THE PURPOSE, CRITERIA AND DEPLOYMENT STRATEGY

Said Munir[a,*], Martin Mayfield[a], Daniel Coca[b], Stephen A Jubb[b]

[a]Department of Civil and Structural Engineering, the University of Sheffield, Sheffield, S1 3JD, UK
[b]Department of Automatic Control and Systems Engineering, the University of Sheffield, Sheffield, S1 3JD, UK
*corresponding author (smunir2@sheffield.ac.uk), Mob: +447986001328, Fax: +44 (0) 114 222 5700

Published: Atmospheric Environment: X 2(2019), 100027
https://doi.org/10.1016/j.aeaoa.2019.100027
Received: 29 June 2018; Received in revised form: 4 March 2019; Accepted: 9 March 2019.

Abstract

Air pollution in large urban areas has become a serious issue due to its negative impacts on human health, building materials, biodiversity and urban ecosystems in both developed and less-wealthy nations. In most large urban areas, especially in developed countries air quality monitoring networks (AQMN) have been established that provide air quality (AQ) data for various purposes, e.g., to monitor regulatory compliance and to assess the effectiveness of control strategies. However, the criteria of structuring the network are currently defined by single questions rather than attempting to create a network to serve multiple functions. Here we propose a methodology supported by numerical, conceptual and GIS frameworks for structuring AQMN using social, environmental and economic indicators as a case study in Sheffield, UK. The main factors used for air quality monitoring station (AQMS) selection are population-weighted pollution concentration (PWPC) and weighted spatial variability (WSV) incorporating population density (social indicator), pollution levels and spatial variability of air pollutant concentrations (environmental indicator). Total number of sensors is decided on the basis of budget (economic indicator), whereas the number of sensors deployed in each output area is proportional to WSV. The purpose of AQ monitoring and its role in determining the location of AQMS is analysed. Furthermore, the existing AQMN is analysed and an alternative proposed following a formal procedure. In contrast to traditional networks, which are structured based on a single AQ monitoring approach, the proposed AQMN has several layers of sensors: Reference sensors recommended by EU and DEFRA, low-cost sensors (LCS) (AQMesh and Envirowatch E-MOTEs) and IoT (Internet of Things) sensors. The core aim is to structure an integrated AQMN in urban areas, which will lead to the collection of AQ data with high spatiotemporal resolution. The use of LCS in the proposed network provides a cheaper option for setting up a purpose-designed network for greater spatial coverage, especially in low- and middle-income countries.

Keywords: air quality monitoring, low-cost AQ sensors, AQ network, sensors deployment, Sheffield.

4.1. Introduction

Air pollution is one of the most serious current threats to health, killing 6.4 million people in 2015 worldwide both in developed and less-wealthy nations (Landrigan, 2016). Air pollution is causing various health problems including respiratory problems, cardiovascular diseases, lung cancer and asthma (WHO, 2013). Particulate matter and nitrogen dioxide (NO_2) pollution may cause premature deaths and hospital admissions for conditions such as cardiovascular problems, allergic reactions and lung cancer (Walters and Ayres, 2001). Air pollution is particularly harmful for children, people with existing health problems and the elderly (Khallaf, 2011). Furthermore, air pollution may reduce visibility, damage historical buildings and monuments, affect vegetation and reduce crop yield and quality (Khallaf, 2011; Ivaskova et al., 2015). Air pollution is considered more of a serious problem in large urban areas in both developed and less-wealthy nations (Brunt et al., 2016; DEFRA, 2015). This is due to the fact that urban areas possess greater numbers of emission sources and densely built-up areas including tall buildings and street canyons which hinder dispersion of locally emitted air pollutants (Wu et al., 2017). Urban areas have various emission sources, e.g., road traffic, point emissions and area emissions emitting high volume of both gaseous (e.g., NO, NO_2, CO, SO_2, and H_2S) and particle pollutants (e.g., PM_{10} and $PM_{2.5}$) (DEFRA, 2017). In addition, more people are exposed to air pollution due to high population density in urban areas.

To improve air quality (AQ), the first step is to improve air quality monitoring networks (AQMN) in large urban areas as the current networks are not dense enough for developing high resolution maps and highlighting local micro-level drivers of air pollution (Castell et al., 2017; Schneider et al., 2017). Urban areas exhibit much greater spatial variability in air pollution levels, which require a dense (ubiquitous) AQMN. Traditional and more accurate AQ monitoring instruments are expensive to purchase and maintain, therefore it is not practical to set up a dense network to capture local-scale spatial variability in air pollution concentrations (Castell et al., 2017; Schneider et al., 2017). Traditional AQMN are sparse having few sites widely spaced around a city therefore historically no or little attention has been paid to selecting monitoring sites by following a formal approach to provide spatial coverage and deploy AQ sensors in various environmental types (e.g., roadside, kerbside, urban and suburban background and green spaces). The current literature lacks a rigorous methodology for determining locations of AQMS (Hoek et al., 2008; Wu et al.,2017).

Assessment of the spatial representativeness of air quality monitoring station (AQMS) is an important subject and is linked to health risk assessment, population exposure to air pollution, the design of AQMN, AQ modelling and data assimilation (Kracht et al., 2017; Martin et al., 2015). The spatial representativeness of a monitoring site is related to the variability of pollutants concentrations around that site (Righini et al, 2014). However, scientific literature and European regulation lacks a clear definition and unified agreement for determining the spatial representativeness of an AQMS. Santiago et al. (2013) have reported that due to the complexity of urban meteorology and emissions distribution, AQ in urban areas cannot be assessed with confidence using only air pollutant measurements from a monitoring station. Air pollution levels estimated by street scale dispersion models and maps of population density and residence time can be used to get a more complete and precise view of the air pollution conditions. To analyse the spatial representativeness of urban AQMS and to complement their measured concentrations, Santiago et al. (2013) have developed a methodology using a set of computational fluid dynamics simulations based on Reynolds-Averaged Navier–Stokes equations (CFD-RANS) for different meteorological conditions in two urban areas Pamplona and Madrid in Spain. They defined the representativeness area of AQMS, as the area where concentrations were within an interval of ±20% of the pollutant concentrations at the monitoring station. Righini et al. (2014) presented a methodology to assess spatial representativeness of an AQMS by analysing the spatial variation of emissions around it. Spatial variability of several air pollutants was carried out using a neighbourhood statistic function in a Geographic Information System (GIS). Low variability of emission around a site showed high spatial representativeness of that site and vice versa. To detect spatial representativeness of several urban background or rural background monitoring sites the methodology was applied in Northern and Central Italy.

There are two types of AQMN: routine networks and purpose-designed monitoring networks (Hoek et al., 2008). They both have their pros and cons. Routine monitoring networks are designed mainly for assessing AQ compliance with regulatory standards. An example of such a network is the Automatic Urban and Rural Network (AURN) in the UK. The AQ monitoring sites in the AURN are continuous sites and most of them have been running over a long period of time (ten years or more), however, the monitoring stations are sparse and not suitable for urban scale modelling and mapping. Purpose-designed monitoring networks are set up for a particular purpose, e.g., for developing a land-use regression (LUR) model or for developing urban scale air pollution maps. In these types of networks, the designers (researchers) have control over the selection of the types and number of monitoring sites required for the purpose, however they could be extremely expensive (Hoek et al., 2008) and unaffordable especially in low- and middle-income countries. Therefore, due to lack of funding many researchers have been using data from the routine (e.g., AURN) networks. More recently due to the introduction of low-cost sensors (LCS) (Schneider et al., 2017; Castell et al., 2017; Borrego et al., 2016; Lewis et al., 2016; Lebret et al., 2000; Goswami et al., 2002) several academic and research organisations have set up purpose-designed AQMN for urban scale modelling and exposure assessment. However, these studies have not followed formal procedures for allocating sites, which means performing the GIS calculation of various variables on the basis of which the sites are selected, e.g., population density, pollution concentrations and their spatial variability. In many studies sites are selected informally i.e. without the GIS calculation of various variables or on an ad hoc basis (favouring the placement of monitors in traffic hot spots or in areas deemed subjectively to be of interest). This is because using a formal procedure requires a significant amount of data on the required variables, which are normally not available.

In this study, we intend to structure an AQMN in Sheffield utilising population-weighted air pollution concentration (PWPC) and weighted spatial variation (WSV), which incorporate population density (social indicator), pollution concentrations and spatial variability of pollution concentrations (environmental indicator). Furthermore, a conceptual model is presented which in addition to PWPC and WSV considers the opinions of experts having local experience and understanding of the purpose of AQ monitoring. This study proposes an integrated air quality monitoring network (IAQMN) in Sheffield using a multipurpose and more robust approach aimed at providing maximum value from a network of sensors by adopting an integrated approach which draws on multiple data sources and techniques to inform decision making. The procedure proposed in this study is based on numerical equations using data of population density, pollution concentrations and their spatial variability. The approach can be applied anywhere and shouldn’t be susceptible to failure due to changes in location or time. The proposed network integrates various layers of AQ monitoring techniques including reference sensors which are the most accurate and are recommended by EU and DEFRA for AQ monitoring, LCS (AQMesh and Envirowatch E-MOTEs), and very low-cost IoT (Internet of Things) sensors. These sensors will be deployed as fixed stations in various layers and mounted on vehicles (mobile monitoring). The aim of the proposed AQ network is to collect AQ data of high spatiotemporal resolution to be used in local-scale high resolution mapping and modelling as a case study in Sheffield. This case study will provide a great example of a purpose-designed monitoring network using mostly LCS, especially for low- and middle-income countries where such networks don’t exist due to the high purchase and maintenance cost of reference AQ instruments.

4.2. Methodology

This paper proposes an IAQMN in urban areas using the city of Sheffield as a case study. The aim is to structure a multipurpose robust and systematic approach based on formal procedure utilising numerical, GIS and conceptual modelling techniques. In the proposed network AQMS are mainly selected on the basis of PWPC and WSV of air pollutant concentrations. PWPC accounts for population density and air pollution concentrations, whereas WSV is the factor of spatial gradients of air pollutant concentrations. In this way, the site selection criteria integrate population density, pollution concentrations and spatial variability of both population and pollution levels. Furthermore, a conceptual model is provided, which in addition to PWPC and WSV focuses on the purpose of the monitoring network, opinion of experts with local experience and financial resources (budget of the project) to determine the number of monitoring sites and to select their locations. The total number of sensors is decided on the basis of project budget, whereas the number of sensors in each output area is a factor of WSV. Once the sensors are deployed and data collected, we will analyse the data to assess spatial representativeness of the sites, which can help us decide how many sensors are redundant and how many more sensors are required in areas where spatial variability of air pollution has not been captured.

Air quality management areas (AQMA) are declared mainly on the basis of NO_2 and PM_{10} levels, which are the cause of primary concern in the urban areas of Sheffield and therefore the project focuses on these two pollutants. However, the measurements of other pollutants (e.g., O_3, CO and SO_2) and meteorological parameters (e.g., wind speed and direction, relative humidity and temperature) will be used to analyse the chemistry and dispersion of air pollutants, which will further help to determine the main drivers of air pollution in Sheffield. In this project the network is designed according to the spatial variability of NO_2. Each pollutant has different spatial variability, therefore a network designed based on the spatial variability of another pollutant (e.g., PM_{10}) will have different characteristics.

In this project the intention is to make use of several layers of AQ sensors including both static (fixed) and mobile monitoring to provide AQ data for high spatial and temporal resolution AQ maps. AQ sensors are installed in vehicles, known as MOBIle Urban Sensing (MOBIUS) vehicle. The monitoring only takes place when the vehicle is stationary. The vehicle is driven to the intended location, parked safely and then the monitoring equipment is turned on. AQ monitoring is not carried out when vehicle is in motion. These layers are shown in Figure 4.1 and their main features are given in Table 4.1. The types of AQ sensors employed include reference sensors, LCS and IoT sensors. IoT sensors are miniature electronic devices that are comprised of sensors, microprocessors and communication integrated circuits that are able to detect changes in the environment. IoT sensors are generally much cheaper, lighter and smaller than the LCS. Generally, their prices are a few tens of pounds for a single pollutant sensor. The quality of data collected by IoT sensors is inferior to the LCS and reference sensors. LCS are more compact, portable and use less power when compared to reference instruments. However, they are larger in size and have much better accuracy than IoT sensors. LCS range in price from a couple of thousand to several thousand pounds (for a relatively sophisticated multi-pollutant and meteorological sensor with communication capabilities). Reference sensors are expensive, both to purchase and maintain, and bulky but are the most accurate units, recommended for use by EU and UK government bodies for AQ monitoring and comply with standards such as MCERTS in the UK. A single unit costs in the region of twenty thousand pounds to monitor a single gas or gaseous species or particle pollutants. IoT, LCS and reference sensors all employ different techniques of air pollutant measurement, which include optical particle counters, light scattering, metal oxide semiconductor sensors, electrochemical sensors, nondispersive infrared sensors, ultraviolet fluorescence, chemiluminescence, infrared photometry and photo-ionisation detection sensors. For more detail see Borrego et al. (2016) and Mead et al. (2013). The LCS used in this project are either Envirowatch E-MOTEs or AQMesh pods. Envirowatch E-MOTEs are deployed either in a local mesh (deployed in a cluster, providing data via ZigBee, within a certain area for high resolution monitoring, no more than 100 m from each other, with a gateway providing uplink capability) or independent (distributed sensors that can be deployed at any distance from each other and can be used for both high and low resolution monitoring, using longer distance communications systems such as GPRS or Wi-Fi providing internet access). AQMesh sensors are independent and can be deployed at both high and low spatial resolution. In this case each sensor independently sends data to a cloud server using GPRS. LCS offer great potential for AQ monitoring in low- and middle-income countries.

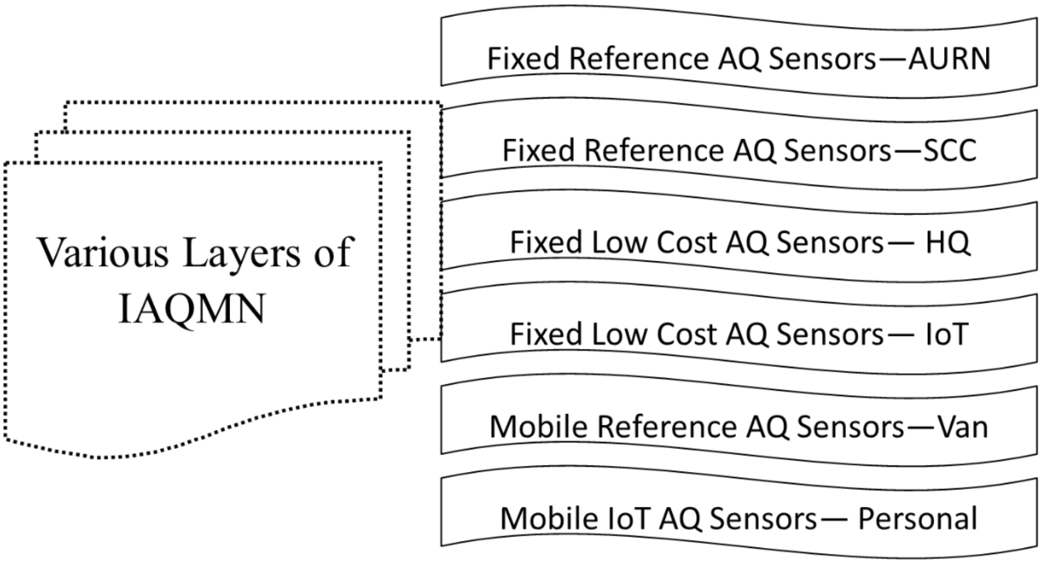

Various Layers of IAQMN
Fixed Reference AQ Sensors—AURN
Fixed Reference AQ Sensors—SCC
Fixed Low Cost AQ Sensors— HQ
Fixed Low Cost AQ Sensors— IoT
Mobile Reference AQ Sensors—Van
Mobile IoT AQ Sensors— Personal

Figure 4.1. Various layers of AQMS. In the diagram AURN stands for automatic urban and rural network, SCC for Sheffield City Council, HQ for High Quality (e.g., AQMesh and Envirowatch E-MOTEs), IoT for Internet of Things, Van – sensors mounted on a vehicle (MOBIUS), and Personal – sensors carried by people.

Table 4.1. Summarising the features of various types of monitoring techniques.

Sensor type	**Temporality (resolution)**	**Spatiality (resolution)**	**Quality**
Reference sensors - AURN	Medium (hourly), long term	Low (fixed)	High
Reference sensors - SCC	Medium (hourly), long term	Low (fixed)	High
LCS	High (minute), long term	High (fixed)	Medium
IoT	High (minute), long term	High (fixed)	Low
Mobile Ref. sensors	High (variable), short term	Variable (determinate)	High
Mobile IoT sensors	High (minute), short term	Variable (indeterminate)	Low

The new generation of sensors such as E-MOTEs and AQMesh pods have the capability to mitigate the effect of climatic factors such as temperature and relative humidity on AQ data collection. The innovation is the addition of a fourth electrode, which is embedded in the sensor electrolyte allowing the reaction from environmental effects to be measured without the effects from the target gas. Furthermore, mathematical algorithms are developed for individual sensor types to compensate for environmental effects and cross-gas interference to provide the best possible precision and accuracy of measurement.

All sensors are pre-calibrated by the manufacturers and, subsequently returned for sensor replacement and recalibration periodically as specified by the manufacturer. In-the-field local calibration of sensors is required in certain circumstances including: (a) following a sensor-pack change the new sensor should be calibrated; and (b) following a large step change in environmental conditions, e.g., a change in average temperature of 10 degrees Celsius or more, relative to when it was originally calibrated. During this project, the sensors will be calibrated locally in two ways: (i) Co-location with reference sensors, and (ii) Using MOBIUS.

Several LCS will be deployed next to reference AQMS including at Devonshire Green and Siemens Close Tinsley. At these sites one E-MOTE and one AQMesh pod will be deployed next to the reference AQMS. The sensors will be placed immediately adjacent or no further than 2 m apart from the reference sensors. This will help correct slopes and offset (intercept) values of the LCS to improve the accuracy of results by comparing data over a period of several months. The manufacturer recommends co-location of sensors with reference sensors for several days or weeks, however, in this project LCS will be co-located for a year with reference sensors and the calibration will be across the seasons (Winter – Nov, Dec, Jan; Spring – Feb, Mar, Apr; Summer – May, June, July; and Autumn – Aug, Sep, Oct), which can help determine the effect of various meteorological parameters on the performance of these sensors. During the calibration LCS measurements are regressed versus reference (Ref) measurements, where LCS data are taken as independent (x-axis) and Ref as dependent (y-axis) variable. Regression model is run and values of slope and intercepts are calculated using the measured LCS and Ref concentrations as shown in equation 4.1 and 4.2.

$$\text{Ref} = \text{intercept} + (\text{slope x LCS}) \quad (4.1)$$

$$\text{NO2\_Ref} = a + (b \text{ x } NO_2\text{\_LCS}) \quad (4.2)$$

The values of slope and intercept are then applied to the whole dataset of LCS. MOBIUS will be used for calibrating LCS around the city in different locations and different seasons. It will be parked for a minimum period of five hours adjacent to the sensors (preferably no more than two metres apart, but as close as practically possible). In situations where the vehicle cannot reach the vicinity of the sensor, it will be removed from its mount and temporarily affixed to MOBIUS for the period of calibration. Installation of the sensors is relatively easy and quick so if necessary this can be accomplished within a matter of minutes. The calibration time period is limited by the auxiliary battery/inverter sets (a total of 2.8 kWh) carried on MOBIUS which power the reference analysers. In some cases, where mains power is available (e.g., sensors deployed at the university campus) the period of colocation can be extended to 24 hours to cover the whole diurnal cycle. Calibration will be carried out at least once in each season (Winter – Nov, Dec, Jan; Spring – Feb, Mar, Apr; Summer – May, June, July; and Autumn – Aug, Sep, Oct). After obtaining the concentrations of the LCS and MOBIUS, the values of intercept and slopes will be calculated and applied as shown in equation 4.1 and 4.2. The main features we want to see in a sensor network are temporal resolution (time), spatial resolution (space) and quality of the data (Figure 4.2). Reference sensors provide high quality data with reasonable time resolution (hourly), however, their spatial resolution is low due to their large size, power requirements and high price. We have only 3 AURN sites and 6 Sheffield City Council (SCC) sites in the whole city of Sheffield, so spatial resolution is low. In contrast, LCS both AQMesh and Envirowatch E-MOTEs, can provide high resolution spatial and temporal data but at relatively lower quality. LCS can provide real-time minute-by-minute data and a high-density network can be set up due to their low price and maintenance cost. Both reference and LCS can provide long-term data (e.g., over a year or longer). On the other hand, mobile networks can utilise both reference sensors and LCS to provide high resolution temporal and spatial data, however they normally provide short term data (e.g., the vehicle can be parked on a specific location for a limited period of time, usually up to five hours) and their spatial resolution is variable. Mobile networks can provide high or low spatial resolution data depending on the need. Furthermore, using MOBIUS we have to monitor roads or streets one by one. This temporal differences (gaps) render the data incomparable with each other due to the fact that AQ levels vary during different hours of the day, days of the week or seasons of the year mainly due to differences in meteorological conditions (e.g., temperature, solar radiation, wind speed and direction) and boundary layer characteristics which affect pollutant dispersion. Therefore, MOBIUS will be mainly used for calibration or for short term monitoring purposes.

As shown in Figure 4.2, in the three-dimensional (3-D) time-space and quality box we want to achieve point (P) ideally, which indicates high quality data with high spatiotemporal resolution. However, this is not always practical due to various reasons, mainly financial budget. Therefore, we need to compromise either on quality, spatial or temporal resolution. The dimension that will be subject to compromise is dependent on the reason for monitoring and the intended purpose of the output data. For example, if we want to determine a long term temporal trend over a ten-year period, there is no need for high temporal resolution (e.g., minute-by-minute or hourly data), daily or even monthly data will suffice. Also, we might not need a dense network of AQ sensors, a small number deployed in urban background, suburban background or rural locations will suffice. In contrast, if the purpose is to investigate how road traffic-flow affects AQ, we will need high temporal resolution, e.g., minute-by-minute data, because in this instance anything with less frequency will be too coarse. Furthermore, in monitoring the effect of traffic on AQ, if the purpose is to see the pattern in air pollution levels, we might not need very accurate readings, so the quality (accuracy) of the readings might not be of a great concern. On the other hand, for urban-scale modelling and mapping, high-resolution spatial data collected by a dense AQMN will be required. Therefore, it can be said that the type of sensors, type of monitoring sites, quality of data, density of monitoring network and temporal resolution are dependent on the purpose of monitoring programme.

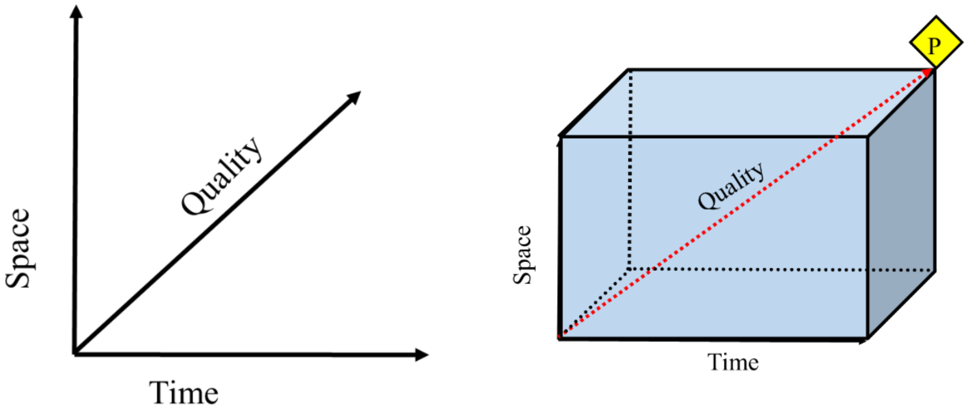


Figure 4.2. Three-dimensional (3-D) box, where x-axis is represented by time (temporal resolution of collected data), y-axis by space (spatial resolution of the collected data) and z-axis by quality (quality of the collected data). Ideally, we want to achieve high spatial and temporal resolution with high quality data (represented by point ‘P’), however this may not be always possible.

4.2.1. Population-weighted pollution concentration (PWPC)

Human exposure to air pollution is a function of population density (residents/km^2) and pollution levels. Therefore, both social (population) and environmental (air pollution) indicators should be considered in structuring an AQMN. Population data are normally more readily available than pollution data, e.g., population data can be obtained from a recent census or local council. In contrast, detailed pollution data of various air pollutants are generally not available, especially in countries with less well developed infrastructure. Therefore, in the absence of air pollution data, as an alternative air pollution emissions or modelling estimations of air pollutants can be used. For example, Righini et al. (2014) have used air pollutant emissions data to optimise AQMN in Italy. It is worth mentioning that air pollutant emissions and concentrations are not the same and the same amount of emissions may result in different concentrations due to differences in meteorological conditions and atmospheric boundary layer height. However, generally emissions of primary pollutants are accepted as a reasonable estimates of pollutant concentrations (Righini et al., 2014). In this study we used population density maps of 2016 to show population density of Sheffield (Figure 4.3). In these maps bright green (6000 – 8000 residents/km^2) represents average population (7000 ±1000), as the average population of Sheffield is about 7000 residents/km^2. Orange and red show areas where population is greater than average, in some case more than double and treble. Areas of high population density are mostly shown in the city centre where people live in multi-storey buildings. Locations of primary and secondary schools were used to represent areas of more vulnerable people (children are more vulnerable and are more likely to be adversely affected if exposed to high levels of air pollution). Schools represent urban and suburban background environmental types. NO_2 diffusion tube locations and annual concentrations (µg/m^3) are shown in Figure 4.4. NO_2 diffusion tubes data are available in Sheffield for the last several years providing reasonable spatial coverage. Therefore, these maps were used to determine spatial variability of NO_2 in the City. In Figure 4.4, orange and red dots show locations where NO_2 levels exceeded annual AQ limits (40 µg/m^3) (Air quality objectives, 2015).

NO_2 concentration is shown in the form of points, whereas population density is shown in the form of polygons, therefore, firstly NO_2 concentration was converted into the same format. In case there were more than 1 point in a polygon, their average NO_2 concentrations were calculated for the polygon. Output areas (polygons) are the lowest geographical levels that are created for Census data. Output areas are built from clusters of adjacent unit postcodes. They are designed to be similar in terms of cultural and demographic characteristics with relatively similar populations for statistical purposes. Output areas do not mix urban and rural areas and should be consisted either entirely of urban postcodes or rural postcodes. An output area should have minimum size of 40 resident households and 100 resident people, however its recommended size is 125 households (Office for National Statistics, 2018).

Averaging NO_2 concentrations across the polygon may introduce a degree of error, especially if the polygon is large and heterogeneous in terms of air pollutant concentrations. However, heterogeneity of air pollutants is minimised by the fact that an output area doesn't mix urban and rural areas. Any polygon without data was excluded from the analysis, which means the polygon was not coloured. Unshaded polygons mean there were no data of NO_2.

To calculate population-weighted pollution concentrations (PWPC), firstly normalised population density (NPD) of each cell was obtained by dividing population density (PD) of each cell by average PD (mean-PD) of all polygons, following the approach used by Carslaw (2015) in the 'openair-manual' (equation 4.3). In the second step NPD was multiplied by pollution concentration (PC) of each cell (equation 4.4) to get PWPC, which is an important indication of people exposure to air pollution. PWPC was mapped using ArcGIS version10.4.1 as shown in Figure 4.5.

$$NPD_i = (PD_i/\text{mean-PD}) \qquad (4.3)$$

$$PWPC_i = NPD_i * PC_i \qquad (4.4)$$

In Figure 4.5, red shows the highest PWPC, whereas blue indicates the lowest PWPC. The number of sensors to be deployed will depend on the budget of the project (economic criteria). The areas with higher PWPC should get priority in deploying AQ sensors. However, it is important to quantify spatial variability of NO_2 concentrations to determine how many sensors will be deployed in each polygon. More sensors should be deployed in the area with greater spatial variability and vice versa, which is discussed in the next section (2.2).

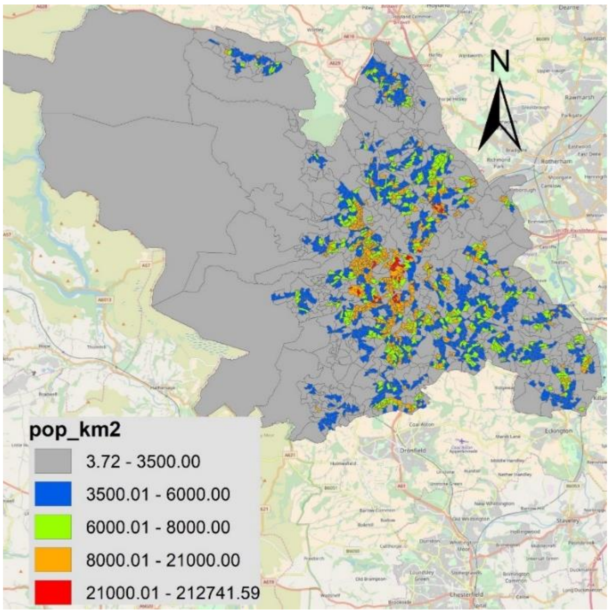


Figure 4.3. Population density (residents/km^2) map of Sheffield, 2016.

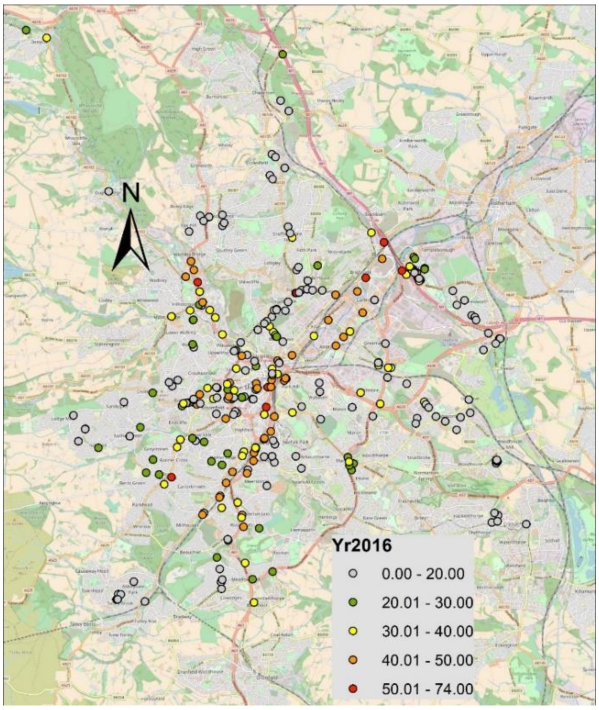


Figure 4.4. Locations and annual average NO_2 concentrations (µg/m^3) of NO_2 diffusion tubes in Sheffield, 2016.

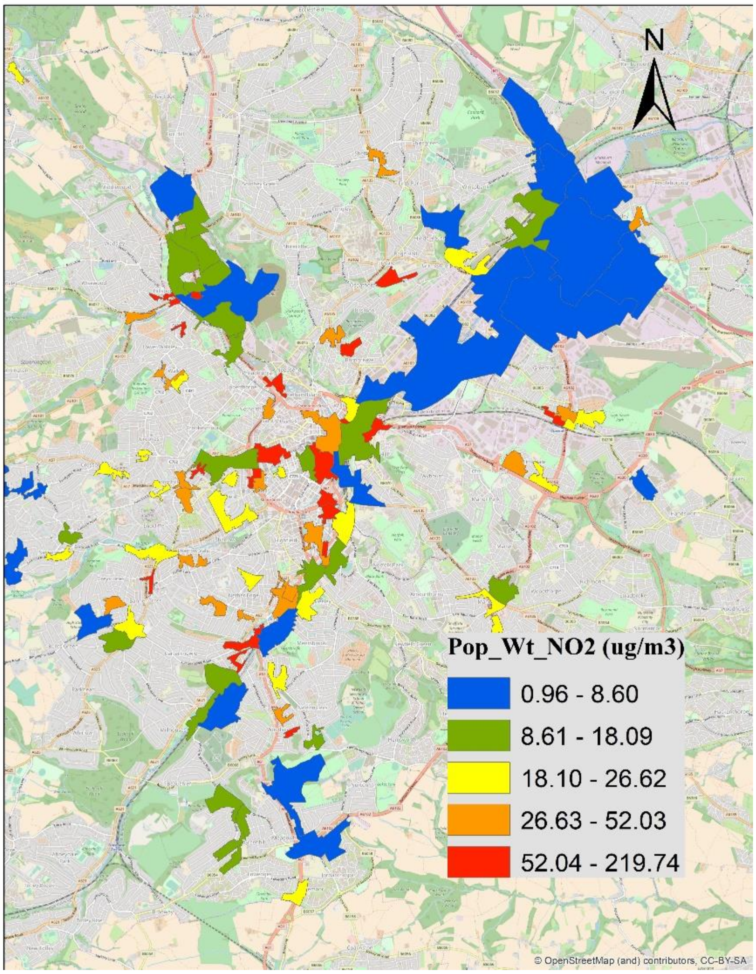


Figure 4.5. Population-weighted NO_2 concentrations (PWPC) ($\mu g/m^3$) Sheffield, 2016.

4.2.2. Spatial variability of NO_2 concentrations ($\mu g/m^3$)

In the previous section PWPC (Figure 4.5) was analysed, which highlights those areas where more people are exposed to NO_2 concentrations. However, to decide where more sensors should be deployed we need to determine spatial variability of NO_2. Areas experiencing high concentrations of PWPC and greater spatial variability require more sensors in contrast to those areas where PWPC is low and are spatially homogenous in terms of pollution concentrations. Therefore, in this section first we quantify spatial variability (SV) of NO_2 concentrations.

SV of NO_2 concentrations ($\mu g/m^3$) are shown in Figure 4.6, which are calculated and mapped in ArcGIS 10.4.1. Standard deviation (STD) determines how NO_2 concentration is dispersed within a given area. To determine SV, we calculated STD of NO_2 concentrations within each 400 m^2 area, having at least 3 observations using equation 4.5.

$$STDi = \sqrt{1/n \sum_{i=1}^{n}(xi-\mu)^2} \sqrt{1/\mathrm{n} \sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{xi}-\mathrm{\mu})^2} \quad (4.5)$$

In equation 4.5, x is NO_2 concentrations, μ is the mean concentrations and n is the number of data points. STDi are mapped in Figure 4.6, where red shows more spatial variability of NO_2 concentrations. To provide a quantitative assessment as to how many sensors should be deployed in each area, in this study we propose an approach, which integrates PWPC (Figure

4.5) with SV (Figure 4.6), thus accounting for population density, pollution concentrations and SV. The resultant variable is termed weighted spatial variability (WSV), which is the product of PWPC and normalised standard deviation (NSTD) as shown in equation 4.6.

$$WSVi = PWPCi * NSTDi \quad (4.6)$$

Finally, the number of sensors in each cell (n_i) is determined by solving equation 4.7 using the WSVi value within each cell and sum of the WSV_i of all cells.

$$n_i = (WSV_i / \sum(WSV_i)) * N_t \quad (4.7)$$

Where N_t is the total number of sensors to be deployed, which is decided on the basis of economic criteria (budget).

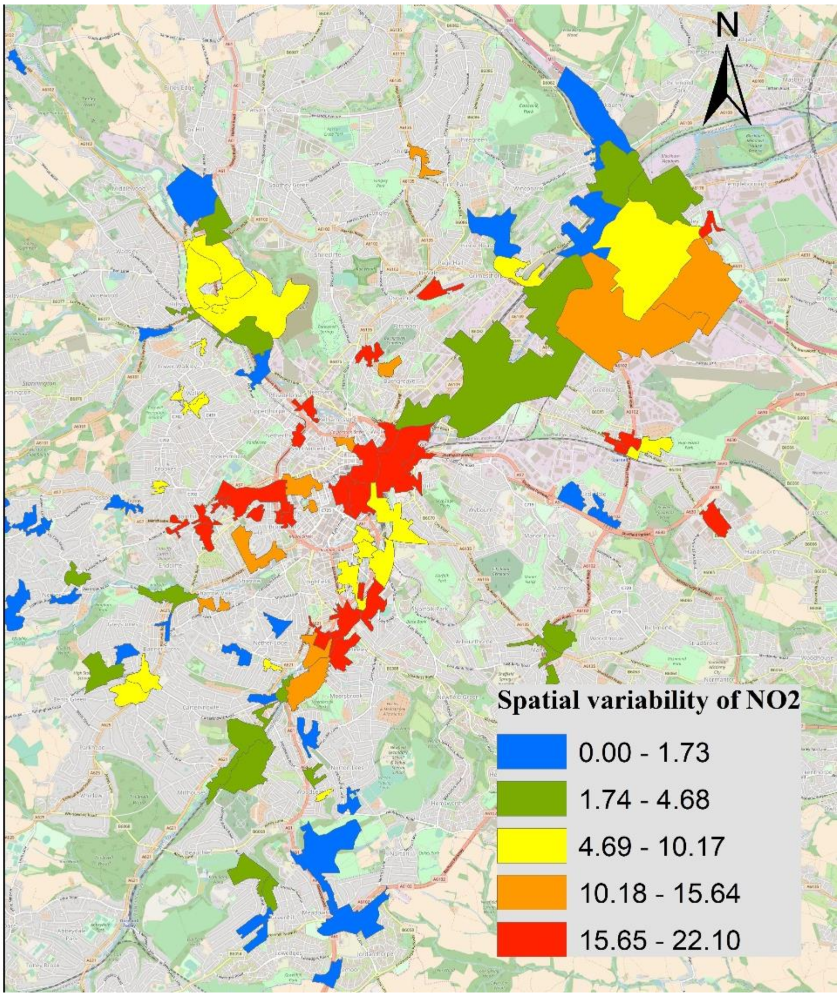


Figure 4.6. Showing spatial variability of NO_2 concentrations (µg/m_3)

The main points considered for site allocation are summarised in Figure 4.7, including population density, pollution levels and pollution spatial variability. Furthermore, the purpose of the AQ monitoring programme and the opinion of experts having experience of the local area are two important factors in siting the AQMS. They inform where exactly the sensors will be deployed in each polygon. Therefore, it is important to analyse the purpose of the AQ monitoring, discussed in next section (2.3).

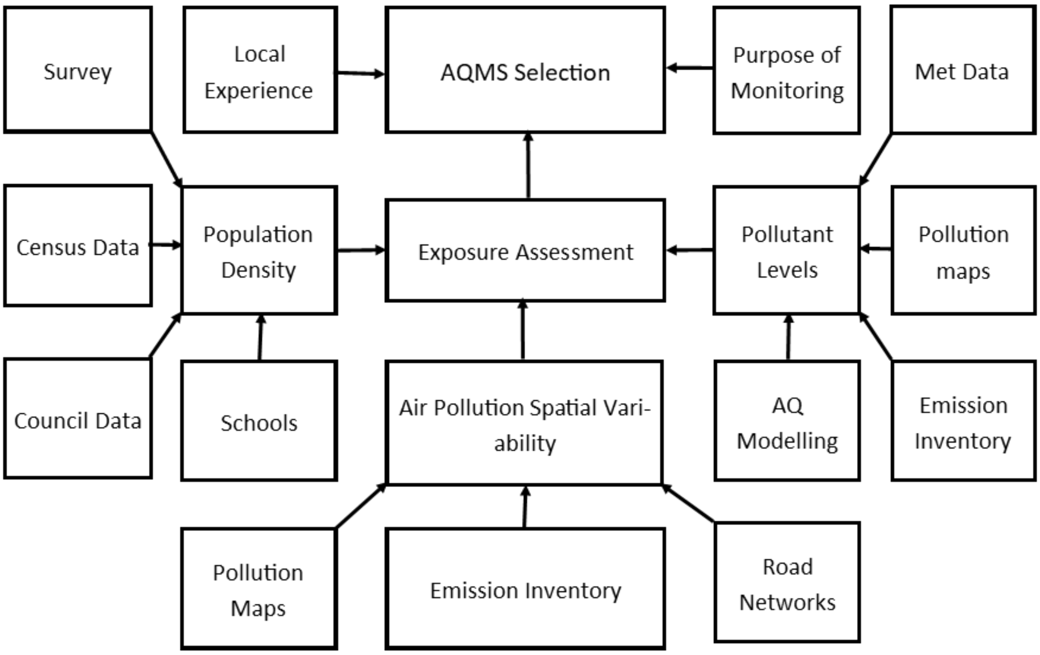


Figure 4.7. Criteria for the selection of AQMS (ESCAPE,2010; LAQM.TG, 2009; Kanaroglou et al., 2005). Also, see Figure 4.8 for the description of 'the purpose of monitoring'.

4.2.3. Purpose of air quality monitoring and its role in site selection

The primary criterion among those that are most important for selecting the locations of AQMS is the purpose of the AQ monitoring programme. Therefore, it is important to briefly describe the main purposes of AQ monitoring and what role they play in determining sites for AQMS deployment. AQ monitoring may be carried out due to the following reasons:

(i) AQ review and assessment (regulatory compliance) (Kanaroglou et al., 2005; ESCAPE, 2010; LAQM.TG16, 2016; LAQM.TG09, 2009): AQ review and assessment involves monitoring current levels of air pollution and modelling how it might change in the near future. The main aim of the review and assessment is to ensure that national AQ objectives are achieved. The purpose of these objectives is to protect human health and environment from the negative impacts of air pollution. Probably the most important reason for AQ monitoring is to assess human exposure to air pollution. This determines the areas where people are exposed to high levels of air pollution. The monitoring programme should take into account air pollution and demographical characteristics of the region under consideration and consider worst-case public exposure both in terms of pollution levels and population density.

(ii) AQ modelling (Raffuse et al., 2007; LAQM.TG16, 2016): AQ monitoring is also carried out to assess the outcome of dispersion modelling studies. In these types of monitoring programmes sensors should be deployed close to the emission sources. For example, if the purpose is to assess the performance of a dispersion model developed for a particular road, the AQ sensors should be deployed at the roadside of that particular road, even if there is no exposure. In additional to dispersion modelling, AQ data collected by the monitoring programmes are used in various AQ statistical, photochemical, mathematical, forecasting and land-use regression models. These are employed in various investigations into AQ related projects and multi-disciplinary projects like transportation models, climate models and other urban system models. In this case AQ sensors should be deployed where the other variables (e.g., weather parameters) are also monitored.

(iii) Temporal trends (Raffuse et al., 2007): Sometimes AQ is monitored to determine how air pollutant levels have changed over a specific period of time, over the last ten years, for instance. In this case monitoring should be carried out at a background site, away from local sources, e.g., an urban background or suburban background site. However, if the purpose is to assess the temporal trend near a particular emission source, then sensors should be deployed as close as possible to that source. Furthermore, monitoring data can be used to determine diurnal, weekly and annual cycles of air pollutants, however for this purpose high resolution temporal concentration measurements are required, such as at intervals of one minute, 15 minutes or hourly.

(iv) Source apportionment (Raffuse et al., 2007): AQ monitoring programmes are also launched to determine various sources of air pollutant emissions. In this case AQ sensors should be deployed in different types of environments including roadside (where both heavy and intermediate traffic loads are encountered), next to point sources, urban background, suburban background and rural sites. Sensors next to local sources determine the contribution of local sources, whereas background and rural sites help determine the contribution of urban level and regional level emission sources.

(v) Spatial coverage (Raffuse et al., 2007): AQ monitoring can help determine spatial trend in air pollution levels. Urban areas demonstrate high spatial variability in air pollution levels due to changes in emission sources and tall buildings which affect air pollutant dispersion processes. Therefore, urban areas in comparison to suburban or rural areas would require more sensors to capture variability in air pollution levels. Sensors should be deployed in different environmental conditions, including next to busy roads, point sources, open streets, street canyons, market places and residential areas. For air pollution mapping at an urban level, a dense network of sensors is required. The density of sensor should be high where air pollution levels are more variable and vice versa (Kanaroglou et al., 2005).

(vi) Identifying the main drivers of AQ: AQ monitoring is sometimes carried out to investigate the effects of various factors on AQ conditions such as various land-use strategies, climate and meteorology, boundary layer height, and topographical and geographical characteristics. If this is the case, then AQ sensors should be deployed in various environment types such as urban background, suburban background, and traffic sites including various altitudes and land-use types.

(vii) Dose-response relationship (Munn, 1981): If an air pollution monitoring programme is used to collect air pollutant data which will be used to establish a dose-response relationship for investigating the effects of air pollution on human health, vegetation, soiling and corrosion of different materials, and economic effect, then AQ sensors should be deployed next to the location where the investigation is taking place.

(viii) Assessing AQ control strategies: Air pollution monitoring is needed for assessing the effectiveness of control strategies, e.g., if an air pollution management and control strategy is implemented in a specific area, then AQ data are required to cover the period just before and after the implementation of the strategy to assess how effective the strategy has been.

These are the main purposes of AQ monitoring, however by no means this list is exhaustive. The purposes of AQ monitoring are summarised in Figure 4.8.

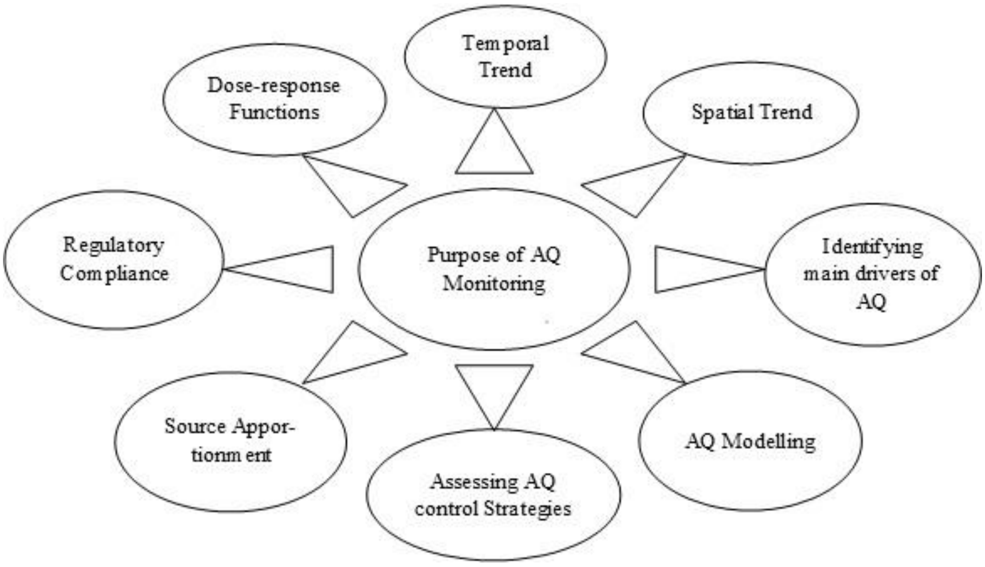


Figure 4.8. The main purposes of AQ monitoring (ESCAPE, 2010; LAQM.TG09, 2009; Raffuse et al., 2007).

The opinion of experts having experience of the local area is a valuable asset in identifying suitable sites for the deployment of AQ sensors. To utilize this, several meetings were arranged with the AQ group and transport team at SCC and researchers from various departments of the University of Sheffield. They had extensive experience of the city and helped identify areas where air pollution has been or is likely to be a problem, where emissions have declined or increased, or where air pollutants emissions are going to change in the near future, e.g. Abbeydale and London Rd, Meadowhall, City Centre and so on. To fully utilise this resource an Air Quality Sensors Network (AQSN) workshop was organised, which was attended by air pollution and environmental science experts from different departments of the University of Sheffield and SCC. Their suggestions were sought on sensor deployment and utilised wherever possible and applicable.

4.3. Results and discussion

In this study following the WSV model and purpose of monitoring, AQ sensors will be deployed in a variety of locations including urban background, suburban background and roadside sites. Some AQMS will be located in the main city centre and others in the suburbs of the city to represent different types of environments. Roadside sites will include both highly and intermediately trafficked roads. Also, both open streets and street canyons will be monitored to analyse the effect of tall buildings on air pollutants dispersion. Urban and suburban monitoring sites will monitor the urban level emission, whereas roadside sites will monitor more local emissions from the traffic. Several sensors will be deployed next to existing reference sites including both AURN and SCC sites for calibration purposes.

Here firstly the existing AQMN is described (section 4.3.1), followed by the proposed AQMN (section 4.3.2).

4.3.1. Existing air quality monitoring network

4.3.1.1. Reference sensors (static)

Reference sensors are the most accurate type and are recommended by the EU and UK DEFRA for monitoring AQ. However, reference instruments are expensive to purchase and maintain and require skilled staff for deployment and calibration. Due to their high purchase and maintenance costs DEFRA and SCC have a sparse network of these sensors in Sheffield. These networks are mainly set up for the purpose of regulatory compliance (in the UK commonly known as review and assessment), mostly providing hourly concentration of various air pollutants over a long period of time (Table 4.1). There are three AURN sites in Sheffield run by DEFRA and six under SCC control. These continuous AQMS are shown in Table 4.2 giving their names, site types, pollutants measured and other details, whereas their locations are shown in Figure 4.9.

Table 4.2. Automatic air quality monitoring stations (AQMS) in Sheffield (SCC, 2016).

Site name	Site type	Easting (X)	Northing (Y)	Pollutant monitored	Monitoring Technique	Distance to road (m)	AURN/SCC
Firvale School (GH1)	Urban BG	436990	390218	NO_2, PM_{10}	CL, TEOM	10	**SCC**
Tinsley Infant School (GH2)	Urban Industrial	440077	390794	NO_2, PM_{10}, , $PM_{2.5}$	CL, TEOM	90 (M1)	**SCC**
Lowfield School (GH3)	Roadside	435181	385366	NO_2, PM_{10}, SO_2	CL, TEOM, UV Fluores.	10	**SCC**
Wicker (GH4)	Urban BG	435959	388021	NO_2, PM_{10}, O_3	CL, TEOM, UV abs.	50	**SCC**
King Ecgbert School (GH5)	Urban BG	430977	380760	NO_2, PM_{10}, O_3	CL, TEOM, UV abs.	100	**SCC**
Waingate (RM1)	Roadside	435750	387647	NO_2, PM_{10}	CL, TEOM	3	**SCC**
Tinsley (SHE)	Urban Industrial	440215	390598	NO_2	CL	120 (M1)	AURN
Devonshire Green (SHDG)	Urban Centre	435158	386885	NO_2, O_3, PM_{10}, $PM_{2.5}$	CL, TEOM, UV abs.	20	AURN
Barnsley Road (SHBR)	Urban Traffic	436276	389930	NO, NO_2, NOx	CL	3	AURN

Table abbreviations are as follows: CL - Chemiluminescence, BG - Background, SCC - Sheffield City Council and AURN - Automatic Urban and Rural Network.

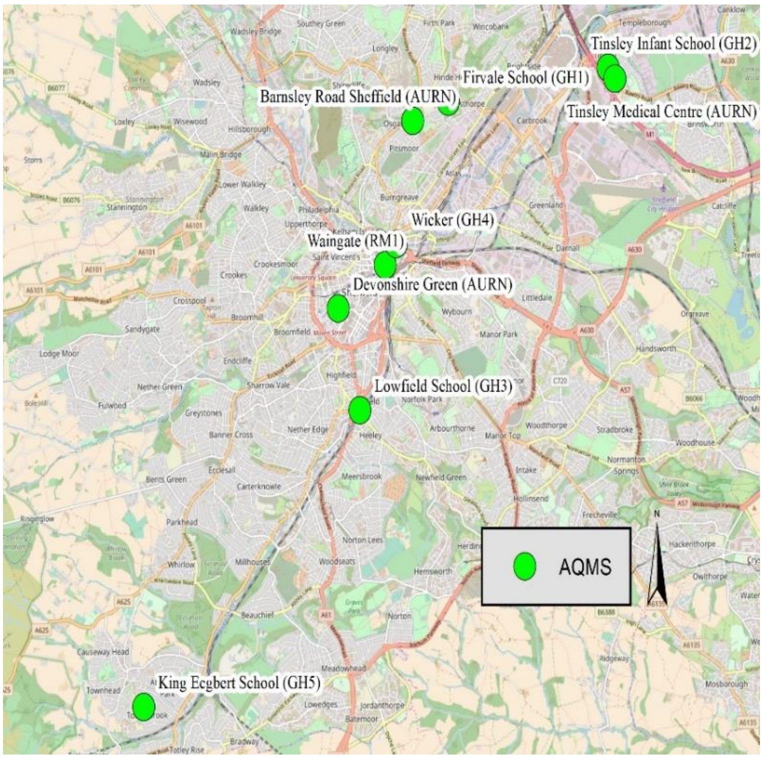

Tinsley Infant School (GH2)
Firvale School (GH1)
Barnsley Road Sheffield (AURN)
Tinsley Medical Centre (AURN)
Wicker (GH4)
Waingate (RM1)
Devonshire Green (AURN)
Lowfield School (GH3)
AQMS
N
King Ecgbert School (GH5)

Figure 4.9. Continuous Air Quality Monitoring Stations (AQMS) in Sheffield comprising of 3 AURN (DEFRA) and 6 SCC sites.

4.3.1.2. Low-cost sensors

An AQMN of LCS in Sheffield is shown in Figure 4.10. In this network the city is divided into three parts: (I) The University of Sheffield Campus; (II) Sheffield City Centre; and (III) Other parts of the city, which include Meadowhall Shopping Centre, Brightside & Attercliffe, Abbeydale and London Roads, southwest of the city, north and northwest of the city, e.g., Penistone Road - A61 and Barnsley Road - A6135, and east and southeast of Sheffield, e.g., Sheffield Parkway.

The University of Sheffield Campus has two meshes (clusters) of sensors, each made of nine (9) E-MOTE sensors. One is deployed along Broad Lane, Portobello Lane and between them (Figure 4.10, middle-panel) and second mesh along A57 (Brook Hill and Western Bank) near Sheffield Children Hospital, Royal Hallamshire Hospital and Sheffield Student Union (Figure4.10, middle-panel). The intention is to capture more micro-level factors causing changes in air pollution concentrations. In addition to multi-storey student accommodation, University students are present in this area much of the time. Further details of the sensor locations are provided in Table 4.3. The western side of the A57 also has many pedestrians due to university students and patients attending the hospitals (who may be sensitive receptors). Several major and minor roads further contribute to the number of emission sources.

Ten (10) Envirowatch E-MOTE sensors are deployed around the city centre (Figure 4.10, lower-panel), covering the busiest area around the train and bus stations, Arundel Gate, Pond Street, Sheaf Street, Sheffield Hallam University and Town Hall. This area experiences high levels of air pollution and WSV due to several busy roads and transport hubs and is surrounded by tall buildings creating dispersion barriers. This area is extensively covered to obtain high spatial resolution readings in order to identify the main drivers of air pollution. The sensor on Arundel Gate between Genting Casino and Hallam Business School is especially deployed to study the street canyon effect. One sensor next to Devonshire Green AQMS, which is part of the UK AURN, is deployed for calibration purposes.

Independent LCS are deployed in the rest of the city in various areas including Meadowhall (2 independent sensors), Brightside & Attercliffe (2 independent E-MOTEs), Abbeydale and London Roads (2 independent AQMesh pods), Southwest of City (2 independent AQ Mesh pods), North and northwest of Sheffield, e.g., Penistone Road (A61) and Barnsley Road (A6135) (3 Independent AQMesh pods), and East and southeast of Sheffield, e.g., Sheffield Parkway (2 independent sensors) (Figure 4.10).

Meadowhall is a very busy shopping centre with large parking areas which remain busy throughout the day and evening. Furthermore, Meadowhall is about to undergo major development in the coming years, therefore AQ monitoring in this area can determine how air pollutant levels will alter once the plans are completed. The Tinsley area is adjacent to a large road network interchange, where several busy roads intersect and access the Motorway (M1), A631 and A6178. In Tinsley a two continuous AQMS are also deployed. Sensors are deployed on Siemens Close, Meadowhall Way and the Meadowhall Interchange. Brightside and Attercliffe areas are also very busy in terms of traffic flow carrying most of the traffic from Sheffield City Centre, bus station, train station and universities to the M1 and Meadowhall Shopping Centre. These roads (A6178 and A6109) are therefore heavily trafficked and highly polluted. For these reasons, sensors are deployed on Savile Street (adjacent to Tesco Extra) and Brightside Lane. Sensors are also deployed in the Abbeydale and London road area, with high population density and pollution levels along these busy main roads. Proposals are in hand to retrofit buses to reduce emissions of NO_2 and PM that should be observed by monitoring these corridors. Furthermore, southwest and north-and-northwest of the city are highly populated and polluted areas where several hospitals and schools are located give rise to possibility of exposure of people who may be more vulnerable to high levels of air pollutants. In the North of City sensors are deployed at the Northern General Hospital, Hills Borough Primary School, and St Marys CoE Primary School. In the west sensors are deployed at Endcliffe Crescent and Cowlishaw Road, and in the east two AQ sensors are deployed at Prince of Wales Road and Maltravers Road.

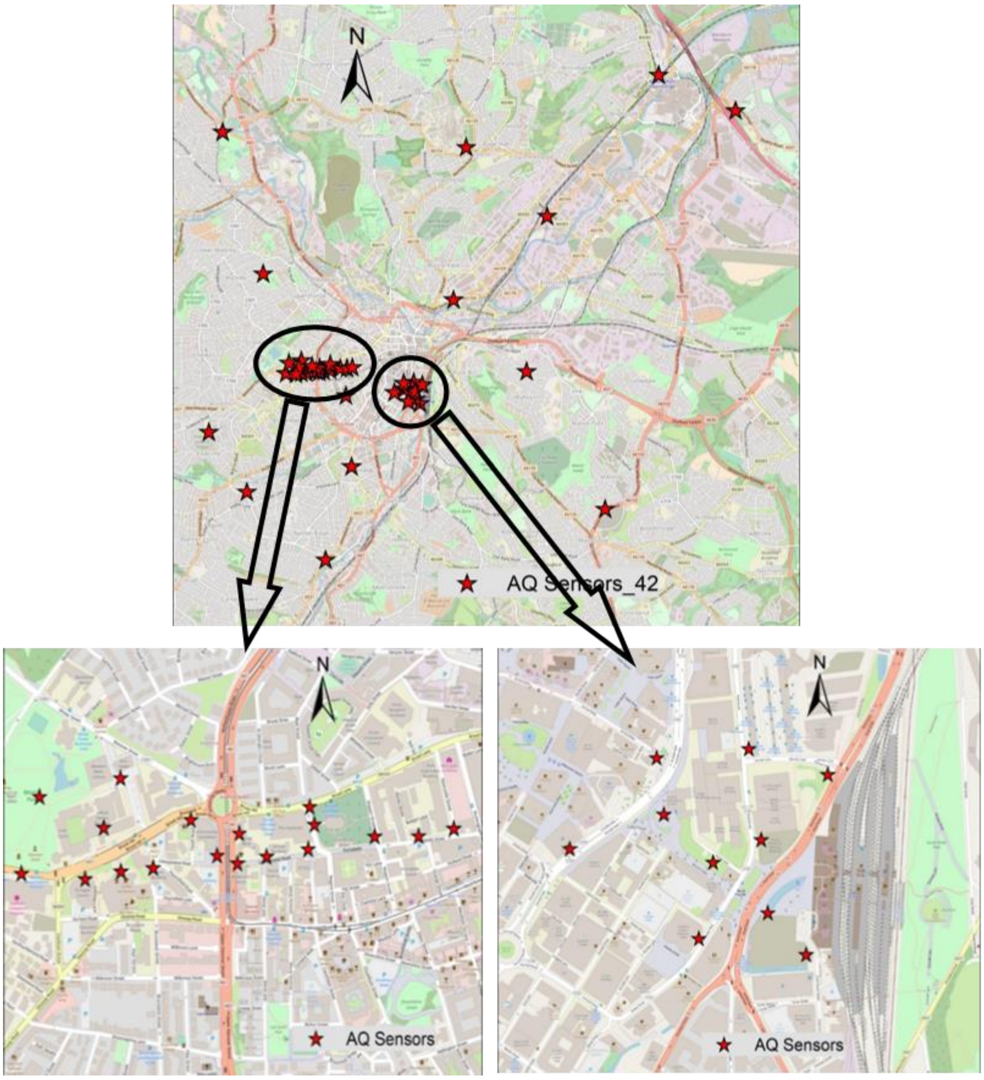


Figure 4.10. EnviroWatch E-MOTE & AQMesh pod sensors (total 42 sensors) deployed around Sheffield (upper-panel). University of Sheffield Campus (lower-left panel) and City Centre (lower-right panel) are magnified to show their details.

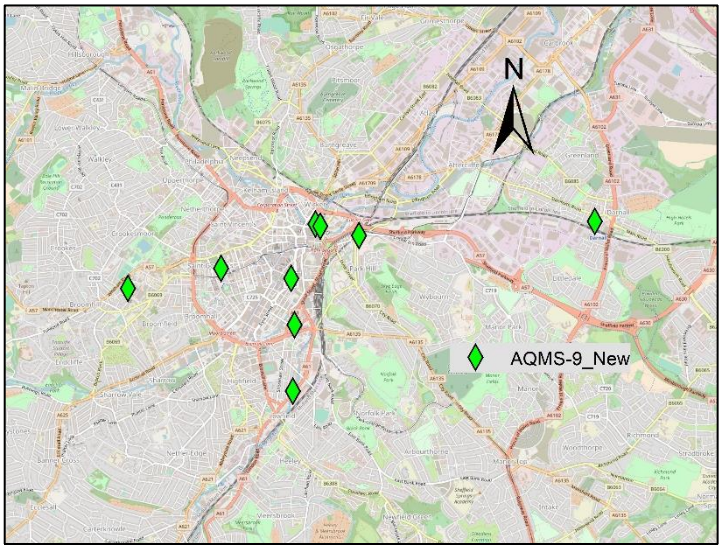
4.3.2. Proposed air quality monitoring network

According to the methodology described in section 4.2.2, the whole city was mapped based on the value of WSV which incorporates PD, pollution levels and SV of air pollution. The number of sensors in each polygon was proportional to the value of WSV, which means more sensors should be deployed in polygons which show greater WSV value using equation 4.7. Firstly, an AQMN (Figure 4.11) is proposed for nine reference instruments deployed by both SCC and DEFRA in Sheffield City (discussed in 4.3.1.1, Figure 4.9). Figure 4.11 (upper-panel) shows the proposed locations of nine AQMS. Comparing this to Figure 4.9, we can clearly see that the proposed network allocates more sensors around the city centre, where pollution levels are higher and more people are exposed to air pollution. Also, near the city centre air pollution levels demonstrate much greater SV probably due to tall buildings and numerous emission sources including many major and minor roads. Six sensors are sited in the city centre including two sensors along Blonk Street and one each along Sheffield Parkway, Arundel Gate, Shoreham Street leading to Sheaf Street and the University of Sheffield along Portobello. Three sensors are allocated outside the city, one each along Whitham Road, Queens Road and Staniforth Road (Figure 4.11, upper-panel).

Figure 4.11 (lower-panel) shows the locations of forty-two (42) LCS. The proposed network provides better spatial coverage, focusing more on the city centre and surrounding area where WSV values were higher. Seven sensors are allocated on the southern side of the city along Queens Road and Chesterfield Road. Two sensors to the east near the junction of Staniforth Road and Prince of Wales Road, and one near Tinsley roundabout. Four sensors are allocated in the north of city along Barnsley Road and five along Penistone Road and Walkley Road. Nine sensors are sited on the western side along Whitham Road and Ecclesall Road. Fifteen sensors are sited in the city centre: St Mary's Road (2), Arundel Gate (2), Sheffield Parkway (2), Blonk Street (4), Haymarket (1), Corporation Street (1), Mapping Street (1), Headford Gardens (1) and Gell Street (1).

The total number of sensors are decided on the basis of financial resources (budget of the project), whereas the number of sensors in each polygon are based on exposure (population and pollution levels) and its SV. Sometimes AQ sensors are deployed for a particular reason, to determine the background level of ground level O_3, for instance. In this case the AQMS should be deployed in a rural background area. Therefore, based on the purpose of monitoring, AQ monitoring authority may deploy a sensor in a specific place against the recommendation of this network. Furthermore, once the general locations and number of sensors are decided using the methodology suggested in this manuscript, the final location for each sensor within the polygon will be decided based on the purpose of the monitoring using the opinion of experts having experience of the local areas.

The final map of LCS and reference sensors is shown in Figure 4.12, where the lower-panel shows the proposed locations according to WSV and the upper-panel shows the existing locations of AQMS in Sheffield. These two maps have been put together to facilitate their comparison.



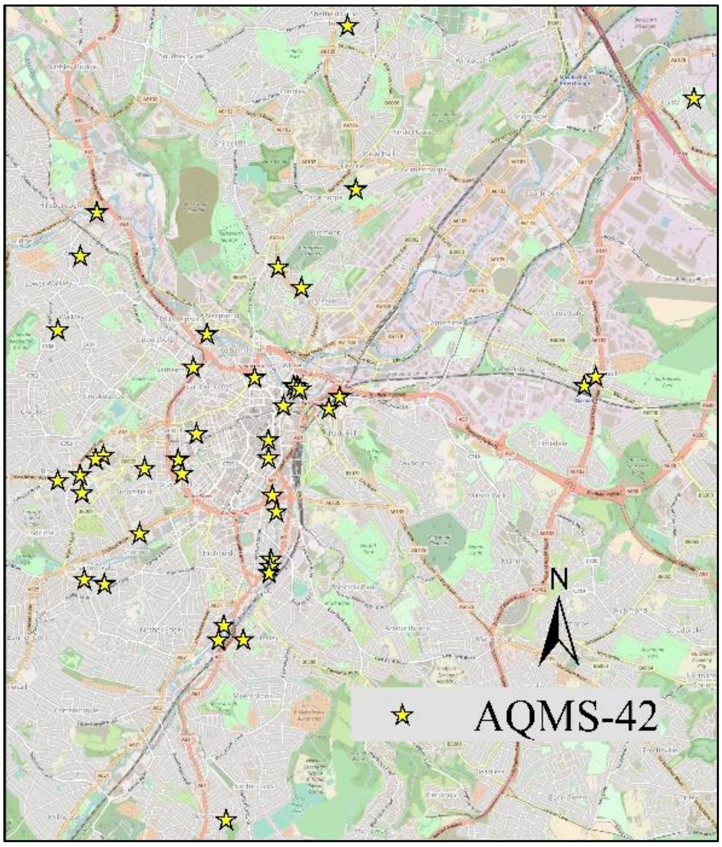
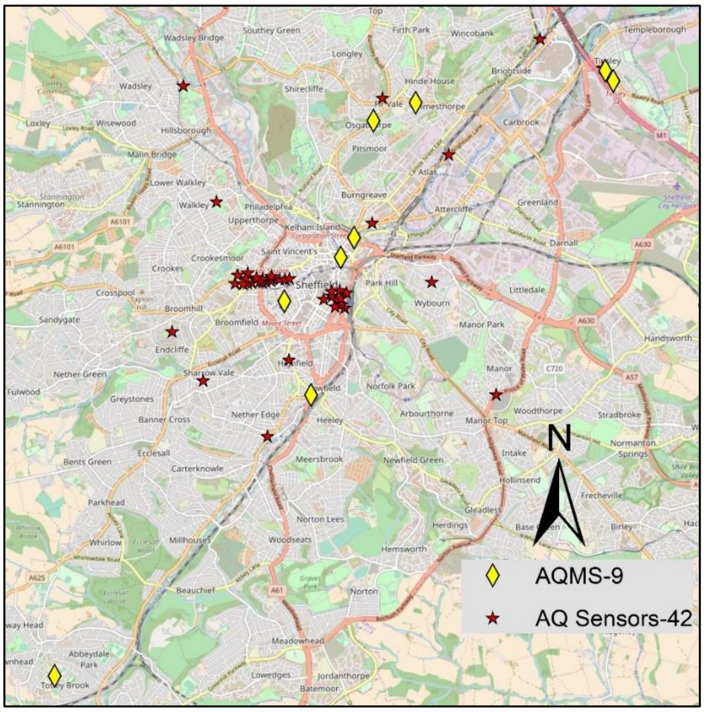


Figure 4.11. Proposed AQ sensor locations based on WSV in Sheffield, where upper-panel shows locations of reference sensors and lower-panel shows locations of LCS.



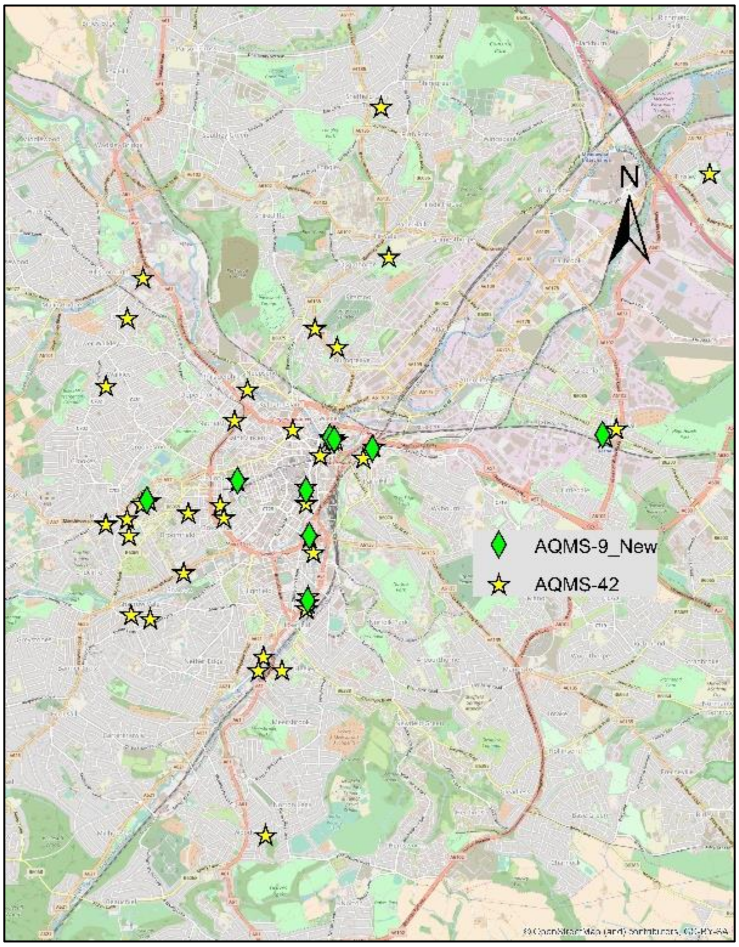


Figure 4.12. AQMN with forty-two (42) LCS and nine (9) continuous AQMS, where the lower-panel shows the proposed network whereas the upper-panel shows the present locations in Sheffield.

In addition to reference sensors and LCS (AQMesh and EnviroWatch E-MOTEs), IoT sensors will be deployed in different parts of the city, however, how many is not yet known. IoT sensors will be deployed mostly next to AQMesh, E-MOTEs and reference sensors so that their performance can be compared. In addition, IoT sensors will be deployed to cover gaps between high quality sensors. MOBIUS will be used to take readings between static sensors. The data points are tagged with location and time utilising an on-board GPS. MOBIUS will be used for AQ monitoring in places where the deployment of fixed sensors is not possible due to limiting factors, e.g., lack of power supply, insufficient space or unsafe location for fixed monitoring stations. MOBIUS is helpful to collect data between various fixed monitoring stations, to highlight hot spots and provide much better spatial coverage, however they are mainly suitable for short term monitoring and sensors calibration. In this way using different types of sensors, an IAQMN will provide high spatiotemporal resolution maps in Sheffield.

The proposed network is more spatially representative than the previous network because it is structured based on WSV. Air pollutant concentrations measured by the proposed network will capture spatial variability of air pollution in locations not captured before and will highlight hotspots of air pollution where more people are exposed to high levels of air pollutants. This proposed network provides a great example for local authorities and DEFRA for structuring an IAQMN. Future work includes collection of AQ data from the proposed network, integrating the data collected by various sensors, and developing a land-use regression model.

4.4. Conclusions

With new developments in AQ monitoring technology including the availability and popularity of LCS, more and more people are setting up purpose-designed AQMN for various purposes, e.g., to produce high resolution AQ maps, to assess human exposure to air pollution, and to review regulatory compliance of air pollutant levels. In this study we proposed an IAQMN based on population density, pollution concentrations and WSV. This is further supported by a conceptual model which analyses the purpose of AQ monitoring and discusses how the purpose is linked with sensors deployments. This study proposes an IAQMN utilising several layers of AQ monitoring approaches including both fixed and mobile techniques employing reference, LCS, and IoT sensors. The aim is to achieve AQ data of high spatiotemporal resolution, however, in many cases there is a compromise made on one or more dimensions due to various constraints, e.g., financial constraints. As a case study, an IAQMN has been proposed in Sheffield, which will monitor AQ levels on roadsides, urban background, suburban background, hospitals, schools and universities. Forty-two (42) LCS along with nine (9) reference sensors will be deployed. The network will be further supported by mobile monitoring using both reference and LCS. The data obtained from various layers of the network will be fused together to be used for developing spatiotemporal high resolution maps and LUR model in Sheffield. This study provides a practical example as to how various types of AQ monitoring sensors can be integrated into one monitoring network in large urban areas to capture local level spatiotemporal variability in air pollution concentrations, especially in low- and middle-income countries where AQMNs either do not exist or are sparse. LCS will be of particular importance in those countries for setting up purpose-designed AQMNs.

Acknowledgment

We would like to express our thanks to the EPSRC and Siemens plc for funding this project. We would also like to thank Sheffield City Council particularly Nicola McHugh and Ogo Osammor for providing additional data including population maps, school locations and NO_2 diffusion-tube data.

Glossary

AQ	Air quality
AQMA	Air quality management area(s)
AQMN	Air quality monitoring network(s)
AQMS	Air quality monitoring station(s)
AURN	Automatic Urban and Rural Network
DEFRA	Department for the Environment, Food and Rural Affairs (UK)
GIS	Geographic information system
GPRS	Global packet radio service
IAQMN	Integrated air quality monitoring network
LUR	Land use regression
NPD	Normalised population density
PD	Population density
PWPC	Population-weighted pollution concentration
SCC	Sheffield City Council
SV	Spatial variability

VOC	Volatile organic compound
WSV	Weighted spatial variability

4.5. References

Air quality objectives, 2015. National air quality objectives and European Directive limit and target values for the protection of human health. Available online: https://uk-air.defra.gov.uk/assets/documents/Air_Quality_Objectives_Update.pdf (accessed: 29/06/2018).

Borrego, C., Costa, A.M., Ginja, J., Amorim, M., Coutinho, M., Karatzas, K., Sioumis, Th., Katsifarakis, N., Konstantinidis, K., De Vito, S., Esposito, E., Smith, P., Andre, N., Gerard, Francis, P., Castell, L.A., Schneider, N., Viana, P., Minguillon, M., Reimringer, M.C., Otjes, W., von Sicard, R.P., Pohle, O., Elen, R., Suriano, B., Pfister, D., Prato, V., Dipinto, M., Penza, M., 2016. Assessment of air quality microsensors versus reference methods: The EuNetAir joint exercise. Atmospheric Environment 147, 246-263.

Brunt, H., Barnes, J., Longhurst, J.W.S., Scally, G., Hayes, E., 2016. Local air quality management policy and practice in the UK: The case for greater public health integration and engagement. Environmental Science & Policy 58, 52–60.

Carslaw, D.C., 2015. The openair manual - open-source tools for analysing air pollution data. Manual for version 1.1-4, King's College London.

Castell, N., Dauge, F.R., Schneider, P., Vogt, M., Lerner, U., Fishbain, B., Broday, D., Bartonova, A., 2017. Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? Environment International 99, 293 – 302.

DEFRA, 2015. Improving air quality in the UK tackling nitrogen dioxide in our towns and cities, UK overview document, December 2015. Available on: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/486636/aq-plan-2015-overview-document.pdf (accessed 09/10/2017).

DEFRA, 2017. Improving air quality in the UK: tackling nitrogen dioxide in our towns and cities, May 2017, Draft UK Air Quality Plan for tackling nitrogen dioxide. Available on: https://consult.defra.gov.uk/airquality/air-quality-plan-for-tackling-nitrogen-dioxide/supporting_documents/Draft%20Revised%20AQ%20Plan.pdf (accessed 05/10\2017).

ESCAPE, 2010. European Study of Cohorts for Air Pollution Effects (ESCAPE), Institute for Risk Assessment Sciences, Utrecht University, the Netherlands. URL: http://www.escapeproject.eu/ (accessed 08/01\2018).

Goswami, E., Larson, T., Lumley, T., Liu, L.-J.S., 2002. Spatial characteristics of fine particulate matter: identifying representative monitoring locations in Seattle. Washington. Journal of Air & Waste Management Association 52, 324–333.

Hoek, G., Beelen, R., De Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., et al., 2008. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmospheric Environment 42, 7561-7578.

Ivaskova, M, Kotes, P., Brodnan, M., 2015. Air pollution as an important factor in construction materials deterioration in Slovak Republic. Procedia Engineering 108, 131-138.

Kanaroglou, P.S., Jerrett, M., Morrison,J., Beckerman,B., Arain, A., Gilbert, N.L., Brook, J.R., 2005. Establishing an air pollution monitoring network for intraurban population exposure assessment: A location-allocation approach. Atmospheric Environment 39, 2399–2409.

Khallaf, M. (Ed.), 2011. The impact of air pollution on health, economy, environment and agricultural sources, InTech, DOI: 10.5772/17660. Available from: https://mts.intechopen.com/books/the-impact-of-air-pollution-on-health-economy-environment-and-agricultural-sources (accessed 29/03/2018).

Kracht O., Santiago J.L., Martin F., Piersanti A., Cremona G., Righini G.,Vitali L., Delaney K., Basu B., Ghosh B., Spangl W., Brendle C., Latikka J., Kousa A., Pärjälä E., Meretoja M., Malherbe L., Letinois L., Beauchamp M., Lenartz F., Hutsemekers V., Nguyen L., Hoogerbrugge R., Eneroth K., Silvergren S., Hooyberghs H.,Viaene P., Maiheu B., Janssen S., Roet D., Gerboles M., 2017. Spatial representativeness of air quality monitoring sites: Outcomes of the FAIRMODE/AQUILA intercomparison exercise. Publications Office of the European Union. ISBN 978-92-79-77218-4. ISSN: 1831-9424. DOI: 10.2760/60611. EUR 28987 EN.

Landrigan, P.J., 2017. Air pollution and health. The LANCET – Public Health 2 (1), 4 – 5. DOI: https://doi.org/10.1016/S2468-2667(16)30023-8.

LAQM.TG16, 2016. Local Air Quality Management, Technical Guidance (LAQM TG 16). Published by the Department for Environment, Food and Rural Affairs. Available online at: https://laqm.defra.gov.uk/documents/LAQM-TG16-April-16-v1.pdf (accessed 16/03/2019).

LAQM.TG9, 2009. Local Air Quality Management, Technical Guidance LAQM TG(9). Published by the Department for Environment, Food and Rural Affairs. Available online at: https://laqm.defra.gov.uk/documents/LAQM-TG-(09)-Dec-12.pdf (accessed 16/03/2019).

Lebret, E., Briggs, D., Reeuwijk, H., Fischer, P., Smallbone, K., Harssema, H., Kriz, B., Gorynski, P., Elliott, P., 2000. Small area variations in ambient NO_2concentrations in four European areas. Atmospheric Environment, 34 (2): 177-185.

Lewis A. C., Lee J. D., Edwards P. M., Shaw M. D., Evans M. J., Moller S. J., Smith K. R., Buckley J. W., Ellis M., Gillot S. R. and Whited A., 2016. Evaluating the performance of low cost chemical sensors for air pollution research. Faraday Discussions 189, 85-103. DOI: 10.1039/C5FD00201J.

Mead, M.I., Popoola, O.A.M., Stewart, G.B., Landshoff, P., Calleja, M., Hayes, M., Baldovi, J.J., McLeod, M.W., Hodgson, T.F., Dicks, J., 2013. The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks. Atmospheric Environment 70, 186-203.

Martin, F., Santiago, J.L., Kracht, O., Garcia, L., Gerboles, M., 2015. FAIRMODE Spatial representativeness feasibility study. JRC Technical Report ISBN 978-92-79-50322-1 (doi: 10.2788/49487). European Comission, Joint Reasearch Centre, Institute for Environment and Sustainability. Luxembourg: Publications Office of European Union. Available at https://ec.europa.eu/jrc/en/publication/fairmode-spatial-representativeness-feasibility-study.

Munn, R.E., 1981. The objectives of air quality monitoring programmes. In: The Design of Air Quality Monitoring Networks. Air Pollution Problems Series. Palgrave Macmillan, London. DOI: https://doi.org/10.1007/978-1-349-05738-2_2.

Office for National Statistics, 2018. RL - https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography#output-area-oa (Accessed 20/11/2018).

Raffuse, S.M., Sullivan, D.C., McCarthy, M.C., Penfold. B.M., Hafner, H.R., 2007. Ambient air monitoring network assessment guidance: analytical techniques for technical assessments of ambient air monitoring networks. US Environmental Protection Agency (EPA-454/D-07-001). Available online: https://www.epa.gov/sites/production/files/2017-02/documents/network-assessment-guidance.pdf (accessed: 16/03/2018).

Righini, G., Cappelletti, A., Ciucci, A., Cremona, G., Piersanti, A., Vitali, L., Ciancarella, L., 2014. GIS based assessment of the spatial representativeness of air quality monitoring stations using pollutant emissions data, Atmospheric Environment 97, 121-129.

Santiago J.L., Martin F., Martilli, A.,2013. A computational fluid dynamic modelling approach to assess the representativeness of urban monitoring station, Science of the Total Environment 454, 61-72.

Schneider, P., Castell, N., Vogt, M., Dauge, F.R., Lahoz, W.A., Bartonova, A., 2017. Mapping urban air quality in near real-time using observations from low cost sensors and model information. Environment International 106, 234 – 247.

SCC, 2016. Sheffield City Council 2016, Air Quality Annual Status Report (ASR), in fulfilment of Part IV of the Environment Act 1995 Local Air Quality Management, June 2016.

Walters, S., Ayres, J., 2001. The health effects of air pollution. in pollution causes, effects and control. In Harrison, R.M. (Ed.), Chapter 11, pp. 275. Fourth Editions, Royal Society of Chemistry, Cambridge, UK. ISBN 0-85404-621-6.

WHO, 2013. World Health Organization, review of evidence on health aspects of air pollution-REVIHAAP. WHO, Copenhagen.

Wu, H., Reis, S., Lin, C., Heal, M.R., 2017. Effect of monitoring network design on land use regression models for estimating residential NO2 concentration. Atmospheric Environment 149, 24 - 33.

CHAPTER 5: ANALYSIS OF AIR POLLUTION IN URBAN AREAS WITH AIRVIRO DISPERSION MODEL - A CASE STUDY IN THE CITY OF SHEFFIELD, UNITED KINGDOM

Said Munir [1,*], Martin Mayfield [2], Daniel Coca [2], Lyudmila S Mihaylova [2] and Ogo Osammor [3]

[1] Department of Civil and Structural Engineering, the University of Sheffield, UK, S1 3JD
[2] Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield, UK, S1 3JD
[3] Air Quality Monitoring & Modelling, Sheffield City Council, Howden House, 1 Union Street, Sheffield, UK, S1 2SH
* Correspondence: smunir2@sheffield.ac.uk

Published: Atmosphere 2020, 11(3), 285
https://doi.org/10.3390/atmos11030285
Received: 25 February 2020 / Revised: 12 March 2020 / Accepted: 13 March 2020 / Published: 15 March 2020.

Abstract

Two air pollutants, oxides of nitrogen (NOx) and particulate matter (PM_{10}) are monitored and modelled employing Airviro air quality dispersion modelling system in Sheffield, United Kingdom. The aim is to determine the most significant emission sources and their spatial variability. NOx emissions (ton/year) from road traffic, point and area sources for the year 2017 were 5370, 6774 and 2425, whereas that of PM_{10} (ton/year) were 345, 1449 and 281, respectively, which are part of the emission database. The results showed three hotspots of NOx, namely the Sheffield City Centre, Darnall and Tinsley Roundabout (M1 J34S). High PM_{10} concentrations were shown mainly between Sheffield Forgemasters International (a heavy engineering steel company) and Meadowhall shopping centre. Several emission scenarios were tested which showed that NOx concentrations were mainly from road traffic, whereas PM_{10} concentrations were from point sources. Spatiotemporal variability and public exposure to air pollution were analysed. NOx concentration was greater than 52 $\mu g/m^3$ in about 8 km^2 area, where more than 66 thousand people lived. Models validated by observations can be used to fill-in spatiotemporal gaps in measured data. The approach used presents spatiotemporal situation awareness maps that could be used for decision making and improving the urban infrastructure.

Keywords: Air quality modelling; urban air quality; Airviro; dispersion modelling; Sheffield; Emissions.

5.1. Introduction

Air pollution has a significant negative impact on urban areas, especially in large megacities and roadside locations, where air pollution has become a growing issue for public health. Air pollution is causing numerous human health and environmental problems. Polluted air, especially with high levels of nitrogen dioxide (NO_2) and particulate matter with aerodynamic diameter of up to 10 μm (PM_{10}) and 2.5 μm ($PM_{2.5}$), is considered the most serious environmental risk to public health in urban areas in the UK [1]. Atmospheric pollutants were estimated to cause three million premature deaths in 2012 worldwide [2], whereas according to Landrigan [3] air pollution caused 6·4 million deaths worldwide in 2015. Air pollutants (e.g., NO_2 and PM_{10}) emitted by various emission sources are reported to cause heart disease, lung cancer, and both chronic and acute respiratory diseases including asthma [2].

Air pollution models are numerical tools for describing the causal relationship of atmospheric pollutant concentrations with emissions, meteorology, deposition, chemical transformation and other factors like topography [4]. Modelling outputs can support Air Quality Monitoring Networks (AQMN) by improving spatial and temporal coverage in urban areas. Quantifying emission sources and predicting air pollutant concentrations, using dispersion modelling approaches like Airviro, AERMOD or ADMS-Urban can help understand the main drivers of air pollution. Outputs of the dispersion model in the form of contour maps highlight spatial variability of air pollutant concentrations in the cities and provide a continuous map that can be used to further fine tune the air quality monitoring network.

Air quality modelling is carried out for several purposes including air quality prediction/forecasting, analysing the dispersion of air pollutants in the atmosphere based on emissions and meteorological parameters, quantifying the impacts of air pollution (e.g., health impacts), modelling the impacts of various factors on air pollution, analysing the relationship between different pollutants, modelling pollution processes and transport, quantifying deposition and environmental fate of pollutants, running and testing emission scenarios, quantifying the emissions of air pollutants from various emission sources, determining long-term trend in air pollutant concentrations, and producing high resolution spatiotemporal maps of air pollution [5 - 12]. Dispersion models are also used for emergency planning of accidental chemical releases [13].

Air quality models can be divided into two main types, namely statistical and dispersion models. There is a wealth of statistical models, including time series (e.g., autoregressive moving average), regression (e.g., multiple linear regression), classification (e.g., logistic and discriminant analysis) and resampling methods (e.g., cross-validation and bootstraps) [14]. Dispersion modelling is further divided into Gaussian, Eulerian, Lagrangian, Computations Fluid Dynamics, Photochemical, Dense-gas, Aerosol Dynamic and Box models [15:17]. A number of techniques exist for modelling the dispersion of air pollutants, which vary in sophistication but all include some sort of simplification of the real dispersion processes. Selecting a particular modelling approach depends on several factors such as the temporal and spatial scale of the model, resolution of the data available to run the model, the purpose of the modelling, skill of the modeller, time, financial and computer resources available [16]. Dispersion modelling techniques for air quality are effective tools for determining downwind concentrations of air pollutants at a given time and space emitted by a known emission source, for example, an industrial plant or a nearby road. Dispersion models are mathematical representation of the atmospheric processes determining the rate at which pollutants are mixed with clean air. Results of dispersion models can be used to replace the need for air pollutant monitoring. However, high cost of purchase and maintenance, high input demands and requirement of skilled staff limit their application.

In this paper the Airviro version 4.01dispersion modelling system [18] is employed to model the emissions of NOx and PM_{10} in Sheffield, United Kingdom. The aim is to analyse different emission sources of air pollutants, investigate spatial variability of the pollutants, identify their main hotspots in Sheffield and assess the performance of Airviro model by comparing the modelled and observed concentrations in Sheffield. Furthermore, several emission scenarios will be tested which help identify the main drivers of air pollution in Sheffield. The rest of the paper is structured as follows: Methodology of this paper is presented in section 5.2, where section 5.2.1 presents the proposed framework, which includes the model for the air quality assessment of the Airviro dispersion. The air quality data and meteorological data are presented in section 5.2.2. The assessment and validation of the proposed model is given in section 5.2.3. Section 5.3 presents results and their analysis. The main outcomes of this work are summarised in section 5.4.

5.2. Methodology

This paper presents a spatiotemporal analysis of the two air pollutants (PM_{10} and NOx) sources, emissions and atmospheric concentrations in the city of Sheffield, United Kingdom. Air quality modelling and monitoring support each other in several ways. Air quality monitoring provides data only for points where the sensors are installed for the past and present time, whereas models provide better spatial coverage and in addition to past and present, can predict air pollution concentrations for the future. Models validated by observations can be used to fill-in spatiotemporal gaps in measured data. Figure 5.1 compares air quality monitoring and modelling and shows how they support each other.

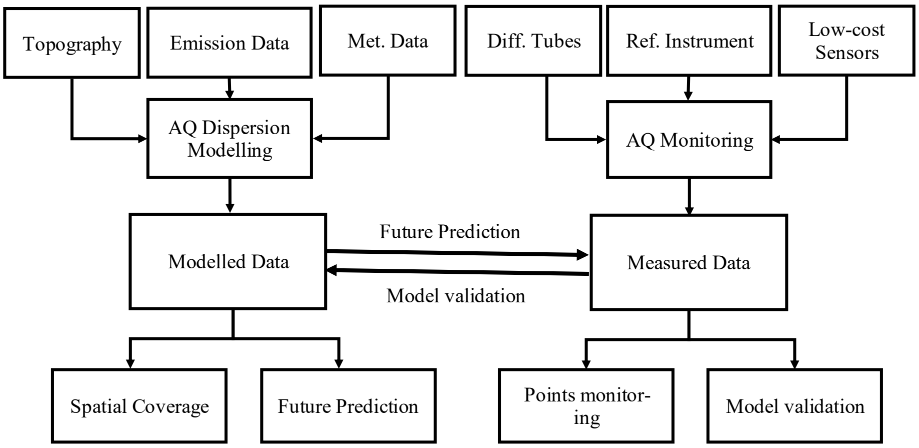


Figure 5.1. Schematic diagram comparing air quality modelling and monitoring and how they support each other.

5.2.1. Dispersion modelling system – Airviro

In this paper Airviro version 4.01 was used, which is an air quality management system developed by Apertum [18, 19]. Airviro is an integrated modelling system for managing emission inventories (Emission Data Base – EDB), modelling dispersion of pollutants and data handling. Airviro is a state of the art dispersion model used by many researchers, consultants and local authorities globally for air quality modelling. Airviro has several modelling options [18, 19]. In this study for urban scale modelling in Sheffield, a Gaussian model is used, which was introduced by Pasquill [20-22] and Briggs [23]. The Gaussian modelling technique is simple, computationally efficient and requires simple input data. Gaussian modelling is normally used for fast screening type calculation of the pollutant dispersion in urban areas from point sources, line sources or area sources. Urban areas can be modelled as a sum of area sources (e.g., domestic and commercial emission), point sources (e.g., factory or power stations), and line sources (e.g., road traffic).

In this paper two air pollutants NOx and PM_{10} are modelled using local topography, emissions and meteorological data to produce air quality maps of the estimated pollutants. In this paper emissions of NOx and PM_{10} from road traffic, point sources and area sources are modelled and the estimated concentrations are compared with measured concentrations from several sensors. Dispersion models convert pollutant emissions to atmospheric concentrations in the form of contour maps and receptor points. This paper presents emission data, real case studies and the developed approaches for air pollution modelling and prediction.

In additional to annual levels, NOx and PM_{10} concentrations are estimated for spring, summer, autumn and winter seasons and compared with measured concentrations at three receptors points (as discussed in section 2.3).

5.2.2. Emission and meteorological data

Sheffield (53°23′N, 1°28′W) is a metropolitan borough and a vibrant city in South Yorkshire, United Kingdom. Historically known as the Steel City, Sheffield no more has the smoking chimney stacks and has emerged as a green and modern cityscape in the proximity of the Peak District National Park. According to 2011 census Sheffield City had a population of 552,700, however, since then the population has grown and according to more recent estimates has reached about 700,000. Its elevation above sea level ranges from 29 m near Blackburn Meadows to 548 m near Margery Hill. Sheffield, like most of the United Kingdom, has a temperate climate with average maximum temperature of 20.8°C in June and July and average minimum temperature of 1.6°C in January and February.

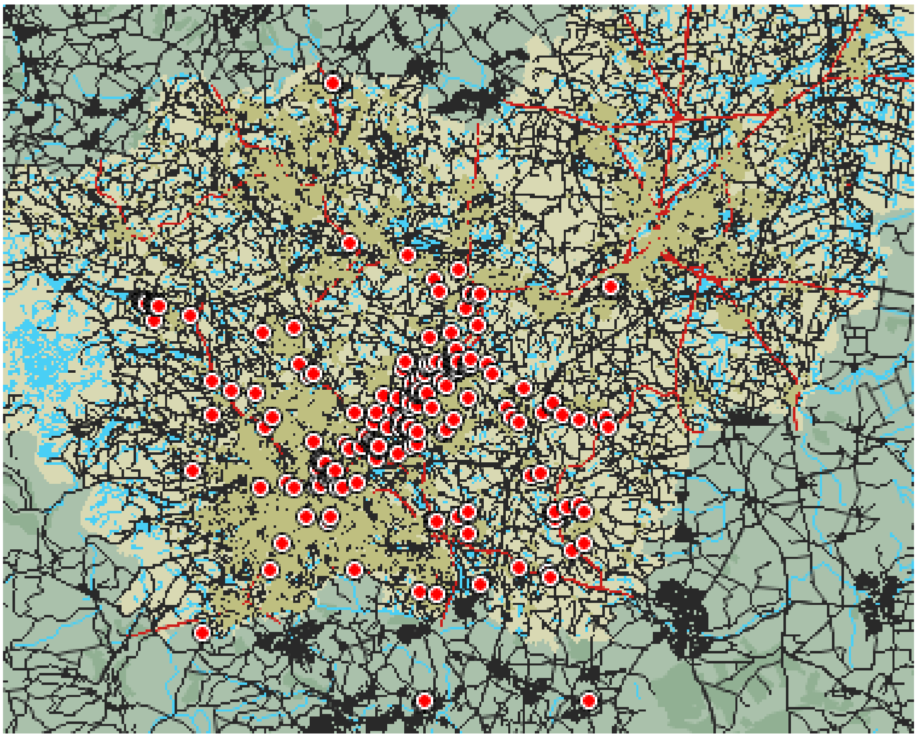
Both spatial road traffic and points air pollutant emission sources are shown in Figure 5.2 (a and b). In this paper the Airviro EDB for Sheffield was used, which contains detailed information on the sources of emissions and allows for emission rates for various types of emission sources such as point sources, area sources or road sources to be calculated. This EDB has an updated 2017 road traffic data (emission factors, traffic counts, vehicle speed, and fleet composition), area sources (commercial and domestic emissions) and point (industrial) sources. The new emission factors (Emissions Factor Toolkit v8.0.1b) include non-exhaust particulate matter such as resuspension of dust particles. The EDB also takes account of spatiotemporal variability in emission rates. Emissions are calculated as a function of day type (e.g., weekday or weekend), hour of day, and month of year. Temporal resolution of emissions is hourly, whereas the model outputs are presented in different time resolutions including hourly, seasonal or annual. For example, to calculate emissions for a road segment, the emission database uses some basic information including road name, road type (e.g., urban road, motorway etc.), vehicle speed, traffic counts, fleet composition, number of lanes, road length, slope and elevation. Likewise, to calculate emission from a point source, the emission database uses various information of the source such as name, location, coordinates, chimney characteristics (e.g., chimney height, outer and inner diameter of chimney, exhaust gas temperature and exhaust gas velocity), and characteristics of the emissions such as substance (e.g., NOx, PM_{10}, SO_2), amount of emissions and time variation of emissions. Full details on Airviro EDB are provided by Airviro User´s Reference [13]. Spatial resolution of area sources, which includes residential and commercial sources is 1km x 1km. Point sources emissions are calculated individually. The Airviro EDB for 2017 shows that there were 268 point sources emitting 12999 ton/year SO_2, 6774 ton/year NOx, 3077 ton/year CO and 1449 ton/year PM_{10} in Sheffield and Rotherham. On the other hand, road sources emitted 5370 ton/year NOx and 345 ton/year PM_{10}. Emission data of some pollutants were missing in the database (Table 5.1). Several pollutants have demonstrated reduction in their emissions, e.g., SO_2 from road sources due to the effects of Directive 2005/33/EC of the European Parliament on the removal of sulphur from road traffic fuels. In addition, improvements in road vehicle fuels and technologies as a result of EU Directives on emission standards also resulted in reductions in CO from 65% (1990) to 16% (2017), in conjunction with reductions from industrial sources "due to the decline in the use of solid fuels in favour of gas and electricity, as well as a decline in the production of steel and non-ferrous metals" [24]. This study considers only two pollutants i.e. NOx and PM_{10}. Emissions of NOx and PM_{10} are shown in Figure 5.3 (a and b), respectively, demonstrating their spatial variability and highlighting hotspots in terms of pollutant emissions [(ton/year)/km^2].

Hourly meteorological data including temperature, relative humidity, wind speed and direction (Figure 5.4) were measured at Woodburn Road weather mast in Sheffield, located at Sheffield Hallam University City Athletics Stadium. These were used for the simulation of hourly NOx and PM_{10} concentrations, whereas for monthly, seasonal and annual scenarios, Airviro used a statistical approach for estimating such meteorological conditions. Before model runs, the meteorological data were used to determine the boundary layer scaling parameters – surface friction velocity and the Monin-Obukhov length. The wind fields were simulated using the diagnostic wind model available in Airviro, which considered the effects of topography, surface roughness and surface adiabatic heating/cooling [25]. Atmospheric conditions are classified into six (6) stability classes in the model: very stable, stable, neutral negative, neutral positive, unstable and very unstable [26].

Table 5.1. Emission of various air pollutants (ton/year) in Sheffield and surrounding areas.

Pollutant	Point sources	Road sources	Area sources	Total Emission
SO_2	12999		1157	14156
NOx	6774	5370	2425	14569
NO_2	122			122
CO	3077		2203	5280
PM_{10}	1449	345	281	2075
$PM_{2.5}$		224		224

(a) Point sources



(b) Road network

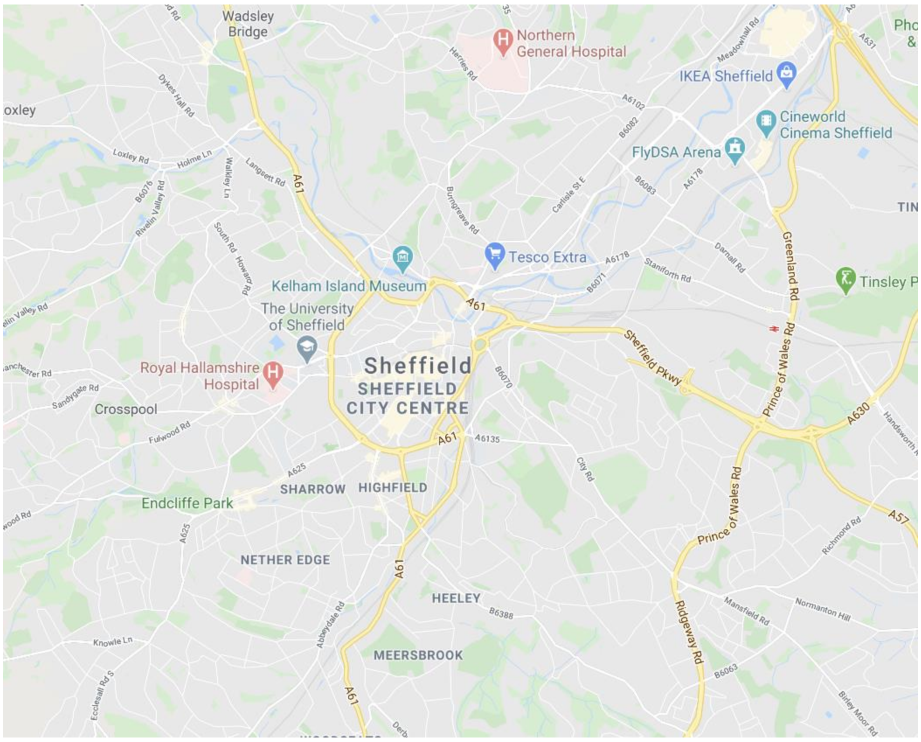
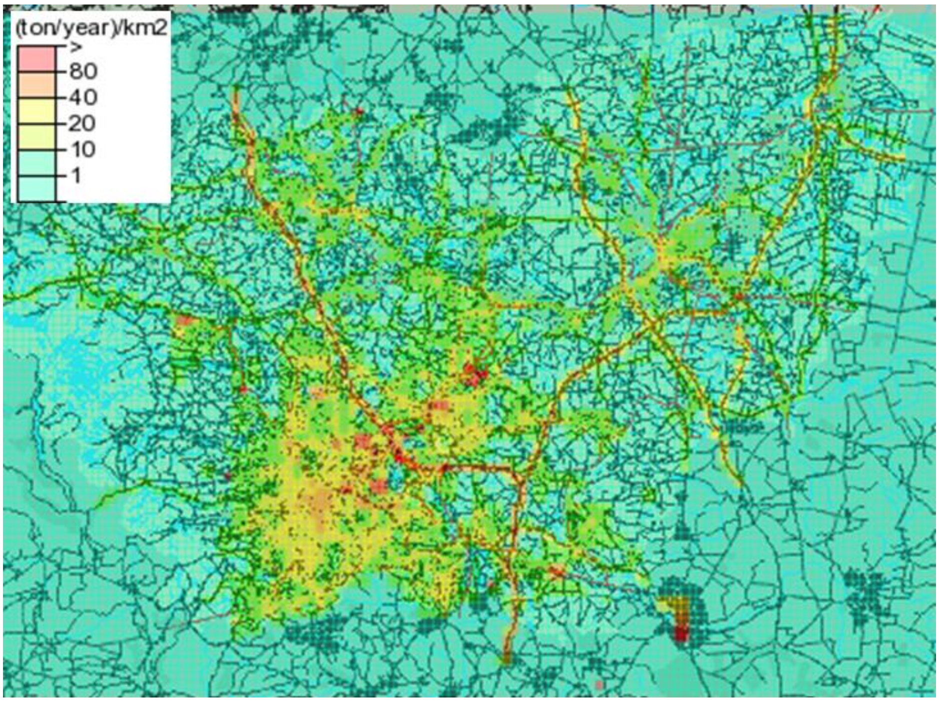
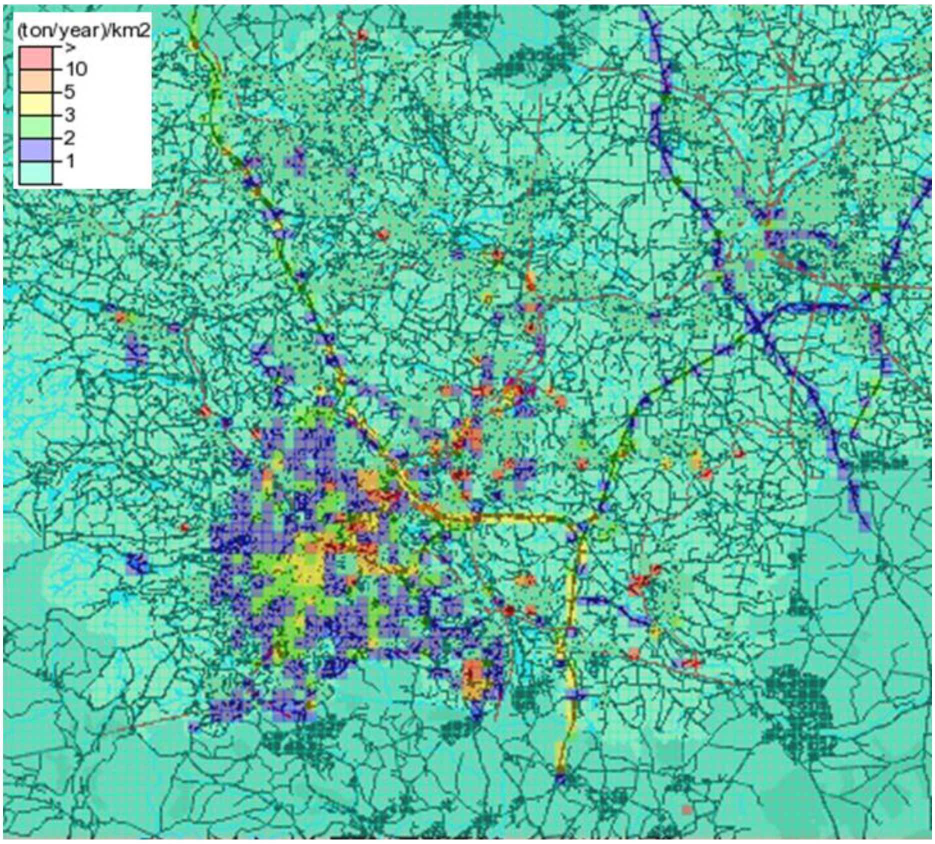


Figure 5.2. Showing emission sources in Sheffield and surrounding areas including Rotherham for year 2017: (a) point sources, (b) road network or line sources.



(a) NOx emissions



(b) PM10 emissions

Figure 5.3. Showing emissions' strength [(ton/year)/km^2] in Sheffield and surrounding areas including Rotherham for year 2017 from all emission sources: (a) NOx emissions, (b) PM10 emissions.

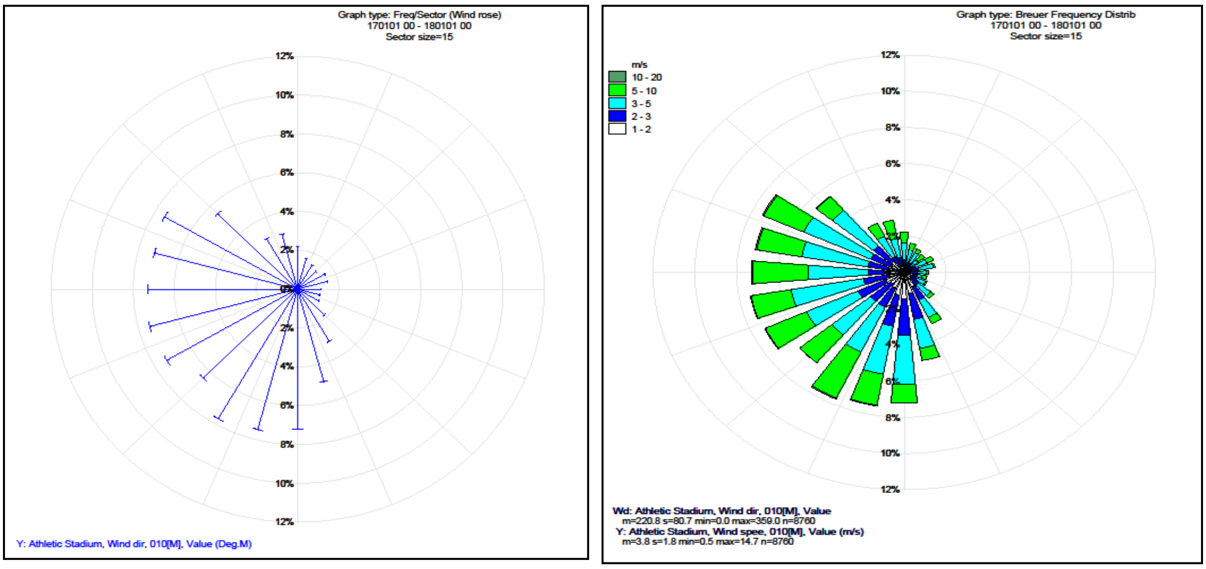


Figure 5.4. Breuer frequency distribution (wind rose) for year 2017 weather Data, used in the Airviro model.

5.2.3. Model assessment

PM_{10} and NOx concentrations (μg/m^3) predicted by Airviro model are presented in the form of contour maps. Predictions are also made for three receptor points namely Devonshire Green, Sheffield Tinsley and Barnsley Road air quality monitoring stations, which are part of the UK Department for Environment, Food and Rural Affairs (DEFRA) Automatic Urban and Rural Network (AURN). Details of these sites are provided in Table 5.2. Measured and modelled NOx and PM_{10} concentrations are compared at these sites to assess the performance of the model. Comparisons are made for annual concentrations as well as for winter, spring, summer and autumn seasons. Furthermore, hourly predicted and observed concentrations are compared for the month of January representing winter and July representing summer season 2017. Several statistical metrics are calculated to assess the performance of the model. The metrics used in this paper are the correlation coefficient (r), Root Mean Square Error (RMSE), Mean Bias (MB), Normalised Mean Bias (NMB), Mean Absolute Error (MAE), Normalised Mean Absolute Error (NMAE), Factor of two (FAC2) and Coefficient of Efficiency (COE).

The RMSE and Root Mean Square Deviation (RMSD) provide a good measure of the model error by calculating how close or far the predicted values are to the observed values. The RMSE measures the difference between the predicted and observed concentrations. The RMSE is a non-negative quantity and ideally we want it to have a zero value which means a perfect fit of the model having no error. MB is simply the average bias between the predicted and observed values. NMB is calculated by adding up the difference between the predicted and observed values (∑ (predi – obsi)) and normalising it by the sum of the observed values (∑obsi). NMB is reported as a percentage (%). NMB estimates average over or under prediction and its value between +0.02 and -0.02 shows acceptable model performance. MAE provides a good indication of the mean absolute error and is in the same units as the quantities being considered. The MAE is normalised dividing it by the observed value. The normalised value is known as NMAE. The correlation coefficient (r) characterises the strength of the linear relationship between two variables i.e. modelled and observed concentrations. The closer to one (±1) the value of 'r' is, the better the similarity is. Generally, a value ranging from ±0.5 to ±0.99 indicates reasonably good performance of the model. FAC2 is the fraction of modelled values within a factor of 2 of the observed values. FAC2 should satisfy the condition that 0.5 ≤ predi/obsi ≤ 2. The ideal value for the FAC2 is 1 (100%). A highly efficient or perfect model should have COE value of 1. However, when analysing real data, a model has a COE value of less than 1. COE having a zero value (COE = 0) means the model prediction is not better than the mean of the observed value, which in other words means its prediction power is zero or it has no predictive advantage. These metrics are further described by [27] and [28].

Table 5.2. Air quality monitoring sites (receptor points) where measured and modelled concentrations are compared.

Site name	Site type	Easting (X)	Northing (Y)	Pollutant monitored	Monitoring Technique	Distance to road (m)	Inlet height (m)
Sheffield Tinsley (SHE)	Urban Industrial	440215	390598	NO_2	Chemiluminescence	120 (M1)	3
Sheffield Devonshire Green (SHDG)	Urban Background	435158	386885	NO_2, PM_{10}, $PM_{2.5}$, O_3,	Chemiluminescence, TEOM, UV Absorption	20	3
Sheffield Barnsley Road (SHBR)	Urban Traffic	436276	389930	NO, NO_2, NOx	Chemiluminescence	4.5	2

5.3. Results and discussion

Employing the Airviro model, emissions (ton/year) of NOx and PM_{10} are modelled to produce atmospheric concentrations (μg/m^3) of these pollutants in the form of contour maps for year 2017 in Sheffield. In sections 3.1 results of NOx and section 3.2 results of PM_{10} are presented and discussed.

5.3.1. NOx maps

Figure 5.5 shows modelled annual average NOx concentrations (μg/m^3) in the form of contour maps using traffic, points and area emission sources employing Gauss module for scenario 2017. Figure 5.5 shows three areas of high NOx concentrations in the city namely Sheffield City Centre, Darnall and near Tinsley Roundabout (M1, J34S). High levels of NOx are also predicted on Sheffield Parkway (A630, A57) and between Meadowhall Shopping Centre and Sheffield Forgemasters International (a heavy engineering steel company). Sheffield City Centre probably experiences the highest levels of NOx, which is mainly due to high level of road traffic but various point and area sources also contribute. Pollution levels are highest in the busiest part of the city including St. Mary's Gate, More Street, Eyre Street, Arundel Gate, Sheaf Street, Pond Street, Exchange Place and Castlegate. The train station, the bus station, the area of the Sheffield Hallam University and a busy shopping centre make this area very busy in terms of road traffics, exposing visitors, workers, students, commuters and residents to high levels of air pollution. Areas adjacent to the city centre also experience considerable amount of air pollution. Generally, pollution levels gradually decrease with distance from the city centre. However, due to prevailing south westerly winds and the locations of some industrial sources, there is a north-eastern trend in air pollution levels i.e. north-eastern region towards M1 experiences considerably high amount of air pollution (39 – 52 μg/m^3) (Figure 5.5). In the north-eastern region, there are three hot spots where air pollution levels are higher (NOx levels > 65 μg/m^3), namely Darnall, near Tinsley Roundabout on M1 J34S and between Sheffield Forgemaster International and Meadowhall shopping centre. The reasons for these hotspots are high traffic levels and some heavy steel industries. The north-eastern region experiences considerably more road traffic due to the Motorway (M1) and several major roads to-and-from Sheffield City Centre, Meadowhall and the M1.

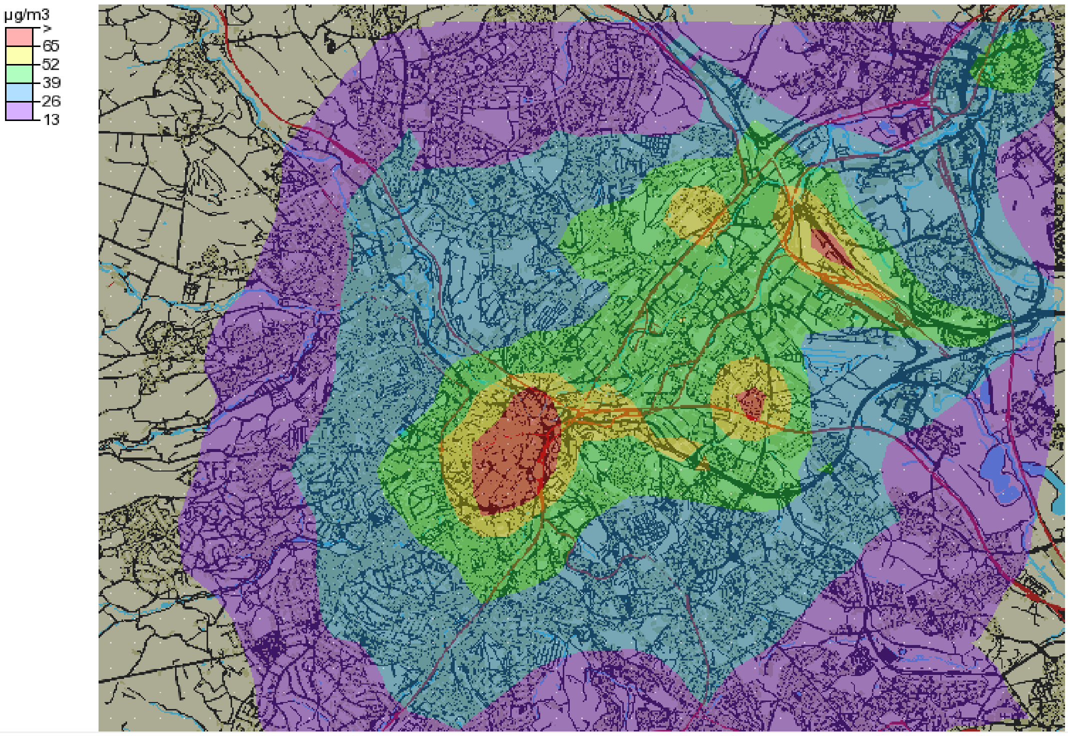
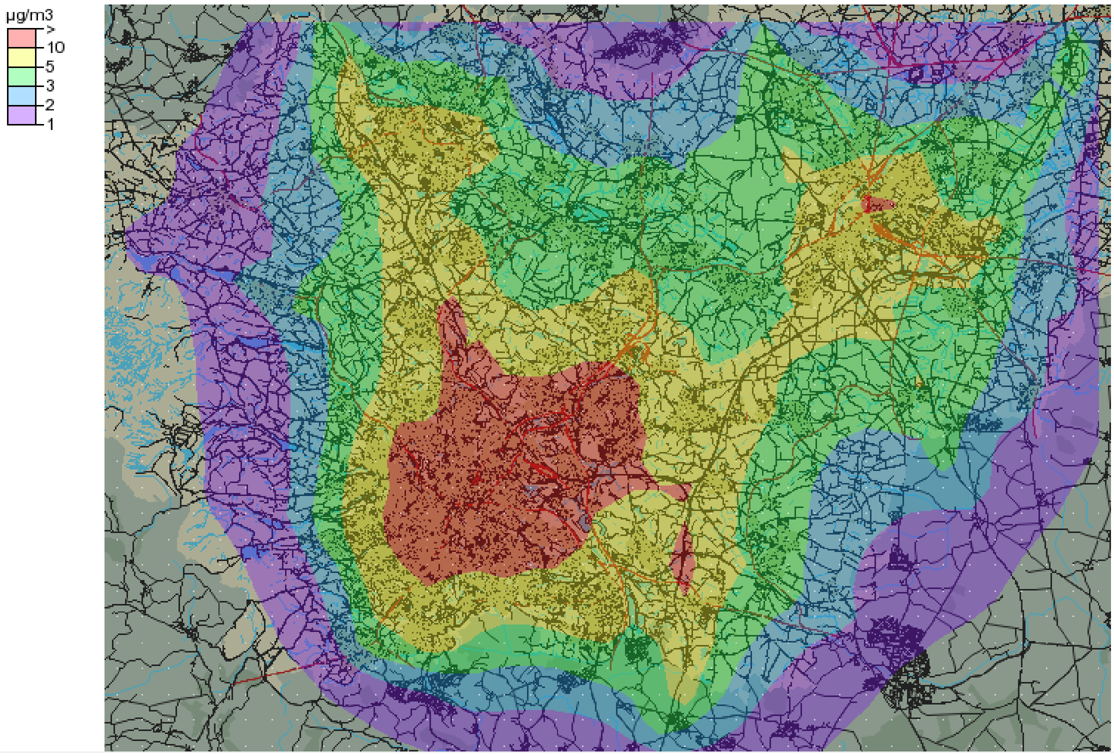


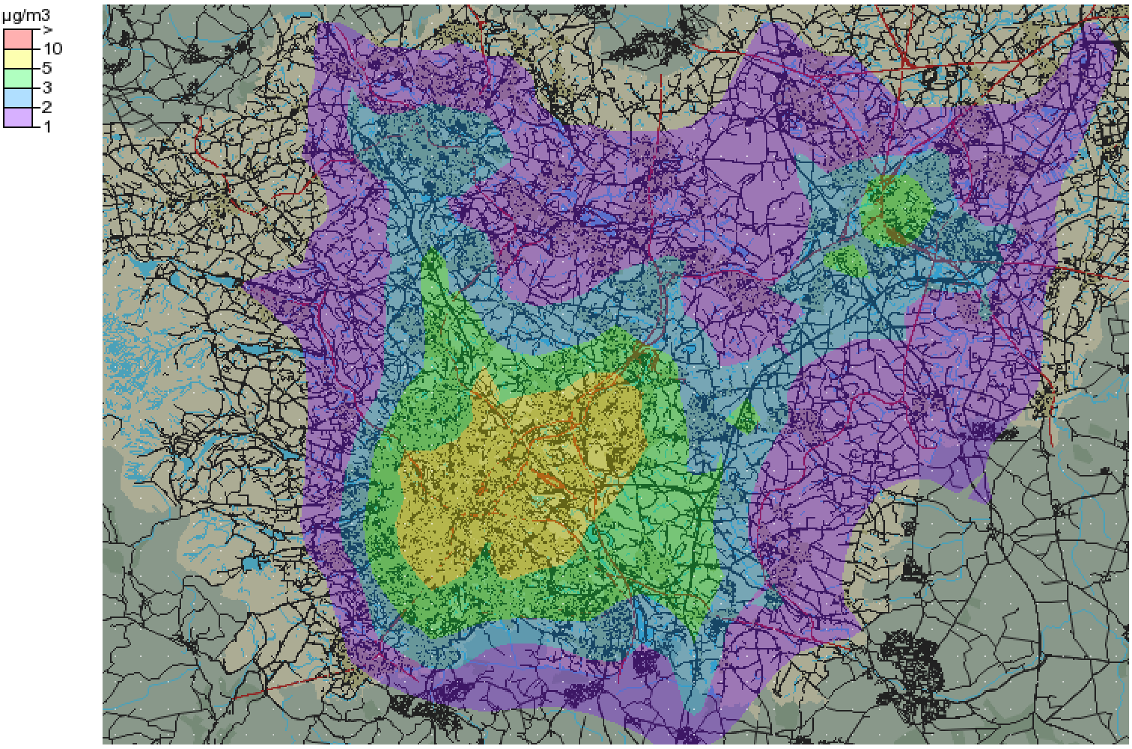
Figure 5.5. Estimated annual mean NOx concentrations (μg/m^3) using all emission sources in Sheffield for year 2017.

Figure 5.6 shows the effect of various emission scenarios on atmospheric NOx concentrations in Sheffield, where Figure 5.6 (a) represents emissions of road traffic, (b) HGV (Heavy Goods Vehicles) and LGV (Large Goods Vehicles), (c) cars both petrol and diesel, (d) buses, and (e) points sources in Sheffield. Here HGV and LGV are European terms used for any vehicles or trucks with a gross weight of over 3500 kg. These different categories of emissions result in different spatial variability of NOx. The outputs of these scenarios show that the levels of NOx pollution are mostly controlled by road traffic. HGVs seem to emit a considerable amount of emissions in the city centre (Figure 5.6b) and their contribution is greater than that of the buses or cars. Scenario (6a) which represents all traffic modes shows high pollution levels in the city centre and surrounding areas. Furthermore, it shows higher pollution levels along Sheffield Parkway leading to Motorway (M1) and along Penistone Road (A61). The other traffic scenarios show moderate levels of air pollution in the city centre and the adjacent areas. Figure 5.6(e) which considers only point sources represents totally different spatial pattern of NOx concentrations. In this scenario the hotspots in the city centre and surrounding areas have disappeared. Here the two hotspots are in Tinsley (near J34SM1) and Attercliffe (near A6109). In Tinsley there are about 15 point sources, whereas on both sides of A6109 (where the hotspot is shown) there are about 13 point sources, which are the most likely reasons for these two hotspots.

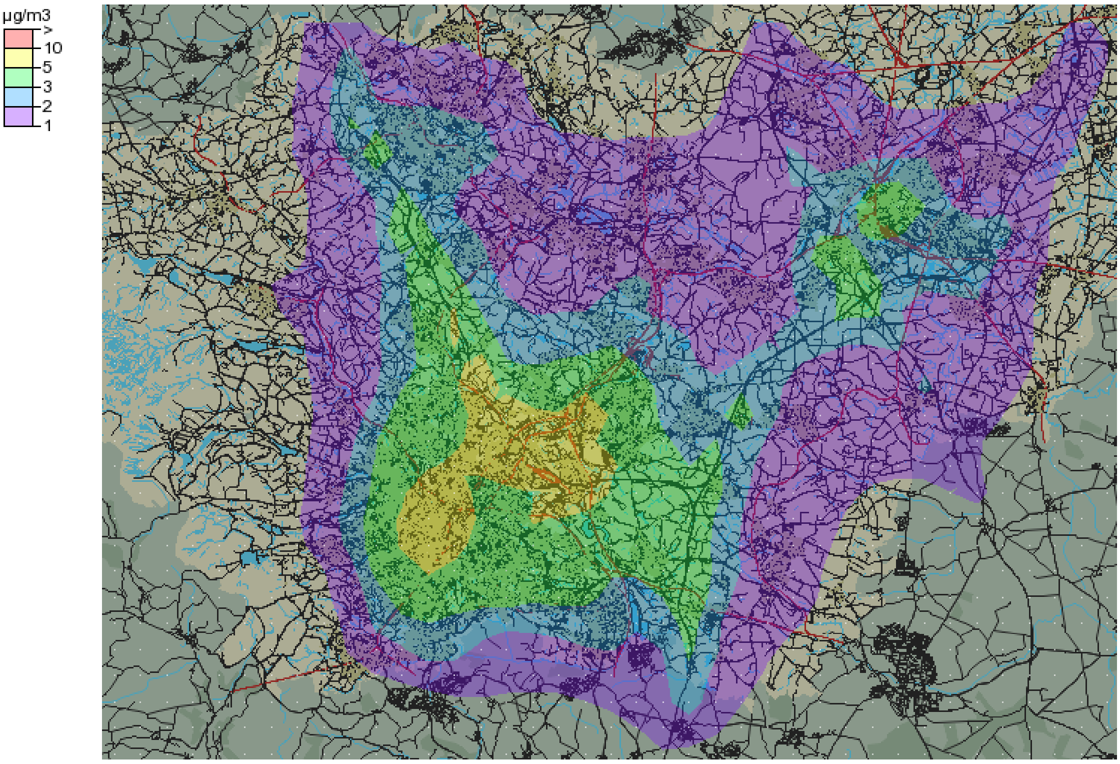
(a) Road traffic

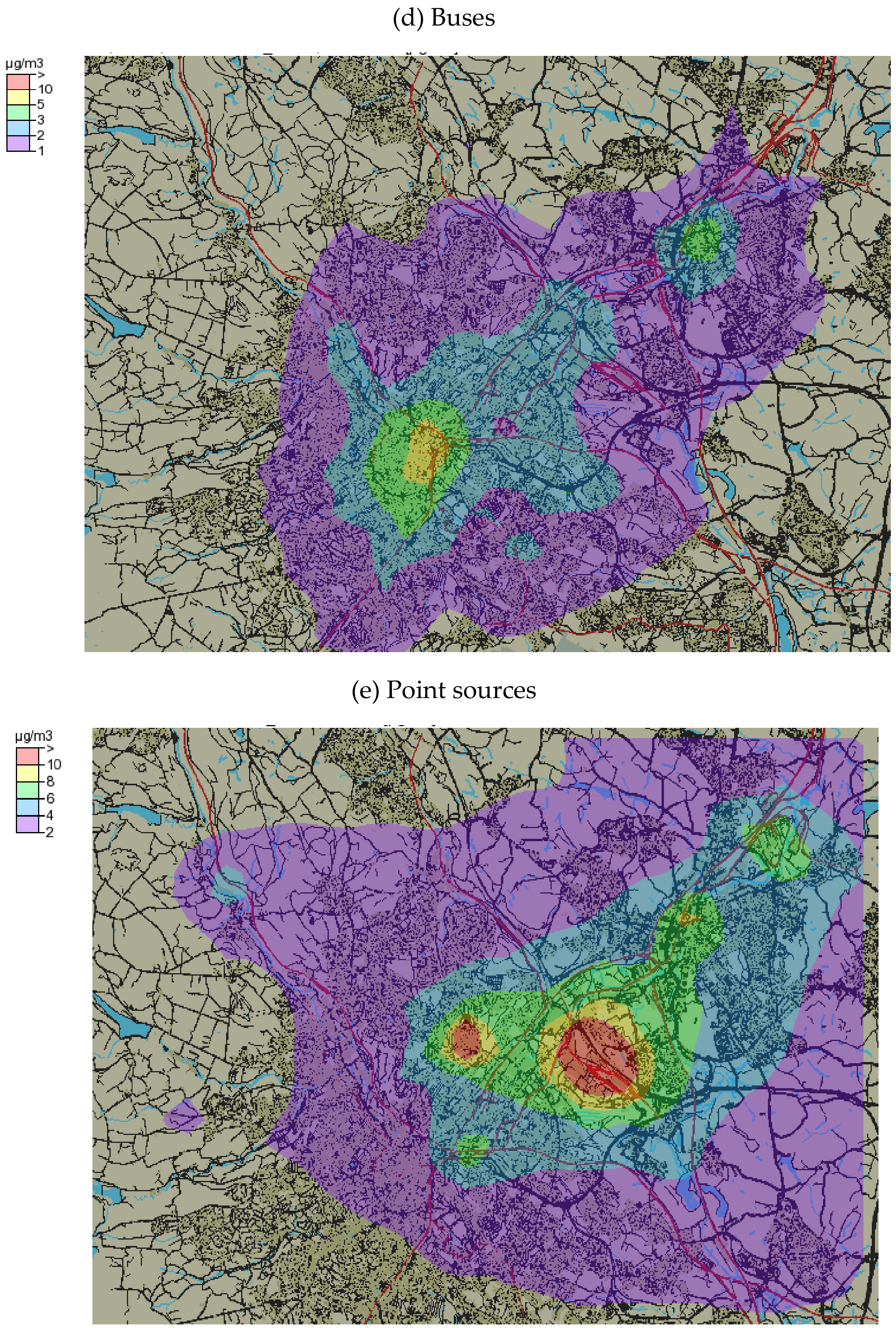


(b) HGV and LGV



(c) Cars





(d) Buses

(e) Point sources

Figure 5.6. Estimated mean annual NOx concentrations (µg/m^3) using emissions from road traffic (a), HGV and LGV (b), cars both petrol and diesel (c), buses (d), and points sources (e) in Sheffield for year 2017.

Atmospheric levels of NOx were estimated in various seasons of the year Figure not shown for brevity) namely: (a) winter (November, December, January), (b) spring (February, March, April), (c) summer (May, June, July), and (d) autumn (August, September, October). Generally atmospheric pollutant levels were higher in colder seasons due to (i) greater combustion of fossil fuels, mainly diesel, petrol, gas and coal to a lesser extent, and (ii) atmospheric stagnation and shallower atmospheric boundary-layer height which discourage pollutant dispersion as compared to hotter seasons when the atmosphere is more turbulent and boundary layer height is wider. Mainly there are four hotspots, which are Sheffield City Centre, Tinsley, near Forgemaster International and Darnall. The four seasons demonstrated slightly different patterns and levels of NOx. Relatively higher levels of NOx were predicted in winter and autumn in which minimum, mean and maximum levels (µg/m^3) were 4, 23, 87 and 2, 21, 91, respectively. Summer and spring showed relatively lower NOx levels in which minimum, mean and maximum NOx levels (µg/m^3) were 2, 16, 68 and 2, 19, 81, respectively.

5.3.2. PM$_{10}$ maps

Figure 5.7 shows the results of modelled annual average PM$_{10}$ concentrations (µg/m^3) in the form of contour maps using road traffic, point and area emission sources. Gauss module of Airviro for scenario 2017 was employed in this study. In contrast to NOx concentrations, the areas with elevated PM$_{10}$ concentrations are shown outside Sheffield City Centre. The highest PM$_{10}$ pollution levels were observed between Meadowhall shopping centre and Sheffield Forgemaster International. This hotspot seems to be due to the point sources (heavy steel and other companies) in this area. However, this is also a busy area in terms of road traffic, which must be contributing a significant amount of emissions. The second hotspot of PM$_{10}$ is shown in Attercliffe, which has six point sources emitting a significant amount of PM$_{10}$ and other pollutants. The third hotspot of PM$_{10}$ is shown in Sheffield Parkway having three point sources. The fourth hotspot is the north-eastern corner of the city centre in the Wicker and West Bar near Derek Dooley Way. The model results demonstrate that PM$_{10}$ levels are affected more by the heavy industries rather than the road traffic, which is in contrast to NOx levels which are more linked with road traffic. Figure 5.8 showing PM$_{10}$ concentrations estimated from point sources (upper panel) and road traffic (lower panel) provides further evidence that PM$_{10}$ levels in Sheffield are more affected by point sources. When only point sources were considered as inputs to the model, the minimum, mean and maximum PM$_{10}$ levels (µg/m^3) were 0.24, 1.3 and 13.9, respectively, whereas when only road traffic was considered these levels were 0.04, 0.56 and 2.5, respectively.

PM$_{10}$ concentrations (µg/m^3) in various seasons of the year were also modelled and compared. PM$_{10}$ concentrations slightly varied and showed slightly different spatial pattern in various seasons. The minimum, mean, and maximum PM$_{10}$ concentrations (µg/m^3) in winter, spring, summer, and autumn were: 1.0, 4.12, 15.4; 0.66, 3.75, 15.4; 0.47, 3.47, 15.5; and 0.75, 4.49, 19.0, respectively. Autumn showed relatively higher average (4.49 µg/m^3) and maximum (19.0 µg/m^3) PM$_{10}$ concentrations compared to other seasons.

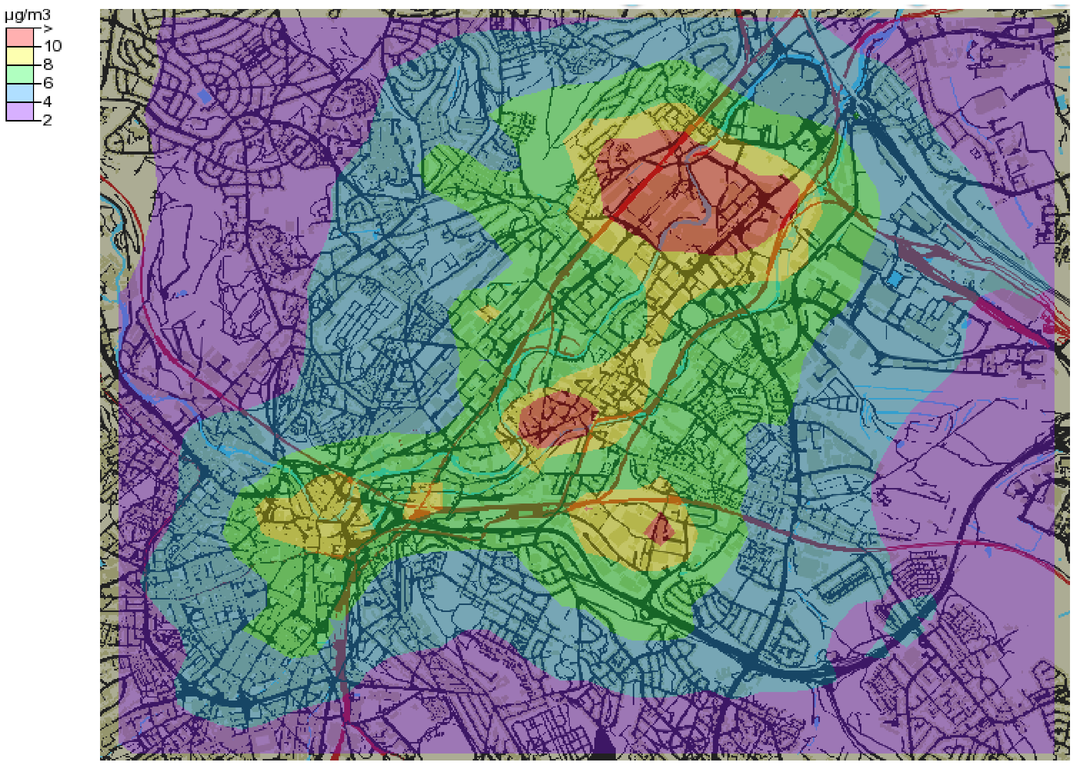
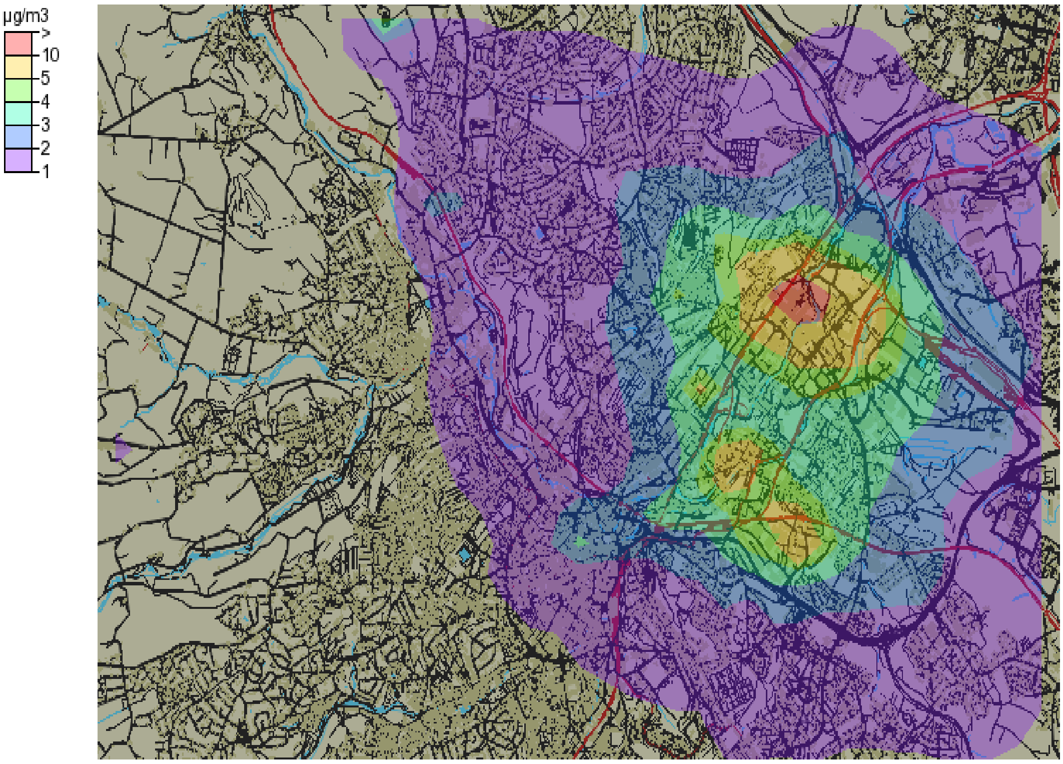


Figure 5.7. Estimated annual mean PM10 concentrations (µg/m^3) using all emission sources in Sheffield City Centre for year 2017.



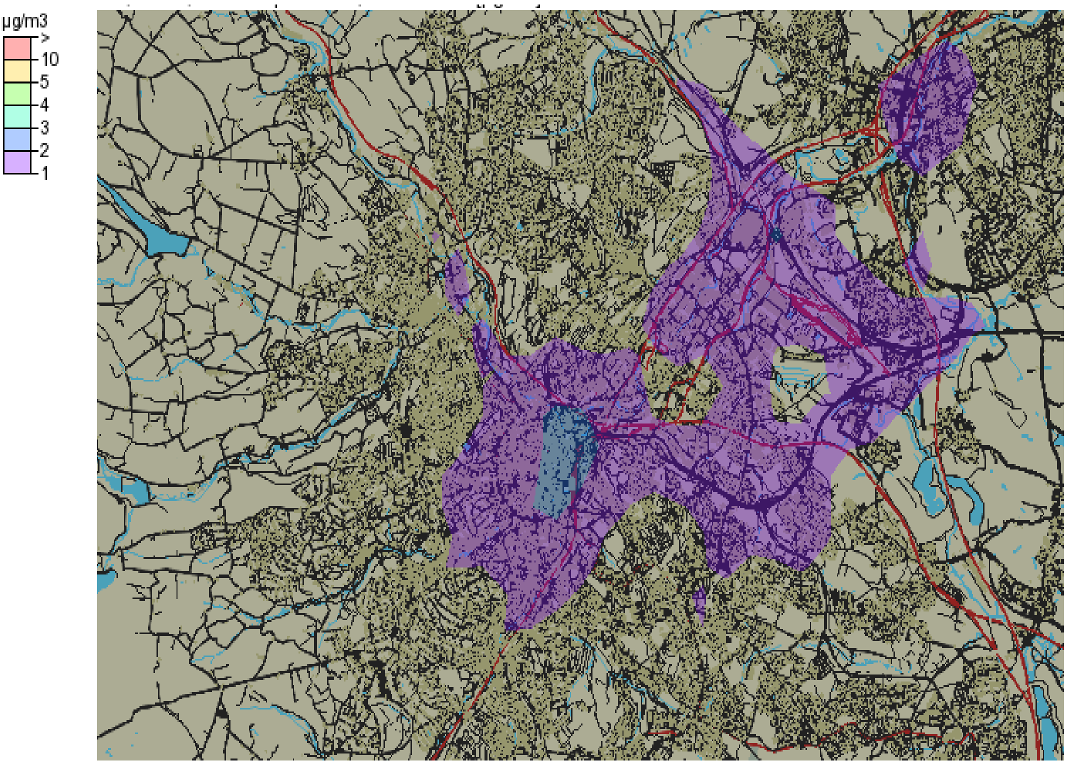


Figure 5.8. Estimated PM_{10} concentrations (µg/m^3) from point sources (upper-panel) and road sources (lower-panel) in Sheffield.

5.3.3. Comparison of predicted and observed concentrations

5.3.3.1. Comparison of seasonal and annual data

Modelled and observed NOx and PM_{10} concentrations cannot be compared in the form of contour maps because observed concentrations are not available in the form of contour maps. To make comparison with observed concentration, both NOx and PM_{10} concentrations were predicted for three receptor points in Sheffield namely Devonshire Green, Sheffield Tinsley and Barnsley Road air quality monitoring stations (AQMS). Details of these sites are provided in Table 5.2. NOx is monitored at all three sites, however PM_{10} is only monitored at Devonshire Green site, and therefore comparison of measured and modelled PM_{10} is made only at Devonshire Green site.

Predicted and measured NOx and PM_{10} concentrations are compared in Table 5.3. Predicted NOx concentrations are higher than the observed concentrations at Devonshire Green and Tinsley and lower at the Barnsley Road AQMS. The encouraging fact is that the model has captured the seasonal trend in NOx concentrations, showing higher levels in winter and lower in summer. Devonshire Green and Tinsley are background sites, whereas Barnsley Road is an urban traffic (roadside) site, therefore lower prediction of NOx as compared to observed concentrations at Barnsley Road site probably indicates that emission inventory for Barnsley Road has under estimated road traffic flow. Observed NOx levels are more than double of the predicted concentrations at Barnsley Road site. However, there is a good positive correlation (r = +0.66) between observed and predicted concentrations. Correlation between observed and modelled NOx at Devonshire Green and Tinsley sites was slightly weaker (r = +0.46 at both sites). A comparison of predicted and observed PM_{10} concentrations at Devonshire Green site showed lower predicted than observed PM_{10} concentrations by a factor of more than two. Furthermore, there was a weak negative correlation (r= -0.18) between the observed and predicted PM_{10} concentrations. However, the data are very limited, only five of each observed and predicted values were available for comparison, therefore to make such comparison meaningful, long term time series observed and predicted NOx and PM_{10} concentrations are required, which are analysed in the next section.

Table 5.3. Comparison of measured and predicted NOx and PM_{10} concentrations (μg/m^3) at Devonshire Green, Tinsley and Barnsley Road AQMS in various seasons in year 2017.

Pollutant	Season 2017	Devonshire Green		Tinsley		Barnsley Rd	
		Observed	Predicted	Observed	Predicted	Observed	Predicted
NOx	Winter	49.2	76.0	69.4	66.3	111.7	47.5
	Summer	23.3	58.3	30.5	50.6	63.6	31.2
	Autumn	27.3	79.6	41.4	73.3	73.8	47.2
	Spring	36.2	70.9	44.2	56.9	86.5	41.2
	Annual	33.9	65.5	46.3	56.4	83.7	38.8
PM_{10}	Winter	15.8	6.98	Not monitored			
	Summer	16.08	5.98				
	Autumn	13.5	7.55				
	Spring	19.0	7.19				
	Annual	16.0	6.63				

5.3.3.2. Comparison of hourly data

To compare hourly observed and predicted data, NOx concentrations were predicted for the months of January and July 2017 for Devonshire Green and Sheffield Tinsley. Comparison was not possible at Barnsley Road AQMS due to missing observed data for both January and July. PM_{10} concentrations are monitored only at Devonshire Green monitoring station, therefore comparison was not possible at the other two sites. To compare predicted and observed concentrations of NOx and PM_{10} both graphical approach and statistical metrics were used.

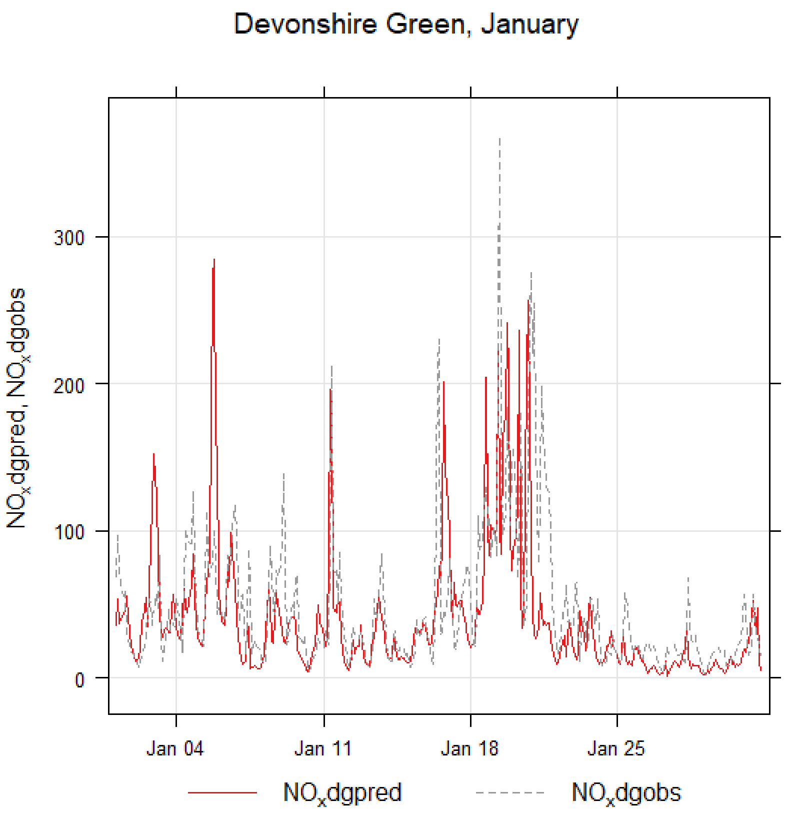
Predicted and monitored NOx concentrations (μg/m^3) are compared for the months of January and July at Devonshire Green (DG) and Sheffield Tinsley (ST) AQMS. January represents winter whereas July represents summer season of the year. The aim is to see how the model prediction varies in winter and summer seasons in comparison to observed data. Figure 5.9 shows the comparison of observed and predicted concentrations in January at Devonshire Green (upper-panel) and at Tinsley (lower-panel). At both Devonshire Green and Tinsley sites the model is slightly under predicting NOx concentrations in January (Figure 5.9). Predicted mean and median concentrations at Devonshire Green in January were 40.29 and 28.20 and observed were 49.04 and 32.17, respectively, whereas at Tinsley predicted values were 51.21 and 27.40 and observed 63.25 and 42.75, respectively (Table 5.4). Negative MB and NMB at both sites also show under prediction (Table 5.5). Metrics showing error of the model (e.g., MB, RMSE and MAE) are slightly greater at Tinsley indicating larger difference between predicted and observed concentrations. However, correlation coefficient value is also higher at Tinsley (0.62) than DG (0.55), showing better linear association between predicted and observed concentrations. The model performance expressed by r (0.65) and FAC2 (0.52) is satisfactory in January at both receptor points.

Summary of predicted and observed NOx concentrations (μg/m^3) for the month of July at both Devonshire Green and Tinsley is shown in Table 5.6. In the month of July predicted mean and median concentrations at Devonshire Green were 35.78 and 32.10, whereas observed concentrations were 19.73

and 16.62, respectively. At the Tinsley site the predicted and observed mean and median concentrations were 38.71, 26.10 and 32.00, 27.48, respectively (Table 5.6). Minimum, mean, median, and maximum values show over prediction of the model in July at both sites. Various statistical metrics calculated for the month of July are shown in Table 5.7, where the values of MB and NMB also show that the model is over predicting NOx concentrations in July. Figure 5.10 graphically compares modelled and observed NOx concentrations, again showing slightly over prediction of NOx concentrations.

Table 5.4. Summary of the observed and predicted NOx concentrations (μg/m^3) in January at Devonshire Green and Tinsley monitoring stations.

Metric	January			
	Devonshire Green		Tinsley	
	NOx_pred	NOx_obs	NOx_pred	NOx_obs
Minimum	0.83	2.73	1.30	3.24
1st Quartile	11.80	18.54	14.15	23.61
Median	28.20	32.17	27.40	42.75
Mean	40.29	49.04	51.21	63.25
3rd Quartile	46.85	61.23	58.05	78.78
Maximum	284.00	367.97	381.00	496.62



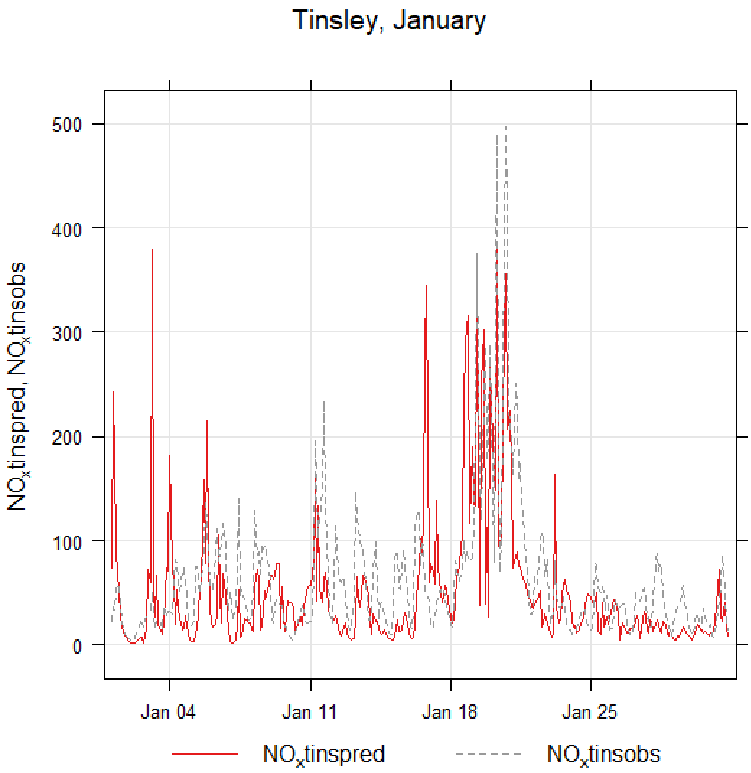


Figure 5.9. Comparison of predicted (pred) and observed (obs) NOx at both Devonshire Green (dg) and Sheffield Tinsley (tins) for the month of January 2017.

Table 5.5. Showing the value of various statistical metrics used for assessing the performance of the model for predicting NOx concentrations in the month of January 2017 at both Devonshire Green and Sheffield Tinsley.

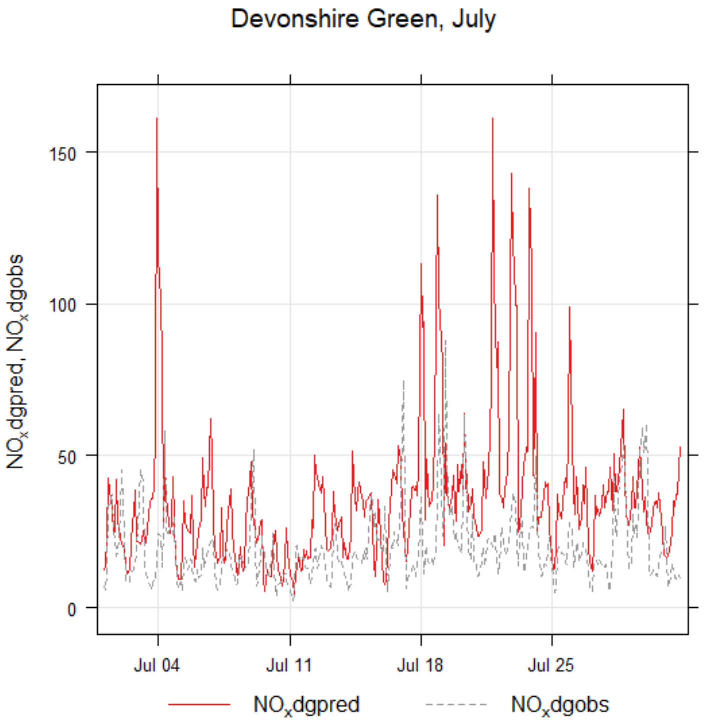
Metric	Devonshire Green	Sheffield Tinsley
FAC2	0.65	0.52
MB	-8.67	-11.96
MAE	25.99	38.18
NMB	-0.18	-0.19
NMAE	0.53	0.60
RMSE	45.62	59.86
r	0.55	0.62
COE	0.22	0.12

Table 5.6. Summary of the observed and predicted NOx concentrations (μg/m^3) in July at Devonshire Green and Tinsley monitoring stations.

Metric	July			
	Devonshire Green		Tinsley	
	NOx_pred	NOx_obs	NOx_pred	NOx_obs
Minimum	3.95	2.10	5.42	4.56
1st Quartile	21.30	12.16	16.50	17.70
Median	32.10	16.62	26.10	27.48
Mean	35.78	19.73	38.71	32.00
3rd Quartile	40.45	23.47	46.05	40.30
Maximum	161.00	87.82	260.00	143.45

Table 5.7. Showing the value of various statistical metrics used for assessing the performance of Airviro model for predicting NOx concentrations in the month of July 2017 at both Devonshire Green and Sheffield Tinsley.

Metric	Devonshire Green	Sheffield Tinsley
FAC2	0.51	0.59
MB	16.05	6.78
MAE	19.40	23.58
NMB	0.81	0.21
NMAE	0.98	0.74
RMSE	28.15	36.36
r	0.32	0.34
COE	-1.24	-0.65



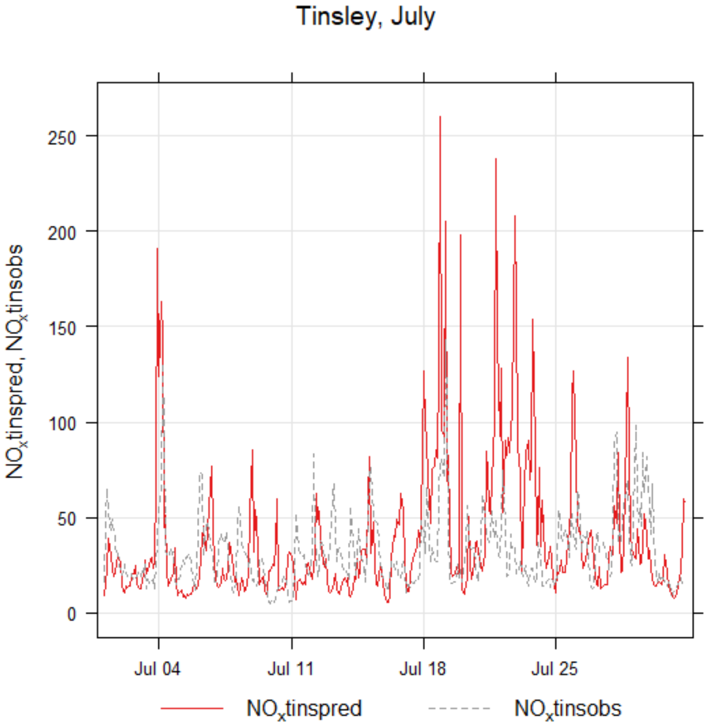
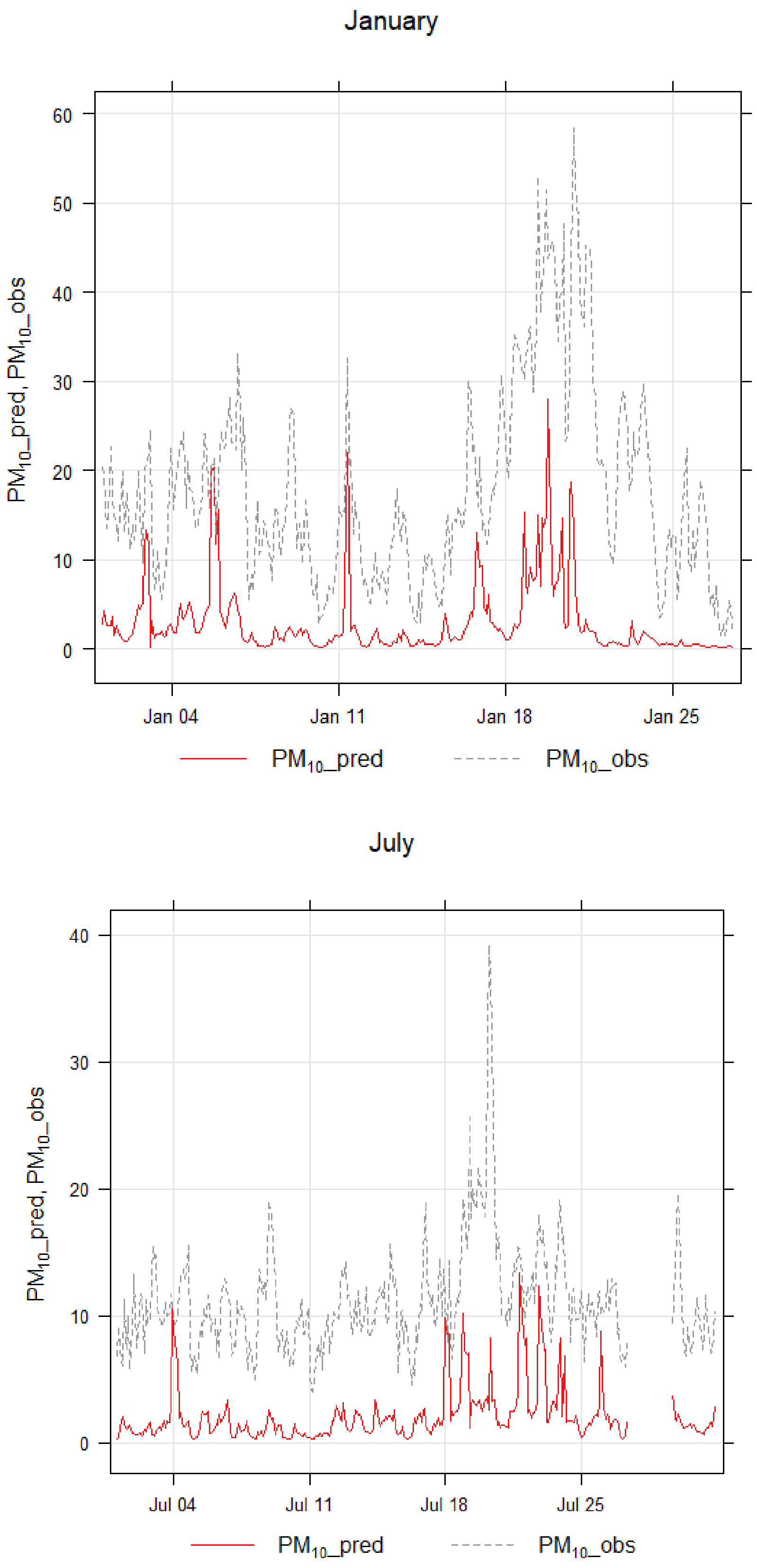


Figure 5.10. Comparison of modelled and monitored NOx at both Devonshire Green (dg) and Sheffield Tinsley (tins) for the month of July 2017.

Predicted and observed PM_{10} concentrations (μg/m^3) at Devonshire Green monitoring station are also compared in the month of January and July for year 2017. Table 5.8 shows summary of the predicted and observed PM_{10} concentrations in both months. In January, the predicted mean and median concentrations were 3.17 and 1.57 and observed mean and median were 17.77 and 15.20, respectively. In July predicted mean and median concentrations were 2.05 and 1.46 and observed mean and median concentrations were 11.12 and 10.35, respectively (Table 5.8), which clearly shows that the model under predicts PM_{10} concentrations in both January and July. This can also be observed in graphical presentations (Figure 5.11) and Table 5.9, which show significant difference in predicted and observed concentrations. FAC2 values are very low in the months of both January (0.05) and July (0.06). Furthermore, NMB value (-0.82) in both January and July shows under prediction by the model. These values show that the model performance is not satisfactory for predicting PM_{10} concentrations at Devonshire Green AQMS. RMSE values for the month of January and July were 17.59 and 9.90, respectively, whereas r values were 0.46 and 0.40 in January and July, respectively. Although the model under predicts PM_{10} concentrations in both months, it captures the trend, therefore, multiplying the modelled concentrations by a constant factor can bring the two time-series (observed and predicted concentrations) close together.

Table 5.8. Summary of the observed and predicted PM_{10} concentrations (μg/m^3) in the month of January and July 2017 at Devonshire Green monitoring stations.

Metric	January		July	
	PM_{10}_obs	**PM_{10}_pred**	**PM_{10}_obs**	**PM_{10}_pred**
Minimum	1.50	0.15	4.00	0.24
1st Quartile	9.83	0.55	8.55	0.812
Median	15.20	1.57	10.35	1.46
Mean	17.77	3.17	11.12	2.05
3rd Quartile	22.80	3.17	12.69	2.38
Maximum	58.35	28.30	39.20	13.30



January

July

Figure 5.11. Comparing measured and predicted PM$_{10}$ concentrations (μg/m^3) at Devonshire Green Sheffield in the month of January (upper-panel) and July 2017 (lower-panel).

Table 5.9. Showing the value of various statistical metrics used for assessing the performance of the model for predicting PM_{10} concentrations in the month of January and July 2017 at Devonshire Green monitoring station.

Metric	Devonshire Green	
	January	July
FAC2	0.05	0.06
MB	-14.60	-9.07
MAE	14.84	9.07
NMB	-0.82	-0.82
NMAE	0.84	0.82
RMSE	17.59	9.90
r	0.46	0.40
COE	-0.74	-2.11

5.3.4. Population exposure to air pollution

Modelled annual NOx concentrations (µg/m^3) in the form of contour map are shown in Figure 5.5, which were obtained using emissions from road traffic, point and area sources employing the Gauss module of Airviro model. The main three hotspots of NOx are shown in the city namely Sheffield City Centre, Darnall and near Tinsley Roundabout on M1 J34S. High levels of NOx are also predicted on Sheffield Parkway (A630, A57) and between Meadowhall Shopping Centre and Sheffield Forgemasters International. Sheffield City Centre is probably the most polluted part of the city, which is mainly due to high level of road traffic.

In this section Figure 5.5 is further analysed and the area of each colour representing a specific level of NOx concentrations is quantified. For this purpose, Figure 5.5 was exported to ArcGIS to calculate area of each colour segment. Red colour which shows NOx levels greater than 65 (µg/m^3) had an area of 2.128 sq. km, whereas yellow colour having NOx levels from 52 to 65 (µg/m^3) had an area of 6.001 sq.km (Table 5.10). To calculate population exposure to NOx pollution, population map (Figure 5.12) was used which was provided by Sheffield City Council. As shown in Table 5.11, 19218 people were estimated to live in the red area and 47517 people were estimated to live in the yellow area. These are the areas mainly in-and-around the city centre of Sheffield. This provides evidence that most of the population exposure to high levels of air pollution in Sheffield is due to NOx pollution which is mainly emitted by road traffic (as discussed above).

Areas with high levels of PM_{10} (Figure 5.7) are mostly outside the city centre, mainly in industrial areas, where population density is low. As discussed above, PM_{10} pollution in Sheffield is mainly caused by point sources, which are mostly located outside the city residential area. Therefore, PM_{10} pollution in Sheffield causes less population exposure compared to NOx pollution.

Table 5.10. Showing area (in km^2) and estimated exposed population to air pollution in Sheffield (according to Figure 5.5 and 5.12).

	Red	Yellow	Green	Blue	Purple	Total
Area (sq.km)	2.128	6.001	20.533	36.269	43.308	108.239
Pop (# residents)	19218	47517	81201	160797	143023	451756
NOx conc. (μg/m^3)	> 65	52 - 65	39 - 51	26 - 38	13 - 25	NA

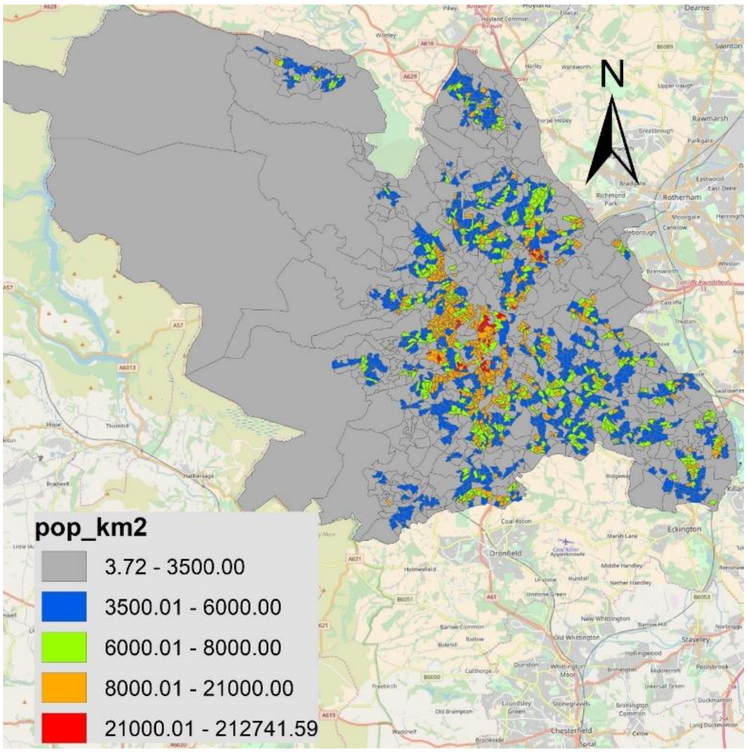


Figure 5.12. Population density (residents/km^2) map of Sheffield, 2016.

5.3.5. Further discussion

Dispersion modelling systems are applied to predict (estimates) pollutant concentrations, which are reflective of emissions, meteorological and topographical data. Dispersion models in the UK are generally divided into: screening, intermediate and advanced models [29]. Advance models are utilised for assessment and review purposes at an urban scale using emissions from point, line and area sources. Among these ADMS-Urban, AERMOD and Airviro modelling system are frequently used throughout the world by local authorities, consultants and researchers.

There is always a degree of uncertainty in the model outputs that is why the predicted concentrations are either lower or higher than the measured concentrations. These uncertainties are mainly due to two reasons [30]: (a) model inputs, including emission inventory, meteorological parameters, parameterisation of boundary layer and stability classes; (b) the model itself. Svensson [30] predicted NOx concentrations in Stockholm, Sweden using Airviro model and compared the results with measured concentrations, reporting that the model under-predicted NOx concentrations, especially in winter season, which is in agreement with the current study. In this paper PM_{10} are significantly under-predicted, which seems to be mainly related with shortcomings in emission inventory. There are several possible sources of uncertainties in emission inventory that may cause the difference between modelled and measured PM_{10} levels including [31]: (a) Emission from diffuse sources, like emissions from coke ovens, metal processing and construction are difficult to be measured satisfactorily as their levels are variable in both time and space. (b) Combustion related emissions are also subject to high uncertainty, especially in cases where PM emissions are very low and difficult to measure (e.g., from gas combustion or emissions from vehicles with a diesel particulate filter). (c) Emissions of PM from non-exhaust traffic sources, such as tyre and brake wear and road abrasion are particularly uncertain. (d) Coarse particles from resuspended soils and road dusts are subject to considerable uncertainties. Furthermore, the emission inventory used in this study does not take into account secondary aerosols, which are formed in the atmosphere. The formation of secondary particles is dependent on the precursor's emissions, e.g., SO_2 and NOx that lead to the formation of secondary particles like nitrate (NO_3^-) and sulphate (SO_4^{2-}) and meteorological conditions like temperature and relative humidity. Secondary particles can contribute significantly into the observed concentrations of particles, especially fine particles ($PM_{2.5}$). Air Quality Expert Group (AQEG) [31] has shown that a significant amount of secondary particles made of nitrate and sulphate adding to the background concentrations is transferred from other large cities in the UK and Europe. Another possible reason for under-predicting PM_{10} is ignoring of small streets in emission inventory, which can directly contribute to the observed concentrations. According to AQEG the estimated uncertainty in total UK emissions is estimated to be between –20% and +30%.

In a recent study Dedele et al. [32] modelled NO_2 concentrations with Airviro in Kaunas, which is the second-largest city in Lithuania. They measured the levels of NO_2 using NO_2 diffusion tubes in 5 streets with different traffic and building characteristics for a two-week period in each season. Measured NO_2 concentration was higher in winter and autumn, and lower in spring and summer seasons than the modelled concentrations. The difference between modelled and measured concentrations was greatest in winter, which was reported to be due to domestic heating in winter that was not accounted for in the model. Dedele et al. [32] reported that because the street canyon model did not take into account emissions from the other emission sources, it resulted in lower estimated values than measured values. Mukharjee et al. [26] also used Airviro to model the levels of NOx, SO_2 and CO in Singapore. They concluded that although road traffic contributed 24% NOx emissions in the city, the exposure caused was 40 % due to the fact that the pollutants were emitted at the ground-levels within the breathing zone. Leksomono et al. [29] also reported that industrial sources produced relatively smaller contribution to ground level NO_2 concentrations per unit of emission. This is because emissions from industrial sources are released at heights well above the ground and therefore subject to more dilution. According to the findings of Mukharjee et al. [26] the predicted and the measured hourly CO concentrations agreed to an accuracy of approximately 19 % with R^2 value of 0.67. The model also captured the changes in the meteorological characteristics. The Airviro model over-predicted the measured NOx concentrations significantly, which was believed to be due to the constraint that the model did not take into account the photochemical transformation of NOx and ozone. Gidhagen et al. [33] employed Airviro model to assess the impact of residential wood combustion on exposure to $PM_{2.5}$ and its health impacts in three urbanised areas in Sweden. Gidhagen et al. [33] estimated that annual mortality due to modelled $PM_{2.5}$ concentrations from residential wood combustion was approximately four people (4 persons/year), corresponding roughly to 0.4 % of the total number of deaths in the region. Leksmono et al. [29] have reported that distance of the site where meteorological data are collected from the area where pollution is to be modelled is important for assessing the levels of a pollutant, especially for modelling short-term concentrations.

The above discussion and the finding of this study indicates that dispersion modelling systems are important tools for air quality management in urban areas, however, care should be taken to minimise the sources of error, which might include: (a) selection of appropriate model, (b) appropriateness of the emission inventory for the purpose, (c) availability of meteorological data and the distance of the meteorological monitoring site from the site of interest, (d) background contribution (both urban and regional), (e) photochemical transformation of pollutants, (f) complexity of the terrain.

5.4. Conclusions

Main aims of the study were: (a) To determine the most significant emission sources of NOx and PM_{10} in Sheffield; (b) To analyse spatiotemporal variability of NOx and PM_{10} in Sheffield and surrounding areas, highlighting the hotspots of air pollution and discussing the main reasons; (c) To assess the performance of Airvrio air quality model for NOx and PM_{10} prediction in different geographical locations in Sheffield during different time of the years.

In this paper NOx and PM_{10} pollutant emissions are modelled and their spatial distribution is analysed employing the Airviro air quality dispersion modelling system. Air pollutant emissions from road traffic, point sources and area sources in Sheffield are modelled for year 2017. Spatial variability of NOx and PM_{10} concentrations is presented in the form of contour maps. Furthermore, NOx and PM_{10} concentrations are predicted for three receptor points. Airviro outputs showed three locations with high NOx concentrations namely Sheffield City Centre, Darnall and near Tinsley Roundabout on M1 J34S. High levels of NOx were also predicted on Sheffield Parkway and between Meadowhall shopping centre and Sheffield Forgemasters International. High PM_{10} concentrations were estimated mainly between Sheffield Forgemasters and Meadowhall, near Sheffield Parkway and Attercliffe. Several emission scenarios were tested for both NOx and PM_{10} which showed that high levels of NOx were mainly linked to road traffic, whereas those of PM_{10} seemed to be linked with point sources. As expected, estimated levels of pollutants were higher in colder season (e.g., winter) than in warmer season (e.g., summer). In case of PM_{10}, predicted concentrations were significantly lower than the observed concentrations at the Devonshire Green monitoring station in both January and July, however, the model successfully captured temporal trends. Furthermore, modelled NOx concentrations showed better association with observed concentrations in terms of both pollutant levels and trends. Also, modelled NOx concentrations were slightly lower in January and higher in July than measured concentrations. Spatial analysis showed that more people were exposed to NOx concentrations mostly emitted by road traffic in the city centre and surrounding areas than to PM_{10} mostly emitted by point sources in Sheffield.

The main outcomes of this study can be summarized as follows: (1) NOx concentrations in Sheffield are mainly from road traffic related emission sources, whereas PM_{10} concentrations are from point sources, e.g., various types of industries such as steel industry. (2) More people are exposed to NOx pollution mainly emitted by road traffics in the city centre. (3) There are three hotspots of NOx pollution in Sheffield namely the Sheffield City Centre, Darnall and near the Tinsley Roundabout (M1 J34S), whereas the high PM_{10} concentrations were shown mainly between Sheffield Forgemasters International and Meadowhall shopping centre. (4) Relatively higher average levels of NOx and PM_{10} were predicted in winter and autumn than in summer and spring compared to measured concentrations. (5) NOx predictions by Airviro were lower in January and higher in July than measured NOx concentrations at both Devonshire Green and Sheffield Tinsley. However, the model under predicted PM_{10} concentrations in both January and July at Devonshire Green site. The difference between measured and predicted PM_{10} concentrations was considerably greater compared to NOx. (6) In Sheffield nearly 19000 people live in areas with NOx levels greater than 65 $\mu g/m^3$ and 48000 people live in areas with NOx levels 52 – 65 $\mu g/m^3$, which together are approximately 15 % of Sheffield population.

Models validated by observations can be used to fill-in spatiotemporal gaps in measured air quality data. Furthermore, dispersion models are important tools for urban air quality management, however, steps should be taken to minimise potential errors in emission data, meteorological data and complexity of the terrain. Particulates generated from vehicle wear and tear, resuspension of dust particles and emission from natural sources require special attention to improve model performance. In addition, further work is required to quantify people exposure to air pollution in Sheffield using dense network of static sensors and personal monitors. People can be exposed to air pollution in their houses (residents), work places and when commuting to-and-from work using various means of transports, e.g., buses, trains, trams, cars, cycles or walking. How exposure levels vary using various transport modes needs to be quantified in Sheffield.

Acknowledgement: We would like to express our thanks to the EPSRC (grant number - EP/R512175/1) and Siemens plc for funding this project, which is part of PhD-work of the first author. We are also thankful to Apertum for freely providing the Airviro model for this study (https://www.airviro.com/airviro).

Author Contributions: Said Munir, Martin Mayfield and Daniel Coca planned and designed the original idea. Said Munir and Ogo Osammor obtained the Airviro model and undertook the modelling. Martin Mayfield and Daniel Coca won the funding for this research project. Said Munir wrote the first draft of the manuscript under the supervision of Martin Mayfield and Daniel Coca. Lyudmila S Mihaylova, Ogo Osammor, Martin Mayfield and Daniel Coca reviewed and edited the manuscript. All the authors have read and approved the content of the manuscript.

Conflicts of Interest: All the authors declare no conflict of interest.

5.5. References

1. DEFRA, 2015. Improving air quality in the UK Tackling nitrogen dioxide in our towns and cities, UK overview document, December 2015. Available online: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/486636/aq-plan-2015-overview-document.pdf (accessed 09/06/2019).
2. WHO, 2013. Health effects of particulate matter, policy implications for countries in eastern Europe, Caucasus and central Asia. Publications of WHO Regional Office for Europe UN City, Marmorvej 51 DK-2100 Copenhagen, Denmark.
3. Landrigan, P.J., 2016. Air pollution and health. The Lancet Public Health, 2, (1): 4 –5. DOI: https://doi.org/10.1016/S2468-2667(16)30023-8.
4. Daly, A., Zannetti, P., 2007. Air Pollution Modeling – An Overview. Chapter 2 of ambient air pollution (Zannetti, P., Al-Ajmi, D., Al-Rashied, S., Editors). Published by the Arab School for Science and Technology and The EnviroComp Institute. Available online: http://home.iitk.ac.in/~anubha/Modeling.pdf (accessed 18/08/2019).
5. Aldrin, M., Haff, I.H., 2005. Generalised additive modelling of air pollution, traffic volume and meteorology. Atmospheric Environment, 39, 2145–2155.
6. Andersen, S. B., Weatherhead, E. C., Stevermer, A., Austin, J., Brühl, C., Fleming, E. L., de Grandpré, J., Grewe, V., Isaksen, I., Pitari, G., Portmann, R. W., Rognerud, B., Rosenfield, J. E., Smyshlyaev, S., Nagashima, T., Velders, G. J. M., Weisenstein, D. K., Xia, J., 2006. Comparison of recent modelled and observed trends in total column ozone. Journal of Geophysical Research, 111. D02303. doi:10.1029/2005JD006091.
7. Arnold, S.R., M.P. Chipperfield, M.A. Blitz, 2005. A three-dimensional model study of the effect of new temperature-dependent quantum yields for acetone photolysis. Geophysics Research, 110, (D22), D22305. doi:10.1029/2005JD005998.
8. Baur, D., Saisana, M., Schulze, N., 2004. Modelling the effects of meteorological variables on ozone concentration-a quantile regression approach. Atmospheric Environment, 38, (28): 4689 – 4699.
9. Berastegi, G.I., Madariaga, I., Elias, A., Agirre, E., Uria, J., 2001. Long term changes of O_3 and traffic in Bilbao. Atmospheric Environment, 35, (2001): 5581 – 5592.
10. Brasseur, G. P., Hauglustaine, D. A., Walters, S., Rasch, P. J., Muller, J. F., Granier, C., Tie, X. X., 1998. MOZART, a global chemical transport model for ozone and related chemical tracers, 1: Model description. Journal of Geophysical Research, 103, 28265–28289.

11. Munir, S., Chen, H., Ropkins, K., 2012. Modelling the impact of road traffic on ground level ozone concentration using a quantile regression approach. Atmospheric Environment, 60, 283–291.
12. Westmoreland, E.M., Carslaw, N., Carslaw, D.C., Gillah, A. and Bates, E., 2007. Analysis of air quality within a street canyon using statistical and dispersion modelling techniques. Atmospheric Environment, 41: 9195–9205.
13. Wilkening, I.H., Baraldi, D., 2007. CFD modelling of accidental hydrogen release from pipelines. International Journal of Hydrogen Energy, 32 (13): 2206-2215.
14. James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. *An introduction to statistical learning: with applications in R.* Springer Texts in Statistics. DOI: 10.1007/978-1-4614-7138-7 7.
15. Salmond, J.A., Clarke, A.G., Tomlin, A.S., 2006. The atmosphere, chapter 2, pp: 8-76. In Harrison, R.M., an introduction to pollution science. The Royal Society of Chemistry.
16. El-Harbawi, M., 2013. Air quality modelling, simulation, and computational methods: a review. Environmental Reviews, 21, (3): 149-179. DOI: https://doi.org/10.1139/er-2012-0056.
17. Modi, M., Ramachandra, V. P., Ahmed, L.S.K., Hussain, Z., 2013. A review on theoretical air pollutants dispersion models. International Journal of Pharmaceutical, Chemical and Biological Sciences, 3, (4): 1224-1230.
18. Airviro User's Reference, 2013. Working with the Dispersion Module - How to simulate the dispersion of pollutants. Swedish Meteorological and Hydrological Institute, SE-601 76 Norrkoping, Sweden. Available online: (https://www.airviro.com/airviro/extras/pdffiles/UserRef_Volume2_Dispersion_v3.23.pdf) (accessed 17/01/2019).
19. Airviro specification, 2015. Airviro Specification v4.00 - Part I: Functions in Airviro. Apertum IT AB, Teknikringen 7- 583- 30. Linköping, Sweden. Available online: https://www.airviro.com/airviro/extras/pdffiles/Specification1_v4.00.pdf (Accessed 17/01/2019).
20. Pasquill, F., 1961. The estimation of the dispersion of windborne material. Meteorological Magazine, 90, 33 – 49.
21. Pasquill, F., 1962. Some observed properties of medium-scale diffusion in atmosphere. Quarterly Journal of Royal Meteorological Society, 88, (375): 70 – 79. doi:10.1002/qj.49708837507.
22. Pasquill, F., 1974. Atmospheric Diffusion. 2nd ed. Published by Horwood, Chichester [Eng.], New York (1974). ISBN 10: 0853120153.
23. Briggs, G.A., 1965. A plume rise model compared with observations. Journal of the Air Pollution Control Association 15, (9): 433 – 438.
24. National Atmospheric Emissions Inventory, 2018. Air pollutant inventories for England, Scotland, Wales, and Northern Ireland, 1990-2016. Available online: https://uk-air.defra.gov.uk/assets/documents/reports/empire/naei/annreport/annrep99/chap1_2.html (accessed:08/07/2019).
25. Airviro User´s Reference, 2018. Working with the Emission Data Base (EDB): How to construct a dynamic emission database and simulate emission scenarios, version 4.00. Available online: https://www.airviro.com/airviro/extras/pdffiles/UserRef_Volume1_EDB_v4.00.pdf (accessed: 20/11/2019).
26. Mukherjee, P., Viswanathan, S., Choon, L.C., 2000. Modeling mobile source emissions in presence of stationary sources. Journal of Hazardous Materials, A76 (2000): 23–37.
27. Carslaw, D., 2011. Defra regional and transboundary model evaluation analysis – Phase 1. King's College London. Version: 15th April 2011. Available online: (https://uk-air.defra.gov.uk/assets/documents/reports/cat20/1105091514_RegionalFinal.pdf (accessed 18/08/2019).
28. Sayegh, A.S., Munir, S., Habeebullah, T.M., 2014. Comparing the Performance of Statistical Models for Predicting PM_{10} Concentrations. Aerosol and Air Quality Research, 14, (3): 653-665.
29. Leksmono, N.S., Longhurst, J.W.S., Ling, K.A., Chatterton, T.J., Fisher, B.E.A., Irwin, J.G., 2004. Assessment of the relationship between industrial and traffic sources contributing to air quality objective exceedences: a theoretical modelling exercise. Environmental Modelling & Software, 21 (2006): 494–500.
30. Svensson, N., 2013. Evaluation of atmospheric dispersion models: Comparison with measurements in Stockholm. Master degree project in meteorology, Stockholm University, Sweden.
31. AQEG, 2012. Fine Particulate Matter (PM2.5) in the UK: https://uk-air.defra.gov.uk/assets/documents/reports/cat11/1212141150_AQEG_Fine_Particulate_Matter_in_the_UK.pdf (accessed: 21/11/2019).

32. Dedele, A., Miskinyte, A., Cesnakaite, I., 2019. Comparison of Measured and Modelled Traffic-Related Air Pollution in Urban Street Canyons. Polish Journal of Environmental Studies, 28 (5): 3115-3123.
33. Gidhagen, L., Bennet, C., Segersson, D., Omstedt, D., 2015. Exposure Modeling of Traffic and Wood Combustion Emissions in Northern Sweden - Application of the Airviro Air Quality Management System. IFIP Advances in Information and Communication Technology, 448, 242-251.1

CHAPTER 6: A NONLINEAR LAND-USE REGRESSION APPROACH FOR MODELLING NO_2 CONCENTRATIONS IN URBAN AREAS – USING DATA FROM LOW-COST SENSORS AND DIFFUSION TUBES

Said Munir [1,*], Martin Mayfield[1], Daniel Coca[2] and Lyudmila S Mihaylova[2]

[1] Department of Civil and Structural Engineering, the University of Sheffield, Sheffield, UK, S1 3JD; martin.mayfield@sheffield.ac.uk (M.M.)
[2] Department of Automatic Control and Systems Engineering, the University of Sheffield, Sheffield, UK, S1 3JD; d.coca@sheffield.ac.uk (D.C.); l.s.mihaylova@sheffield.ac.uk (L.S.M.)
* Correspondence: smunir2@sheffield.ac.uk (S.M.); Mob: +447986001328

Published: Atmosphere 2020, 11, 736
https://doi.org/10.3390/atmos11070736
Received: 8 June 2020; Accepted: 9 July 2020; Published: 11 July 2020.

Abstract

Land Use Regression (LUR) based on multiple linear regression model is one of the techniques used most frequently for modelling the spatial variability of air pollution and assessing exposure in urban areas. In this paper, a nonlinear generalised additive model is proposed for LUR and its performance is compared to a linear model in Sheffield, UK for the year 2019. Pollution models were estimated using NO_2 measurements obtained from 188 diffusion tubes and 40 low-cost sensors. Performance of the models was assessed by calculating several statistical metrics including correlation coefficient (R) and root mean square error (RMSE). High resolution (100m x 100m) maps demonstrated higher levels of NO_2 in the city centre, eastern side of the city and on major roads. The results showed that the nonlinear model outperformed the linear counterpart and that the model estimated using NO_2 data from diffusion tubes outperformed the models using data from low-cost sensors or both low-cost sensors and diffusion tubes. The proposed method provides a basis for further application of advanced nonlinear modelling approaches to constructing LUR models in urban areas which enable quantifying small scale variability in pollution levels.

Keywords: air quality modelling; land-use regression; nonlinear regression; Sheffield; spatial analysis; low-cost sensors; nitrogen dioxide.

6.1. Introduction

Air pollution is one of the serious environmental issues that affects human health and may cause mortality. Landrigan [1] reported that air pollution caused 6·4 million deaths worldwide in 2015, which shows the significance of the impact of poor air quality on human health. Several air pollutants are shown to have negative impacts on human health, however, among gaseous pollutants nitrogen dioxide (NO_2) is considered the most serious pollutant causing both chronic and acute respiratory diseases including asthma, hospital admission and mortality [2]. NO_2 is considered the most serious gaseous pollutant in urban areas and many air quality management areas (AQMAs) in the UK are based on the exceedances of NO_2 [3]. Therefore, it is important to carry out different monitoring and modelling investigations to analyse its spatial variability, especially micro-level variability. This could be done by developing high resolution maps in urban areas that help in understanding the main drivers of NO_2 levels and quantifying exposure to elevated NO_2 concentrations in urban areas. It is not feasible to capture micro-level spatial variability in NO_2 concentration with the help of monitoring network in a large urban area as this requires a huge number of sensors. Alternatively, a certain number of sensors can be installed, and the concentrations can be extrapolated to other area using modelling techniques. This is exactly what this study intends to do.

One of the challenges of air pollution in urban areas is the quantification of small-scale local level exposure by analysing spatial variability in pollutant concentrations. The small scale variability in urban areas is controlled by emission sources, land use features, building density-and-height and geographical characteristics. To characterise spatial variability in pollutant concentrations, three approaches are most commonly used: GIS based interpolation methods, Dispersion models, and LUR models.

Hsu et al., [4] reviewed different methods used for interpolation and divided them into two categories: Geostatistical techniques and Non-geostatistical methods. The geostatistical techniques include Ordinary Kriging, Universal Kriging, Simple Kriging, Empirical Bayesian Kriging and Original CoKriging. The non-geostatistical techniques include Splines, Trend Surface Analysis, Inverse Distance Weighting and Natural Neighbour. According to the findings of Hsu et al. [30] the geostatistical interpolations showed better prediction than the non-geostatistical interpolation techniques. Interpolation methods are used to interpolate air pollutant concentrations between various air monitoring stations to provide better spatial coverage. However, when these approaches are applied at a local scale such as intra-city scale, these methodologies are known to produce considerable variations in air pollutant concentrations within a small area and are more effective at large scales, such as national or regional scale [5]. Therefore, at urban scale interpolation approaches should not be the priority for analysing spatial variability of air pollution.

Dispersion models are probably the most advanced modelling techniques for determining the spatial variability of air pollutants in urban areas. Dispersion models combine the data of pollutant emission from different sources (e.g., point, line and area sources), the geophysical characteristics of the study area and meteorological parameters. Dispersion models have the potential to incorporate both temporal and spatial variations to replace the need for air pollution monitoring. However, high cost of purchase and maintenance, high input demands, and requirement of skilled staff limit their application.

LUR models provide an effective alternative to GIS interpolation and dispersion modelling techniques on urban scale. LUR is a spatial modelling approach used most frequently for analysing the spatial variability and quantifying public exposure to air pollution in urban areas. Different modelling, mapping and data fusion techniques are available for modelling spatial variability of air pollution in urban areas [6]. Several authors have preferred Land Use Regression (LUR) over other approaches. Briggs et al. [7] compared the performance of LUR with spatial interpolation methods (e.g., kriging, TIN-contouring and trend surface Analysis) and reported that LUR performed much better than the interpolation techniques. The reason was that in urban areas spatial variability of air pollutants is more controlled by the local emission sources and geographical characteristics, rather than a smoothly varying field which is assumed by the interpolation methods. Hoek et al. [8] reviewed several LUR models and reported that the performance of the LUR model in urban areas was either equivalent or better than dispersion and geostatistical approaches. LUR is a widely employed approach for air pollution exposure estimation using Geographical Information Systems (GIS) and statistical analyses to determine the association between geographic features and measured atmospheric pollutant concentrations [9]. LUR is easy to implement and can provide an effective alternative to geostatistical and dispersion modelling techniques on urban scale. LUR approach was introduced by Briggs et al. [9] and since then has been used in numerous studies around the world [8, 10-16]. LUR models associate pollutant concentrations, such as NO_2 to site specific geographical characteristics, e.g., topography, land use, traffic, population density, altitude and meteorological parameters. The use of these variables in the regression model are known to capture small scale variability on city scale [17]. Recent development in GIS technology has also added to the popularity of LUR approaches.

Rahman et al. [13] developed a LUR model in Brisbane, Australia to predict NO_2 and NOx during 2009 – 2012. The model was able to explain 64 % and 70 % variations in NO_2 and NOx, respectively.

Distance to major roads and industrial areas were the common predictor variables for both NO_2 and NOx, suggesting an important role of road traffic and industrial emissions. Rahman et al. [13] used the following independent variables in their model: distance to coast (km), distance to port (km), distance to airport (km), distance to nearest major road (km), distance to nearest minor road (km), major road length (km), minor road length (km), population density (person/km^2), land use by type (km^2), and elevation (m). Muttoo et al. [12] used several geographic predictor variables to predict NOx levels in Durban, South Africa employing an LUR model. They used length of minor roads within a 1000 m radius, length of major roads within a 300 m radius, and area of open space within a 1000 m radius in the model as independent variables. The LUR model was able to explain 73% variance in NOx concentrations, however cross validation resulted in R^2 value of 0.59. Hoek et al. [8] have provided a detailed review on LUR models, identifying 25 studies on the subject. They have identified several significant predictors for LUR models, including various traffic characteristics, population characteristics, land use, physical geography, and climatic conditions. Gillespie et al. [18] developed an LUR model to estimate exposure to NO_2 in Glasgow, Scotland and reported that the use of more than 60 training sites had a considerable beneficial effect on model performance. Mostly, LUR approaches are applied typically at a city scale e.g. [17, 19], however, some researchers have applied LUR models to entire countries. Beelen et al. [15] and Stedman et al. [14] applied LUR models in Netherland and UK, respectively, whereas Vienneau et al. [20] developed an LUR model for both UK and Netherland and compared its outputs in both countries.

Sheffield City Council (SCC) has declared most of the urban area in Sheffield City as AQMA due to the high levels of NO_2. Detailed modelling investigations are required to model NO_2 concentrations and to create high resolution maps in the city for quantifying public exposure to NO_2 and determining small scales NO_2 spatial variability in the city. In this paper LUR models are developed to predict NO_2 concentrations and make high resolution maps (100m x 100m) using NO_2 data measured by a network of diffusion tubes (DT) and low-cost sensors (LCS). The literature reviewed above all have employed multiple linear regression models (MLRM) for developing LUR models, which work on the assumption that response and predictor variables have linear association. Here an advanced nonlinear approach Generalised Additive Model (GAM) is proposed, which is more suitable for air quality data analysis. The rest of the paper is structured as follows: Methodology of this paper is presented in section 6.2, wherein section 6.2.1 describes monitoring sites and predictor variables; section 6.2.2 describes LUR model development; section 6.2.3 describes model specifications; section 6.2.4 describes model validation; and section 6.2.5 describes statistical software used in this study. Results and discussion are presented in section 6.3 and the main outcomes of this work are summarised in section 6.4.

6.2. Methodology

In this study three LUR models are developed for modelling the spatial variability of NO_2 concentrations and producing high resolution maps (100m x 100m) in Sheffield for the year 2019. Sheffield (53°23′N, 1°28′W) is a historical metropolitan borough in South Yorkshire, United Kingdom and has emerged as a green and modern cityscape in the proximity of the Peak District National Park. According to 2011 census Sheffield City had a population of 552,700, however, since then the population has grown and according to more recent estimates has reached about 700,000. Among gaseous pollutants, NO_2 is the pollutant of concern in Sheffield and most of the AQMA in Sheffield is declare due to the elevated levels of NO_2 concentrations mostly emitted by road traffic [21].

6.2.1. Predictor variables and NO_2 monitoring sites

In this paper NO_2 concentration (μg/m^3) is modelled using data from DT and LCS. There were 188 DT and 40 LCS measuring NO_2 concentrations (μg/m^3) in Sheffield. The locations of DT and LCS are shown in Figure 6.1. There are two types of LCS: 13 AQMesh pods and 27 Envirowatch E-MOTEs. For a brief description of these sensors see Munir et al. [21, 22]. DT and LCS have relatively high uncertainty

(25 to 30 %) as compared to reference sensors (15%). Generally, DT are exposed for a period of 2-4 weeks (no longer than 5 weeks and no shorter than 1 week). After this period, the old DT are replaced with new ones. In this way, the monitoring is carried out for the whole year to get annual average. LCS (both AQMesh and Envirowatch E-MOTEs) were installed around the city (Figure 6.1), providing high resolution temporal data (e.g., 5 minutes to hourly), which were converted to annual average. A summary of NO_2 concentrations measured by DT and LCS for year 2019 is provided in Table 6.1.

Table 6.1. Descriptive statistics of NO_2 measured by DT and LCS in Sheffield.

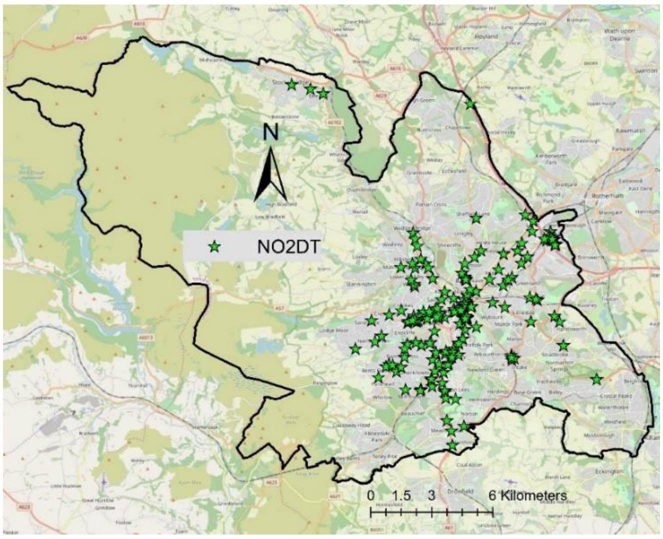
Metrics	DT NO_2	LCS NO_2
Minimum	13.58	12.74
1st Quartile (25th percentile)	28.29	25.23
Median	33.77	33.19
Mean	34.23	41.23
3rd Quartile (75th percentile)	40.00	45.28
Maximum	62.03	146.54
Standard Deviation	9.65	27.69

To model NO_2 concentration, data of different land use, traffic and population variables were collected to be used as predictor (independent) variables. Maps of the predictor variables were downloaded from the ordinance survey UK and provided by the Sheffield City Council. ArcGIS version 10.7.1 and its LUR tools were used to extract different values within 100 m x 100 m grid. These variables are given below:

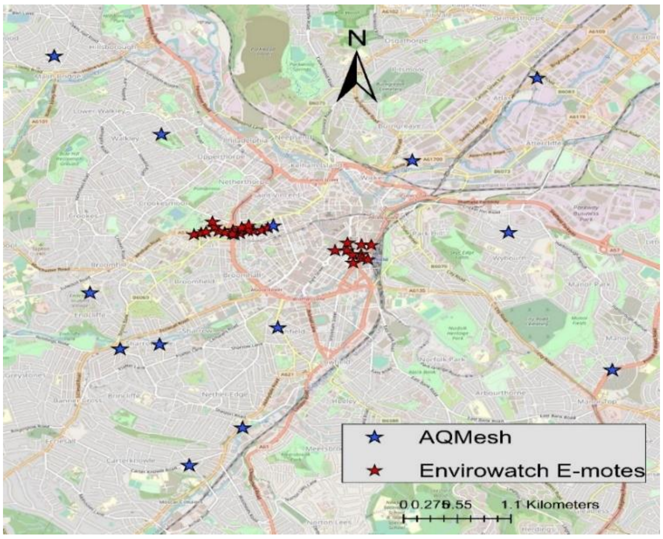
(a) Area (m^2) of industrial land use, residential area, commercial area, parks and green area, and building area;
(b) Length (m) of motorways, major roads, and minor roads;
(c) Distance (m) to motorway, major road, minor road, building, industry, bus stop, parks, commercial area, and residential area;
(d) Population (persons per km^2), Altitude (m), number of bus stops, easting (m), northing (m), and street intersection.

As the impact distance varies among different variables, buffers of multiple radii (10, 50, 100, 200, 300 500, and 1000 m) were created for industrial area, commercial area, park and green area, residential area, major roads, minor roads, motorways, and bus stops. Figure 6.2 shows the spatial distribution of different predictor variables in the city.

(a)


N
NO2DT
0 1.5 3 6 Kilometers

(b)


N
AQMesh
Envirowatch E-motes
0 0.275 0.55 1.1 Kilometers

(c)

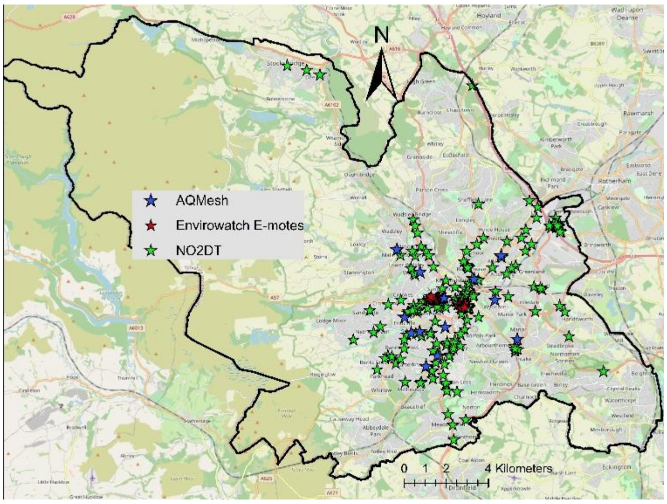
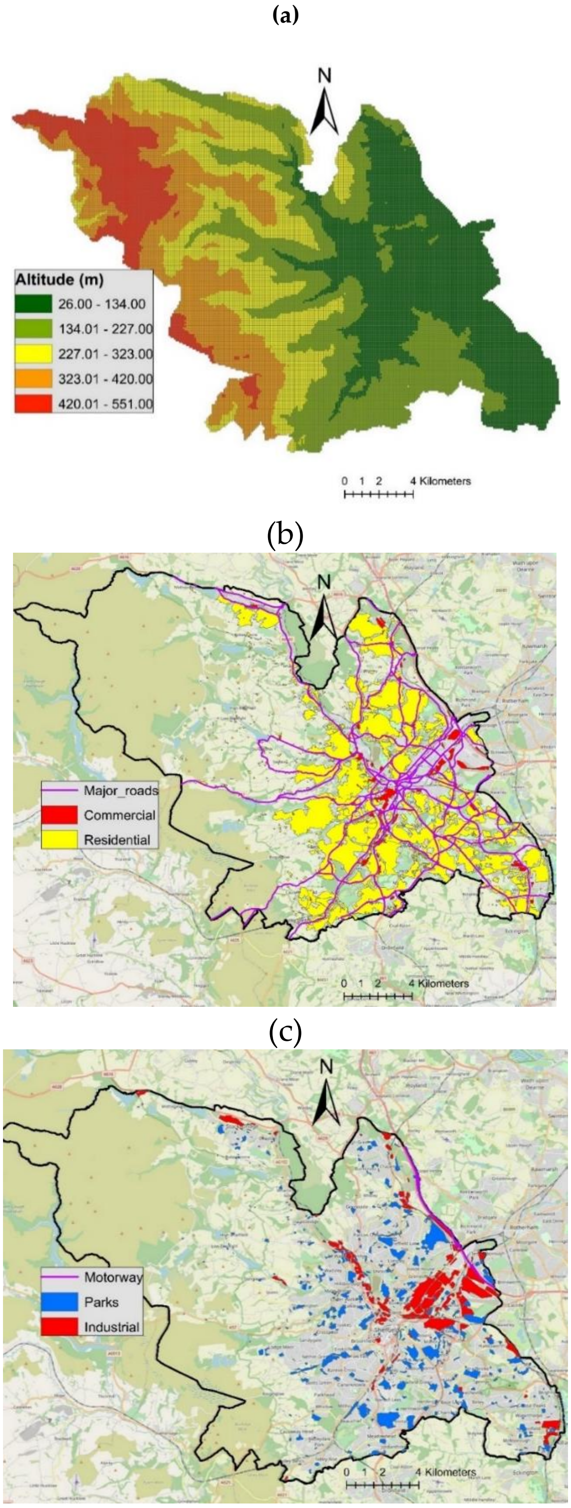

N
AQMesh
Envirowatch E-motes
NO2DT
0 1 2 4 Kilometers

Figure 6.1. Showing the locations of 188 NO_2 diffusion tubes (DT) (a), 40 low-cost sensors (LCS): 13 AQMesh pods and 27 Envirowatch E-motes (b), and both LCS and DT (c) in the City of Sheffield.



(a)

(b)

(c)

Figure 6.2. Altitude (a), major roads, commercial and residential area (b) and motorway, green parks and industrial area (c) in Sheffield.

6.2.2. LUR model development

In this paper LUR models were developed for modelling NO_2 concentrations in Sheffield. NO_2 concentration was regressed against the predictor variables. Both linear and non-linear LUR models were developed using three NO_2 datasets:

(1) Measurements of NO_2 obtained from 188 DT;
(2) Measurements of NO_2 obtained from 40 LCS;
(3) Combined NO_2 measurements obtained from both LCS and DT (228).

It should be remembered that in each case 75 % randomly selected data were used for model training (fitting) and 25 % data were hold-out for model testing (cross validation). The novelty of this study is that in addition to a Multiple Linear Regression Model (MLRM), a nonlinear Generalised Additive Model (GAM) is proposed for developing LUR model. The performance of MLRM and GAM is compared. Secondly, in addition to DT measurements, NO_2 data from a network of 40 LCS are also used in this paper. It should be noted that both MLRM and GAM were developed and validated in R programming language [23]. However, to extrapolate predicted NO_2 concentration to the entire Sheffield city to produce continuous heat maps, MLRM and GAM used different approaches.

MLRM and GAM are shown in equation (6.1) and (6.2) (these are just examples, total 72 predictor variables were used in the initial model, which were minimised by stepwise regression model).

i. MLRM

$$\begin{aligned}&NO_2Conc \sim \beta o + \beta 1\ (Population) + \beta 2\ (BusStops) + \beta 3\ (Altitude) + \beta 4\\ &(IndustrialArea) + \beta 5\ (ResidentialArea) + \beta 6\ (BuildingArea) + \beta 7\ (ParksArea) + \beta 8\\ &(Easting) + \beta 9\ (Northing) + \beta 10\ (MajorRd) + \beta 11\ (MinorRd) + \beta 12\ (St\_Intersect) +\\ &\beta 13\ (CommercialArea) + \beta 14\ (Dist\_MajorRd) + \beta 15\ (Dist\_MinorRd) + \beta 16\\ &(Dist\_Building) + \beta 17\ (Dist\_Industry) + \beta 18\ (Dist\_Parks) + \beta 19\ (Dist\_Motorway) +\\ &\beta 20\ (Dist\_commercialArea) + \beta 21\ (Dist\_ResidentialArea) + \beta 22\ (Dist\_BusStop) + \beta 23\\ &(Motorway) + \varepsilon\end{aligned} \qquad (6.1)$$

In equation (1), βo is the intercept, β1 to β23 are the coefficients (slopes) of the predictor variables and ε is the error term (the difference between modelled and measured concentrations).

ii. GAM

$$\begin{aligned}&NO_2Conc \sim \alpha + s1\ (Population) + s2\ (BusStops) + s3\ (Altitude) + s4\ (IndustrialArea)\\ &+ s5\ (ResidentialArea) + s6\ (BuildingArea) + s7\ (ParksArea) + s8\ (Easting) + s9\\ &(Northing) + s10\ (MajorRd) + s11\ (MinorRd) + s12\ (St\_Intersect) + s13\\ &(CommercialArea) + s14\ (Dist\_MajorRd) + s15\ (Dist\_MinorRd) + s16\ (Dist\_Building)\\ &+ s17\ (Dist\_Industry) + s18\ (Dist\_Parks) + s19\ (Dist\_Motorway) + s20\\ &(Dist\_commercialArea) + s21\ (Dist\_ResidentialArea) + s22\ (Dist\_BusStop) + s23\\ &(Motorway) + \varepsilon\end{aligned} \qquad (6.2)$$

In equation (6.2), α is the intercept, 's' term is the smoothing function of the covariates and ε is the residual or error term, the difference between measured and predicted values. For smoothing term the degree of smoothing was automatically assigned by the generalised cross validation (GCV) method described by Wood and Augustin [24]. For more details on smoothing functions see Wood [25, 26]. MLRM explicitly assume normality of the error term and linearity of the relationship between response variable and predictor variables. GAM is an advanced model and relaxes such restrictions. Therefore, GAM is able to successfully handle nonlinearities in the association between response and predictor variables, which is important for air quality data as the relationship is not always linear. For more details on GAM see Hastie and Tibshirani [27], Wood [25] and Wood [26].

6.2.3. Model specification

Model specification (also refer to as model selection) is the process of determining which predictor variable to include or exclude from the model. In this study we employed stepwise regression algorithm (both forward and backward) for model selection. The aim was to find the best performing model (minimising prediction error) with minimum number of predictors (Parsimonious model) by selecting only those predictors whose contribution was significant in controlling the variations of NO_2 concentrations. For this purpose MASS-package [28] in R-programming language [23] was used.

6.2.4. Model validation

Model validation is testing the goodness of fit of the fitted model. In this process we compare the predicted concentrations with measured one. Here we used cross validation process which is a generalisation of the model to an independent dataset, not used in the model fitting. Randomly selected 75 % data was used for model fitting and 25 % for model validation.

6.2.5. Mapping modelled NO_2 concentration

After both MLRM and GAM were fitted and validated, NO_2 concentration was predicted for the entire Sheffield City for 37605 square grids each with size 100m x 100m. To do this MLRM and GAM used different techniques. In case of MLRM to predict NO_2 concentrations the coefficients of predictor variables were used in ArcGIS using 'field calculator' function in attribute table. However, GAM being a nonlinear model, doesn't produce a single coefficient (slope) for each predictor variable. Therefore, it was not possible to use the same approach. Instead, NO_2 concentration was predicted in mgcv-package [25] in R-programming language [23] and then exported to ArcGIS. Once in ArcGIS, the layer containing the predicted NO_2 concentration was joined with the polygon layer having the squared grids. For this purpose the whole study area was divided into 37605 square grids, each100m x 100m resolution using ArcGIS 10.7.1 software.

6.2.6. Statistical software

In this study mainly two statistical and mapping software were used: (a) ArcGIS version 10.7.1 and its LUR tools. LUR Tools is an ArcGIS toolbox having some important functions for constructing land use predictor variables for developing LUR model. (b) R programming language ([23] and several of its packages including 'MASS' [28], 'openair' [29] and 'mgcv' [25]. The 'mgcv-package' was used for running GAM, 'MASS' was used for running stepwise regression and 'openair-package' was used for general data analysis and developing different plots.

6.3. Results and discussion

In this study both MLRM and GAM are employed to model the spatial variability of NO_2 concentrations in Sheffield. NO_2 data were collected from two main sources: DT and LCS. Firstly, MLRM and GAM were developed using data from 188 DT (section 6.3.1), followed by the models using NO_2 data from LCS (section 6.3.2), and both DT & LCS (section 6.3.3). In all three case the data were divided into training dataset (randomly selected 75% data, used in the fitted model) and hold-out testing dataset (randomly selected 25 % data, used in the modle cross validation). Both GAM and MLRM were fitted using all predictor variables first, and then stepwise regression was used for model specification aiming to select only those predictors which had significant effect.

6.3.1. LUR model using NO_2 data from diffusion tubes

Both GAM and MLRM were used to regress NO_2 concentrations from DT against the predictor variables. Stepwise regression showed that the following covariates had significant effect: distance to major road (m), distance to minor road (m), residential area (m^2), commercial area (m^2), distance to bus stops (m), building area (m^2), and altitude (m). Among them, building and commercial area had positive coefficients, whereas residential area, distance to major road, distance to minor road, distance to bus stop and altitude had negative coefficients.

Three types of roads were used as covariates in the models: motorway, major roads (A-roads and B-roads), and minor roads. Among these the effect of major and minor roads was significant, the reason is obvious that these roads carry most of the traffic and are spread throughout the study area. In contrast motorway, although very busy has little length inside the study area as shown in Figure 6.2 (c) and therefore had an insignificant effect. It should be noted that negative coefficients of distance to major roads, minor roads and bus stops show positive effect of these three variables. In other words, as the distance between DT and these variables increase, NO_2 concentrations decrease. Therefore, areas near roads and bus stops have higher concentrations compared to areas away from the roads and bus stops, which is expected. Some researchers have suggested to take inverse or squared-inverse of the distance, e.g., [30] to turn the coefficients positive, however, it does not change the output of the model. Therefore, we stick to the original values. The negative effect of altitude on air quality is well known [31], meaning NO_2 concentration decreases at higher altitude. According to the data used in this study, minimum altitude was 26 m and maximum 551 m (Figure 6.2a). The effect of altitude was negative on NO_2 concentrations and highly significant. The western side of the city including Peak District National Park had higher altitude and lower NO_2 concentrations as compared to the city centre and eastern side, which had lower altitude and higher level of NO_2. The effect of residential area was negative, probably due to the fact that NO_2 levels were relatively lower at residential areas than at commercial areas and roadside locations. The effect of industrial area was insignificant probably due to the reasons that not many sensors are installed near industrial area. In addition, due to tall chimneys the emissions from the industry are not read by the local sensors. More recently Munir et al. [21] reported that industrial emissions in Sheffield had significant effect on the level of PM_{10}, but not on the levels of NO_2.

Coefficients and level of significance of different predictor variables for MLRM are shown in Table 6.2. Nonlinear model does not provide a single coefficients for each predictor variable, the output of the GAM model showing the association between NO_2 and predictor variables is shown in Figure 6.3 (only two predictor variables are shown for brevity). Figure 6.3 shows that NO_2 concentrations decrease drastically as distance from the roadside increases up to approximately 500 m, afterwards the curve is flattened as the distance increase further and there are fewer data points, resulting in wider error bars. In case of altitude, the negative effect is stronger at higher altitudes (altitude >180 m).

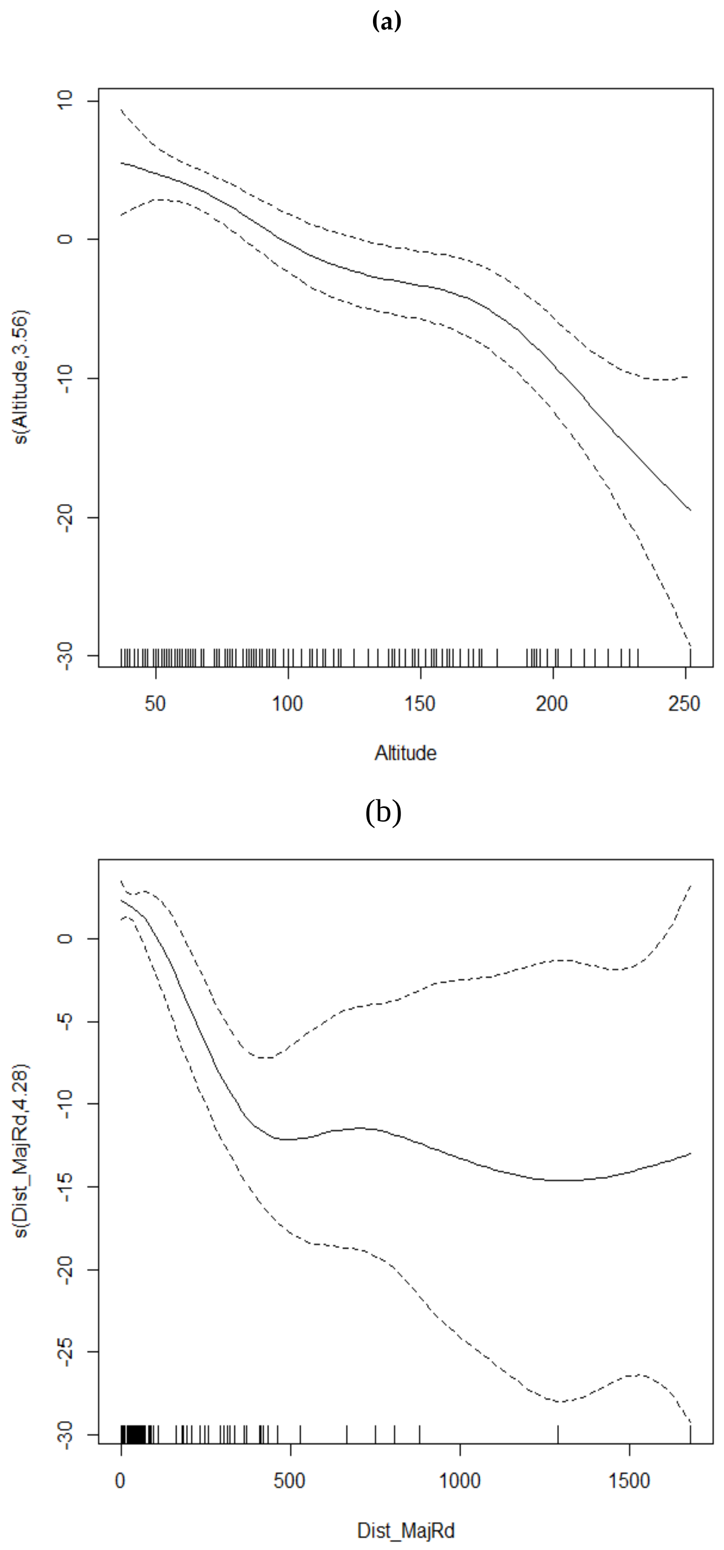
Several statistical metrics were calculated for model assessment for both fitted model (FM) and cross validation (CV) (Table 6.3). Statistical metrics used here were correlation coefficients (r, unitless), fraction of predictions within a factor of two of observations (FAC2, unitless), root mean squared error (RMSE, same units as the quantity being considered, here μg/m^3), mean bias (MB, same units as the quantity being considered, here μg/m^3), mean gross error (MGE, same units as the quantity being considered, here μg/m^3), normalised mean bias (NMB, unitless) and normalised mean gross error (NMGE, unitless). Generally, GAM with r-values 0.73 and 0.70 for training and hold-out dataset, respectively showed better performance than MLRM with r-value 0.67 for both training and hold-out dataset. Other metrics showing error of the model (e.g., NMGE, MGE and RMSE) are slightly greater for MLRM than GAM, indicating better performance of GAM in terms of smaller difference between predicted and observed concentrations. Furthermore, RMSE, NMGE and MGE have positive values, showing slightly over prediction of the model. FAC2 having value of 1 shows acceptable model performance for both MLRM and GAM. Results showed that models performance was not deteriorated considerably when applied to independent testing dataset. Comparison of predicted and measured NO_2 concentrations is made graphically in the form of scatter plots in Figure 6.4, showing strong correlation between modelled and measured concentrations for both models. Both models have over-predicted at lower levels of NO_2 (< 40 μg/m^3) and under-predicted at higher levels (> 40 μg/m^3), especially for the testing datasets (Figure 6.4 c & d).

Table 6.2. Showing coefficients and significance levels of different predictor variables for MLRM.

Predictor variable	Coefficient	p-value
Intercept	43.7025	< 2e-16 ***
Building	0.0020	0.075+
Dist_MajorRd	-0.0114	0.001 **
Dist_MinorRd	-0.0968	0.052+
Residential	-0.0006	0.002 **
Commercial	0.0005	0.053+
Altitude	-0.0543	0.000 ***
Dist_Bstop	-0.0308	0.026 *

Note: p.stars relate to how statistically significant the effect is: p-value < 0.001 = ***, p-value < 0.01 = **, p-value < 0.05 = * and p-value < 0.1 = +.

Furthermore, NO_2 concentration was predicted for 37605 square grids with 100m x 100m resolution to produce maps of NO_2 concentrations for the entire Sheffield City. The resultant maps of predicted NO_2 concentrations for both MLRM and GAM are shown in Figure 6.5. The model successfully captured spatial variability of NO_2 concentrations in Sheffield, showing higher levels of NO_2 in the city centre and on busy roads around the city. The city centre and the area between the motorway (M1) and the city centre is particularly highlighted with NO_2 levels higher than EU annual limits of 40 μg/m^3 (Figure 6.5). Western part of the city especially Peak District National Park has shown lower level of NO_2 concentrations due to high altitude and limited amount of minor and major roads. Recently, Munir et al. [21] using Airviro dispersion modelling system, have reported similar results in Sheffield. However, the results of this study are much more detailed (100m x 100m resolution), successfully capturing micro-level local variations in NO_2 concentrations, intended for quantifying public exposure. The maps show how NO_2 levels vary from street to street. Spatial trends of predicted NO_2 levels closely match with measured NO_2 levels (Figure 6.5).



(a)

(b)

Figure 6.3. Showing the nonlinear association of NO_2 concentrations (µg/m^3) with altitude in meters (a) and distance to major roads in meters (b). The dashed lines are the estimated 95% confidence intervals. The vertical lines on the x-axis show the presence of data.

Table 6.3. Showing various statistical metrics for assessing the performance of MLRM and GAM for both fitted model (FM) using training dataset and cross-validation (CV) using hold-out dataset based NO_2 data from DT. MB, MGE and RMSE have the same units as the quantity being considered (here μg/m^3), whereas FAC2, r, NMGE and NMB are unitless.

Metrics	MLRM		GAM	
	FM	**CV**	**FM**	**CV**
FAC2	1	1	1	1
MB	1.51e-15	-0.95	-3.68e-11	-1.17
MGE	4.98	6.89	4.80	6.55
NMB	4.45e-17	-0.03	-1.09e-12	-0.03
NMGE	0.15	0.19	0.14	0.18
RMSE	6.44	8.98	6.13	8.69
r	0.67	0.67	0.73	0.70

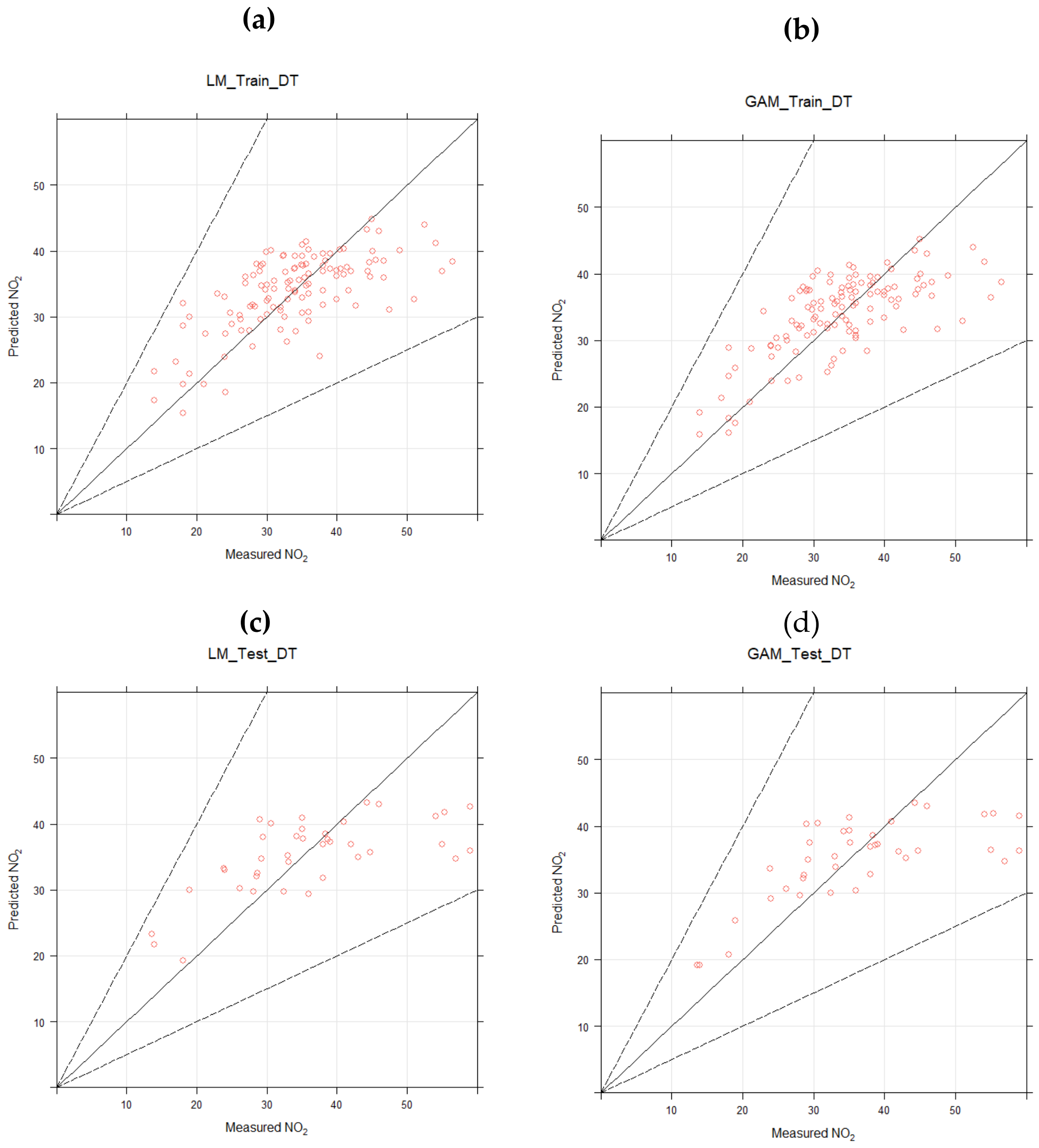


Figure 6.4. Comparing measured and predicted NO2 concentrations using multiple linear regression model (MLRM) and generalised additive model (GAM) for training (a and b) and testing dataset (c and d) using NO2 data from diffusion tubes (DT) only.

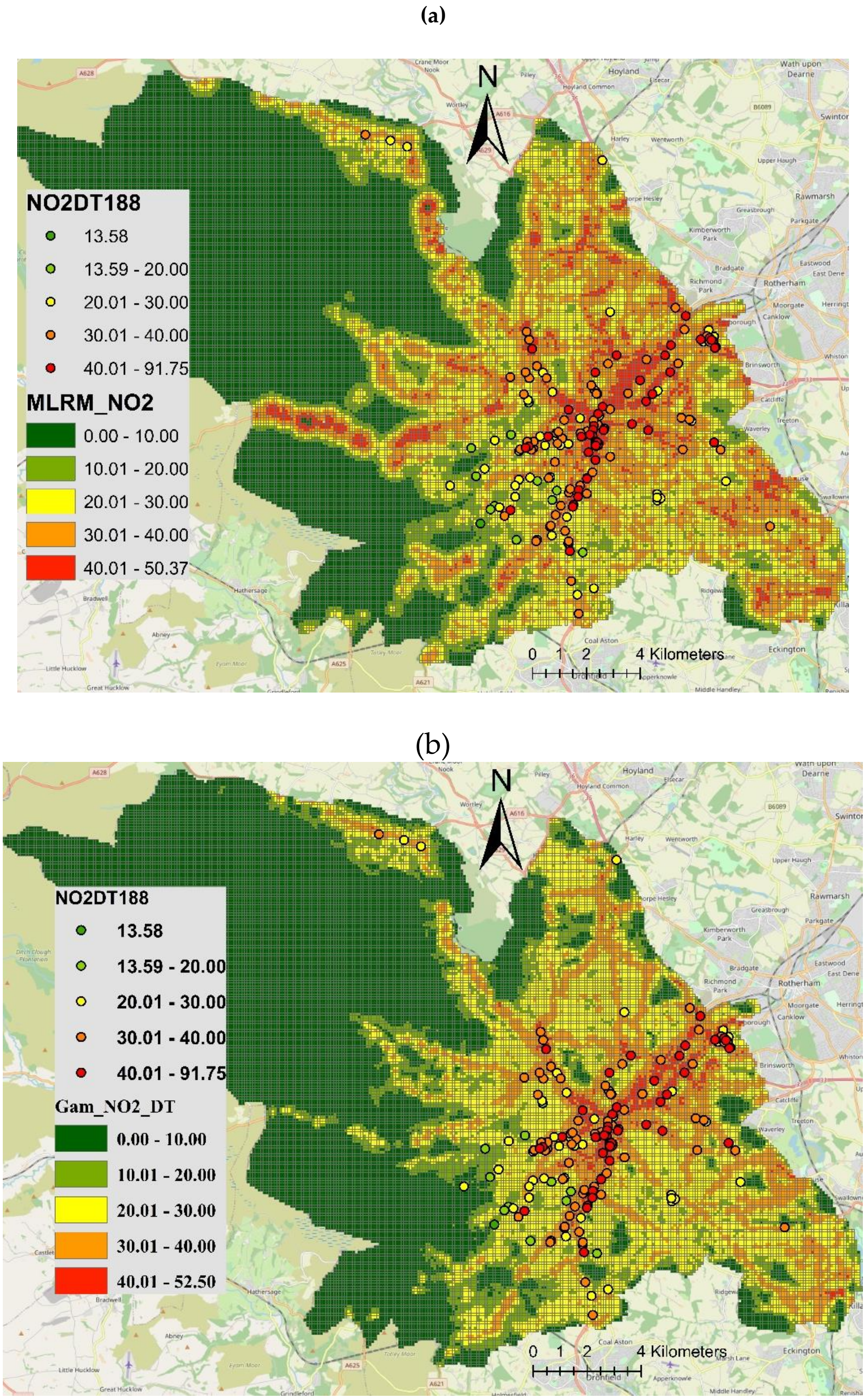


Figure 6.5. High resolution maps (100m x 100m) of NO_2 concentrations (μg/m^3) predicted by (a) MLRM and (b) GAM in Sheffield. The points show values of NO_2 measured by DT.

6.3.2. LUR model using NO_2 data from low-cost sensors

Employing MLRM and GAM, NO_2 concentration (μg/m^3) from 40 LCS was modelled using the same predictor variables as in section 6.3.1. Using a stepwise regression algorithm 5-covairaites were found to have a significant effect on NO_2 concentrations, which were: distance to major road, distance to minor road, distance to commercial area, distance to residential area, and altitude. Several statistical metrics for both fitted and cross validated models are shown in Table 6.4. Comparing fitted model, GAM with r-value of 0.89 showed better performance than the MLRM with r-value 0.55. However, when the models were applied to an independent set of data (hold-out dataset), MLRM showed better performance (r-value = 0.78) as compared to GAM (r-value = 0.56). Likewise, metrics (e.g., RMSE, MGE, NMGE, NMB) showing error of the models were lower for training dataset and higher for testing dataset for GAM than MLRM. FAC2 having value of 1 or nearly 1 shows acceptable model performance for both MLRM and GAM. The results of MLRM on hold-out dataset do not seem genuine as r-value is greater and RMSE is less than the fitted model, which probably shows that the model is over-fitted. This is also confirmed by the fact that overall both fitted and cross-validated model showed good performance, however, when the model results were extrapolated to the entire city, both MLRM and GAM models failed to predict the expected spatial variability of NO_2 concentrations in the city (Figure 6.6), which should have been more like Figure 6.5. This is probably due to the fact that LCS are mostly installed in the city centre and at the University of Sheffield, and are not representing the wider area of the city. Therefore, extrapolation of the modelled NO_2 outside the calibration range, results in unreasonable values. Secondly, there were only 40 LCS probably not providing enough spatial coverage to represent the whole city. Therefore, the models are over fitted and are not successfully fitting the data from the rest of the city.

Gillespie et al. [18] developed an LUR model to estimate exposure to NO_2 in Glasgow, Scotland. They used 135 NO_2 passive diffusion tubes, which were divided to four groups (32 – 35 sites per group) and models were developed using a combination of 1 to 3 groups as training sites to assess how the number of training sites affected the model performance. The explanatory variables used in the models were major road length, minor road length, all urban areas, building volume, distance to nearest major or minor road, green rural area, minor road length, and street configuration. The models were able to explain moderate to high variance in the data, where R^2 ranged from 0.62 to 0.89 for training dataset and 0.44 to 0.85 for hold-out dataset. Precision of estimated exposure was increased with increasing number of training sites. Gillespie et al. [16] concluded that the use of more than 60 training sites in LUR model has quantifiable benefits in epidemiological application. Therefore, this might suggest that probably in present study the number of sites (40 LCS) were not enough and resulted in over fitting of the model.

Table 6.4. Showing various statistical metrics for assessing the performance of MLRM and GAM for both fitted model (FM) using training dataset and cross-validation (CV) using hold-out dataset based on NO_2 data from LCS.

Metrics	MLRM		GAM	
	FM	CV	FM	CV
FAC2	0.97	1	1	0.90
MB	-9.18e-15	-0.48	-1.67e-10	-3.76
MGE	11.07	8.82	7.60	16.35
NMB	-2.42e-16	-0.01	-4.42e-12	-0.10
NMGE	0.29	0.24	0.20	0.44
RMSE	19.31	12.56	10.40	22.21
r	0.55	0.78	0.89	0.56

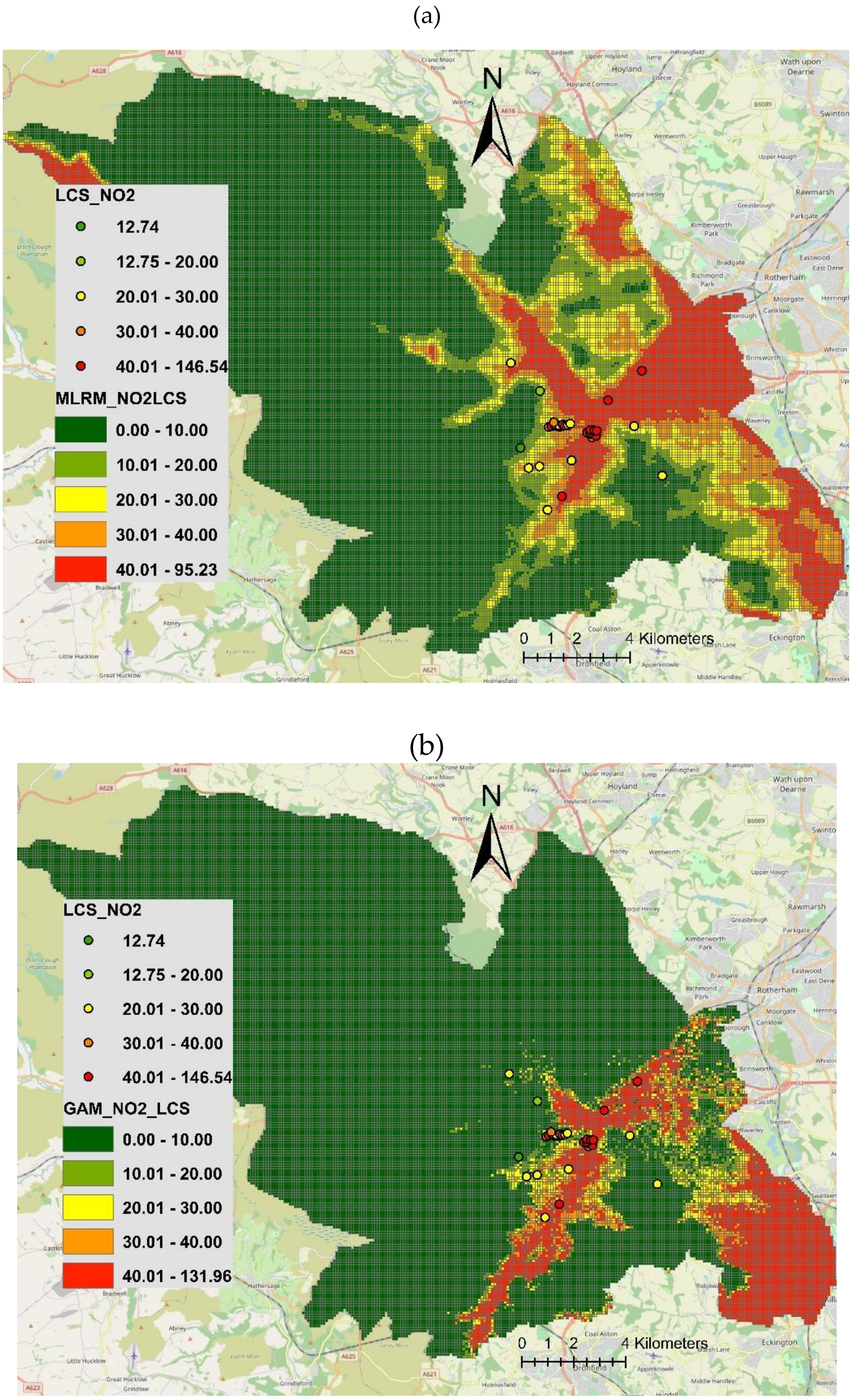


Figure 6.6. High resolution (100m x 100m) maps of NO_2 concentrations (μg/m^3) in Sheffield, predicted by (a) MLRM and (b) GAM. The points show values of NO_2 measured by LCS.

6.3.3. LUR model using NO_2 data from diffusion tubes DT and low-cost sensors

In this section NO_2 data from 188 DT and 40 LCS were combined to increase the number of monitoring sites and see how it affects the model outputs. At the model selection stage six predictors showed significant effect, which were distance to major road, minor roads length, commercial area, population, distance to commercial area and altitude. Performance of the model was assessed by calculating several statistical metrics as shown in Table 6.5. In case of MLRM, r-values were 0.60 and 0.52 for training and hold-out dataset, whereas in case of GAM r-values were 0.69 and 0.53, respectively. Metrics (e.g., RMSE, MGE, NMGE, MB and NMB) expressing error of the models have smaller values for GAM than MLRM, showing GAM outperforms MLRM. FAC2 also shows that prediction of GAM is closer to the measured values. Furthermore, comparing Table 6.5 with Table 6.3, it can be clearly observed that combining DT and LCS did not improve the models performance compared to the models using DT data only. This is probably due to the fact that DT and LCS use different techniques for measuring NO_2 concentrations and combining their data might not be the right thing to do. Second, DT provide long-term NO_2 concentrations (e.g., annual mean), whereas LCS provide short-term average (e.g., 5 minutes to 1 hour mean), which is converted to annual mean. It probably indicates that ideally all sensors should be of the same type and mixing of sensors of different grades might cause conflict and affect the model outputs. Finally, 188 DT are enough to build an LUR model, further increasing the number does not improve the model outputs.

Figure 6.7 shows the maps of predicted NO_2 concentrations, where NO_2 concentrations predicted by MLRM range from 0 to 50 μg/m^3, whereas NO_2 concentrations predicted by GAM range from 0 to 70 μg/m^3. City centre and the eastern side of the city towards the motorway are highlighted as having high levels of NO_2 due to lower altitude and greater length of major and minor roads and commercial areas.

Table 6.5. Showing various statistical metrics for assessing the performance of the MLRM and GAM for both fitted model (FM) using training dataset and cross-validation (CV) using hold-out dataset based on NO_2 data obtained from DT and LCS.

Metrics	**MLRM**		**GAM**	
	FM	**CV**	**FM**	**CV**
FAC2	0.93	0.89	0.99	0.91
MB	7.41e-15	-2.28	-4.01	-2.03
MGE	7.31	9.32	6.49	9.21
NMB	2.36e-16	-0.09	-1.28	-0.07
NMGE	0.23	0.31	0.21	0.30
RMSE	9.47	12.33	8.62	12.22
r	0.60	0.52	0.69	0.53

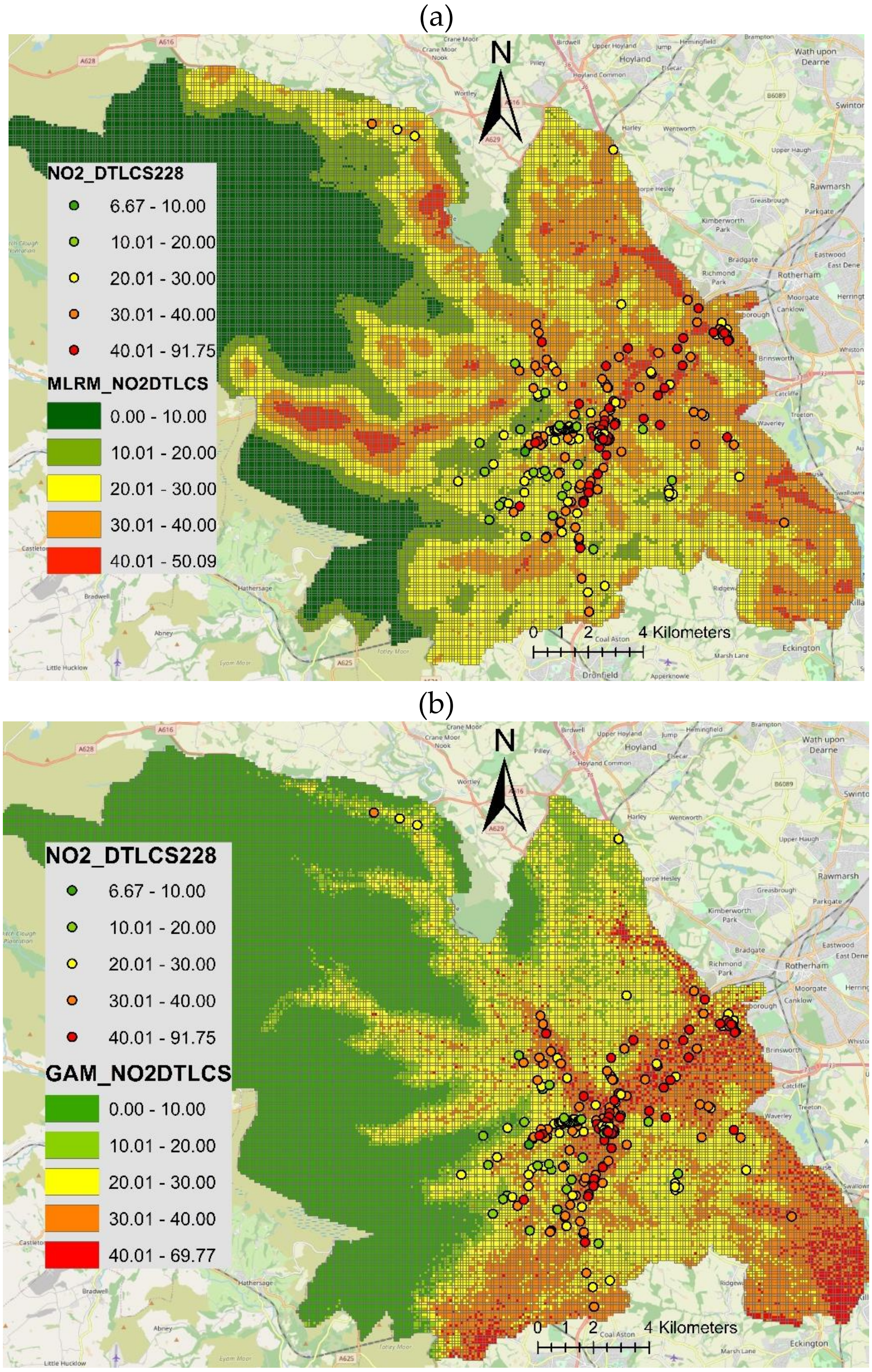


Figure 6.7. High resolution (100m x 100m) maps of predicted NO_2 concentrations (µg/m^3) in Sheffield, estimated by (a) MLRM, and (b) GAM. The points show values of NO_2 measured by both DT and LCS.

Major roads, minor roads, commercial areas and altitude had significant effects in all three models (described in sections 6.3.1, 6.3.2 and 6.3.3). Residential areas had significant effect in two models, whereas bus stops and population has significant effect only in one model. This shows that in addition to emission sources, topography especially altitude plays an important role in controlling the levels of air pollution. However, the effect of predictor variables is not the same everywhere and may change from region to region. To build an LUR model, researchers have used different sets of land use and traffic related features. The most common predictor variables used are topography, land use, traffic, population density, altitude and meteorological parameters. The use of these variables in the regression model are known to capture small scale variability [17]. The effect of predictor variables may vary from one area to another, depending on the nature of geographical conditions, size of urban area, the type of environment of the monitoring site (e.g. roadside, urban background or rural) and type of the pollutant modelled (e.g., NO_2, O_3, VOCs, PM_{10}, $PM_{2.5}$ etc.). The main constraint for building an LUR model is the unavailability of reasonable size monitoring network to provide measured data of pollutant for fitting and validating the model. This is particularly a problem in poor and low income countries having no air quality monitoring networks. To overcome this problem, Molter et al. [32] have suggested to use existing data from a dispersion air quality model. However, this is a problem in itself as dispersion model requires emission, geographical and meteorological data.

To the best of our knowledge, almost all previous investigations have used linear regression approach for building LUR model. Therefore, linear regression has become a default methodology for developing LUR models. This is due to the fact that in contrast to linear regressions, nonlinear regressions do not produce a single coefficient for each explanatory variable and therefore need to be applied in a different way to the traditional linear methods. With advances in IT and data analysis software (e.g., R programming language and Python), nonlinear LUR models can be developed and predicted in these software. These software have several special packages for spatial analysis and producing maps of measured or predicted pollutants. Therefore, maps of the predicted pollutants can be either produced in these software or alternatively the predicted pollutants can be exported to GIS software (e.g., ArcGIS) for further analysis. In this paper, the nonlinear model was developed using both R programming language and ArcGIS. The proposed method provides a basis for further application of advanced nonlinear modelling approaches to constructing LUR models in urban areas which enable quantifying small scale variability in pollution levels.

6.4. Conclusions

It has been a common practice to use the linear regression for developing LUR models, however, the association between air pollutant levels and spatial features is not always linear, therefore, ideally nonlinear modelling approaches should be used for developing LUR models, which can help in understanding small scale spatial variability of different air pollutants in urban areas. In this paper, the GAM was fitted and predicted in R programming language and the predicted concentrations was then transferred to ArcGIS for producing maps of NO_2 in Sheffield. Alternatively, R has several special packages that can be used for mapping and spatial analysis of the predicted concentrations.

In this paper spatial variability of NO_2 concentration is modelled in the city of Sheffield for year 2019. MLRM is a traditional and most commonly used approach for developing LUR models, here in addition to the linear approach, a nonlinear GAM model is employed and its benefits and the way it is applied are discussed. Three datasets of NO_2 measurements were used: (a) NO_2 data from 188 DT, (b) NO_2 from 40 LCS, and (c) NO_2 data from both DT and LCS. The first group performed better than the other two groups. Among predictor variables altitude (negative effect), major roads (positive effect), minor roads (positive effect), and commercial area (positive effect) had significant effect in all three groups. The model successfully captured the spatial variability of NO_2 in Sheffield, estimating high levels of NO_2 in the city centre and on major roads around the city. The eastern area between the city centre and motorway (M1) showed particularly high levels, whereas the western area (Peak District National Park) demonstrated lower levels of NO_2 concentrations.

The main contributions of this work are summarised as follows: (a) An advanced nonlinear GAM is proposed for developing an LUR model, which outperforms the linear counterpart. (b) High resolution maps (100 m x 100 m) of NO_2 are developed in Sheffield using a nonlinear LUR model for quantifying public exposure to NO_2 and determining how the exposure varies at small scales in the city. (c) NO_2 data measured by a network of DT and LCS are integrated to developed high resolution maps. (d) It is confirmed that Sheffield City Centre and its eastern sides experience relatively higher levels of NO_2 pollution.

Future work could include developing a spatiotemporal LUR model using meteorology, traffic counts and fleet composition data. In this study, the effect of major and minor roads is analysed on NO_2 levels, however, road traffic and composition may vary from time to time on a given road. It is, therefore, important to capture temporal and spatial variability in meteorology and traffic data and feed it to the nonlinear LUR models. This will probably further improve the model performance, depending on the quality and temporal resolution of the data.

Author Contributions: S.M., M.M. and D.C. planned and designed the original idea. S.M. obtained the data and did the modelling. M.M. and D.C. won the funding for this research project. S.M. wrote the first draft of the manuscript under the supervision of M.M. and D.C. L.S.M., M.M. and D.C. reviewed and edited the manuscript. All the authors have read and approved the content of the manuscript.

Funding: This research was funded by the Engineering and Physical Sciences Research Council (EPSRC) (grant number - EP/R512175/1) and Siemens plc. The APC was paid by UKRI/EPSRC.

Acknowledgement: We would like to express our thanks to the Engineering and Physical Sciences Research Council (EPSRC), grant number - EP/R512175/1 and Siemens plc for funding this project. We are also thankful to Sheffield City Council for providing NO_2 diffusion tubes data.

Conflicts of Interest: All the authors declare no conflict of interest.

6.5. References

1. Landrigan, P.J. Air pollution and health. The Lancet Public Health 2016, 2, (1), 4 –5. DOI: https://doi.org/10.1016/S2468-2667(16)30023-8.
2. WHO. Review of evidence on health aspects of air pollution-REVIHAAP project: final technical report. World Health Organziation Regional Office for Europe, 2013. http://www.euro.who.int/en/health-topics/environment-and-health/airquality/publications/2013/review-of-evidence-on-health-aspects-of-air-pollutionrevihaap-project-final-technical-report (Accessed on 12 February 2020).
3. DEFRA. Improving air quality in the UK Tackling nitrogen dioxide in our towns and cities, UK overview document, December 2015. Available online: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/486636/aq-plan-2015-overview-document.pdf (accessed on 09 April 2020).
4. Hsu, S.; Mavrogianni, A.; Hamilton, I. Comparing spatial interpolation techniques of local urban temperature for heat-related health risk estimation in a subtropical city. *Procedia Eng.* 2017, 198, 354 – 365.
5. Briggs, D.J. The Role of Gis: Coping With Space (And Time) in Air Pollution Exposure Assessment. J. Toxicol. Environ. Health (A) 2005, 68(13-14), 1243-61.
6. Schneider, P.; Castell, N.; Vogt, M.; Dauge, F.R.; Lahoz, W.A.; Bartonova, A. Mapping urban air quality in near real-time using observations from lowcost sensors and model information. *Environ. Int.* 2017, 106, 234-247.
7. Briggs, D.J.; de Hough, C.; Gulliver, J.; Wills, J.; Elliott, P.; Kingham, S.; Smallbone, K. A regression-based method for mapping traffic-related air pollution: Application and testing in four contrasting urban environments. *Sci. Total Environ.* 2000, 253, 151–167.
8. Hoek, G.; Beelen, R.; De Hoogh, K.; Vienneau, D.; Gulliver, J.; Fischer, P.; et al. A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmos. Environ.* 2008, 42, 7561-7578.
9. Briggs, D.J.; Collins, S.; Elliott, P.; et al.. Mapping urban air pollution using GIS: a regression-based approach. *Int. J. Geogr. Inf. Sci.* 1997, 11, 699–718.
10. Eeftens, M.; Beelen, R.; de Hoogh, K.; et al. Development of land use regression models for $PM_{2.5}$, $PM_{2.5}$ absorbance, PM_{10} and PM coarse in 20 European Study areas; results of the ESCAPE project. *Environ. Sci. Technol.* 2012, 46 (20), 11195–11205.

11. Lee, J.H.; Wu, C.F.; Hoek, G.; et al., 2013. Land use regression models for estimating individual NOx and NO2 exposures in a metropolis with a high density of traffic roads and population. *Sci. Total Environ.* 2013, 472, 1163–1171.
12. Muttoo, S.; Ramsay, L.; Brunekreef, B.; Beelen, R.; Meliefste, K.; Naidoo, R.N. Land use regression modelling estimating nitrogen oxides exposure in industrial south Durban, South Africa. *Sci. Total Environ.* 2017, 610–611, 1439–1447.
13. Rahman, Md. M.; Yeganeh, B.; Clifford, S.; Knibbs, L.D.; Morawska, L. Development of a land use regression model for daily NO_2 and NOx concentrations in the Brisbane metropolitan area, Australia. *Environ. Modell. Softw.* 2017, 95, 168 - 179.
14. Stedman, J.; Vincent, K.; Campbell, G.; Goodwin, J.; Downing, C. New high resolution maps of estimated background ambient NOx and NO2 concentrations in the U.K. *Atmos. Environ.* 1997, 31, 3591–3602.
15. Beelen, R.; Hoek, G.; Fischer, P.; van den Brandt, P.A.; Brunekreef, B. Estimated long-term outdoor air pollution concentrations in a cohort study. *Atmos. Environ*. 2007, 41, 1343–1358.
16. Ryan, P.H.; LeMasters, G.K.; Biswas, P.; Levin, L.; Hu, S.; Lindsey, M.; Bernstein, D.I.; Lockey, J.; Villareal, M.; Hershey, G.K.K.; Grinshpun, S.A. A comparison of proximity and land use regression traffic exposure models and wheezing in infants. *Environ. Health Perspect.* 2007, 115, 278–284.
17. Ryan, P.H.; LeMasters, G.K. A review of land-use regressionmodels for characterizing intraurban air pollution exposure. *Inhal. Toxicol.* 2007, 19 (Suppl. 1), 127–133.
18. Gillespie, J.; Beverland, I.J.; Hamilton, S.; Padmanabhan, S. Development, Evaluation, and Comparison of Land Use Regression Modeling Methods to Estimate Residential Exposure to Nitrogen Dioxide in a Cohort Study. *Environ. Sci. Technol.* 2016, 50, 11085–11093.
19. Beelen, R.; Hoek, G.; Vienneau, D.; Eeftens., M.; Dimakopoulou, K.; Pedeli, X.; Tsai, M.Y. Development of NO_2 and NOx land use regression models for estimating air pollution exposure in 36 study areas in Europe e-The ESCAPE project. *Atmos. Environ*. 2013, 72, 10 - 23.
20. Vienneau, D.; deHoogh, K.; Beelen, R.; Fischer, P.; Hoek, G.; Briggs, D. Comparison of land-use regression models between Great Britain and the Netherlands. *Atmos. Environ.* 2010, 44, 688-696.
21. Munir, S.; Mayfield, M.; Coca, D.; Mihaylova, L.S.; Osammor, O. Analysis of air pollution in urban areas with Airviro dispersion model - A Case Study in the City of Sheffield, United Kingdom. *Atmos.* 2020, 11(3), 285.
22. Munir, S.; Mayfield, M.; Coca, D.; Jubb, S.A. Structuring an Integrated Air Quality Monitoring Nework in Large Urban Areas – Discussing the Purpose, Criteria and Deployment Strategy. *Atmos. Environ: X* 2019, 100027.
23. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria 2019. URL: https://www.R-project.org/.
24. Wood, S. N.; Augustin, N. H. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecol. Modell.* 2002, 157(2-3), 157-177.
25. Wood, S.N. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. Royal Stat. Soc. (B)* 2011, 73(1), 3-36.
26. Wood, S.N. Generalized additive models: an introduction with R 2006. Chapman and Hall/CRC.
27. Hastie, T.J.; Tibshirani, R.J. Generalised Additive Models 1990. Chapman & Hall, London.
28. Venables, W. N.; Ripley, B. D. Modern applied statistics with S. Fourth edition 2002. Springer, New York. ISBN 0-387-95457-0.
29. Carslaw, D.C.; Ropkins, K. openair - an R package for air quality data analysis. *Environ. Modell. Softw.* 2012, 27-28, 52-61.

30. Korek, M.; Johansson, C.; Svensson, N.; Lind, T.; Beelen, R.; Hoek, G.; Pershagen, G.; Bellander, T. Can dispersion modeling of air pollution be improved by land-use regression? An example from Stockholm, Sweden. *J. Expo. Sci. Environ. Epidemiol.* 2017, 27, 575–581.
31. Ji, H.; Chen, S.; Zhang, Y.; Chen, H.; Guo, P.; Zhao, P. Comparison of air quality at different altitudes from multi-platform measurements in Beijing. *Atmos. Chem. Phys.* 2018, 18, 10645–10653.
32. Molter, A.; Lindley, S.; de Vocht, F.; Simpson, A.; Agius, R. Modelling air pollution for epidemiologic research — Part I: A novel approach combining land use regression and air dispersion. *Sci. Total Environ.* 2010, 408, 5862–5869.

CHAPTER 7: UNDERSTANDING SPATIAL VARIABILITY OF NO_2 IN URBAN AREAS USING SPATIAL MODELLING AND DATA FUSION APPROACHES

Said Munir[a,*], Martin Mayfield[a], Daniel Coca[b]

[a]Department of Civil and Structural Engineering, the University of Sheffield, Sheffield, S1 3JD, UK
[b]Department of Automatic Control and Systems Engineering, the University of Sheffield, Sheffield, S1 3JD, UK
*corresponding author (smunir2@sheffield.ac.uk), Mob: +447986001328, Fax: +44 (0) 114 222 5700

Published: Atmosphere 2021, 12(2), 179
https://doi.org/10.3390/atmos12020179
Received: 3 January 2021; Revised: 23 January 2021; Accepted: 25 January 2021; Published: 29 January 2021.

Abstract

Small-scale spatial variability in NO_2 concentrations is analysed with the help of pollution maps. Maps of NO_2 estimated by the Airviro dispersion model and land use regression (LUR) model are fused with measured NO_2 concentrations from low-cost sensors (LCS), reference sensors and diffusion tubes. In this study, geostatistical universal kriging was employed for fusing (integrating) model estimations with measured NO_2 concentrations. The results showed that the data fusion approach was capable of estimating realistic NO_2 concentration maps that inherited spatial patterns of the pollutant from the model estimations and adjusted the modelled values using the measured concentrations. Maps produced by the fusion of NO_2-LCS with NO_2-LUR produced better results, with r-value 0.96 and RMSE 9.09. Data fusion adds value to both measured and estimated concentrations: the measured data are improved by predicting spatiotemporal gaps, whereas the modelled data are improved by constraining them with observed data. Hotspots of NO_2 were shown in the city centre, eastern parts of the city towards the motorway (M1) and on some major roads. Air quality standards were exceeded at several locations in Sheffield, where annual mean NO_2 levels were higher than 40 µg/m^3. Road traffic was considered to be the dominant emission source of NO_2 in Sheffield.

Keywords: nitrogen dioxide; spatial variability; urban air quality; data fusion; dispersion modelling; land use regression; Sheffield

7.1. Introduction

Poor air quality is one of the growing environmental issues in urban areas. Long-term exposure to air pollution is known to cause various health issues including both chronic and acute respiratory diseases such as asthma, cancer, cardiovascular diseases, hospital admission and mortality [1-3]. Air pollution caused 6.4 million deaths worldwide in 2015, showing the significance of the impact of poor air quality on human health [4]. Both particulates (e.g., PM_{10} and $PM_{2.5}$) and gaseous pollutants have negative impacts on human health; however, among gaseous pollutants, nitrogen dioxide (NO_2) is considered the most serious pollutant for human health [3]. Many Air Quality Management Areas (AQMAs) in the UK were declared due the high levels of NO_2 exceeding air quality standards set for human health protection [5]. Among the other challenges of air pollution in urban areas is the quantification of small-scale spatial variability, which is controlled by local emission sources, land use features, building density and height and geographical characteristics. To characterise spatial variability in pollutant concentrations, three approaches are used most commonly [6]: GIS-based interpolation methods, dispersion models and land use regression (LUR) models.

Interpolation methods are used to interpolate (predict) air pollutant concentrations between various air quality monitoring stations to provide better spatial coverage. One of the basic and probably most widely used approaches for interpolation in geostatistics is kriging. Kriging interpolates air pollutant levels (e.g., NO_2 concentrations) between two points by modelling them with Gaussian processes. However, when these approaches are applied at a local scale, such as an intra-city scale, these methodologies are known to produce considerable variations in air pollutant concentrations within a small area and are more effective at large scales, such as national or regional scale [6, 7]. For handling these shortcomings, more advanced kriging techniques (e.g., cokriging and universal kriging) can be used to combine modelled and measured values of NO_2, which further improve the performance of the interpolation [8, 9]. The second approach for spatial modelling is dispersion modelling techniques, which are probably the most advanced modelling techniques for determining the spatial variability of air pollutants in urban areas. Dispersion models use emission data of different sources (e.g., point, line and area sources), the geophysical characteristics of the study area and meteorological parameters. Dispersion models have the potential to incorporate both temporal and spatial dimensions to complement air pollution monitoring. The third approach for modelling spatial variability is LUR models, which provide an effective alternative to GIS interpolations and dispersion modelling techniques on urban scales. LUR is a spatial modelling approach used most frequently for analysing the spatial variability and quantifying public exposure to air pollution in urban areas. Several authors have preferred LUR to the other approaches [6, 10] due the fact that it produces realistic and detailed maps and is easy to apply.

Integrating and merging data and information from several sources is known as data fusion [11]. In the literature, data fusion is also referred to as decision fusion, data combination, data aggregation, multisensor data fusion and sensor fusion. Data fusion techniques merge observed data with modelled data in a mathematical, objective way, adding value to both observed and modelled data. The observed data are improved by filling spatiotemporal gaps, whereas the modelled data are improved by constraining them with observed data [9]. Therefore, data fusion of observed data with modelled data can improve urban-scale air quality mapping. Schneider et al. [9] reported that the accuracy of fused data normally depends on a range of factors, which include (a) the total number of observations, (b) spatial distribution of the network, (c) uncertainty of the measured data, and (d) the ability of the model to accurately predict air pollutant levels with high spatial and temporal resolution.

In this study, firstly, the maps produced by the LUR and Airviro dispersion models are compared for analysing the spatial variability of NO_2 concentrations in Sheffield [12, 13]. Secondly, using the universal kriging technique, modelled NO_2 concentrations are fused with the measured concentrations in a view to further improve the spatial maps of NO_2. Measured NO_2 concentrations in the form of points provide absolute values, whereas modelled concentrations provide spatial patterns for the fused high-resolution maps, which are helpful for understanding local-level spatial variability of air pollution and quantifying public exposure in urban areas. This study used measured NO_2 concentrations from three sources (reference sensors, low-cost sensors (LCS) and diffusion tubes) and estimated concentrations from two sources (Airviro dispersion model and LUR model). Measurements of LCS and diffusion tubes are not as reliable as of reference sensors; however, due to cheaper price and maintenance they provide better spatial coverage and a unique opportunity for the spatial modelling of different air pollutants.

The rest of the paper is structured as follows: the methodology of this paper is presented in Section 7.2. In the Methodology, Section 7.2.1 provides a brief description of the study area, Section 7.2.2 describes the air quality monitoring network (AQMN) in Sheffield, Section 7.2.3 describes the NO_2 map estimated by Airviro, Section 7.2.4 describes the NO_2 map estimated by LUR and Section 7.2.5 describes kriging and universal kriging techniques. Results and discussion are presented in Section 7.3 and the main conclusions of this study are presented in Section 7.4.

7.2. Methodology

In this paper, NO_2 concentrations are analysed from the AQMN in Sheffield. In the first part of the paper, spatial variability of NO_2 concentrations (µg/m^3) is analysed using three modelling approaches: kriging interpolation, Airviro dispersion model [13] and LUR model [12]. In the second part, NO_2 concentrations measured by various sensors are fused with the NO_2 concentrations estimated by both Airviro and LUR models. For data fusion, one of the geostatistical techniques known as universal kriging is used, which was employed in R programming language using "automap" package [14]. The aim is to develop high-resolution NO_2 maps for understanding the spatial variability of NO_2 in the city of Sheffield, UK.

Below, a brief description of the study area and the AQMN in Sheffield is provided, followed by a description and comparison of the Airviro and LUR NO_2 estimated maps.

7.2.1. Brief Description of the Study Area

Sheffield (53°23′ N, 1°28′ W) is one of the oldest and most historical cities in South Yorkshire, UK. Sheffield is the second largest city in the Yorkshire and Humber region and had a population of 584,853 in 2019. Sheffield is known as a green city and 61% of its area is composed of green space. The Peak District National Park and Pennine upland range lie to the west of the city and constitute one third of the city's area. Sheffield has a temperate climate. Air pollution is a serious environmental issue in Sheffield and most of the urban area has been declared as an Air Quality Management Area (AQMA) due to the high levels of NO_2 and particulate matter (PM_{10}). Sheffield has a large AQMN, which is briefly described below (Section 7.2.2).

7.2.2. Air Quality Monitoring Network (AQMN) in Sheffield

There is a large network of one hundred and eighty-eight (188) NO_2 diffusion tubes (DT) in Sheffield (Figure 7.1a). The Urban Flows Observatory, at the University of Sheffield, has made a network of forty-one (41) low-cost sensors (LCS) which has twenty-eight (28) Envirowatch E-MOTEs and thirteen (13) AQMesh pods, providing continuous hourly data of NO_2 concentrations (Figure 7.1b). Furthermore, in Sheffield, there are three (3) Automatic Urban and Rural Network (AURN) reference sites run by the UK Department for Environment, Food and Rural Affairs (DEFRA) and five (5) reference stations run by the Sheffield City Council (SCC) (Figure 7.1b). All these sensors make a multi-layer network providing NO_2 data. In this paper, NO_2 data for over a year (August 2019 to September 2020) are analysed. A summary of the NO_2 data from the network is provided in Table 7.1. Further details on the network can be found in [12].

Table 7.1. Summary of annual mean NO_2 concentrations (µg/m^3) measured by diffusion tubes (DT), automatic urban and rural network (AURN), Sheffield City Council (SCC) and low-cost sensors (LCS) (AQMesh and Envirowatch) in Sheffield (August 2019 to September 2020).

Metrics	DT NO_2	AURN	SCC	AQMesh/Envirowatch E-MOTE
Minimum	13.58	19.09	8.12	12.69
1st Quartile	28.29	21.00	24.05	25.08
Median	33.77	22.90	24.62	34.23
Mean	34.23	24.69	21.53	38.70
3rd Quartile	40.00	27.50	25.02	42.36
Maximum	91.75	32.09	25.86	136.81
Standard Deviation	9.65	6.68	7.53	27.69
Number of Sensors	188	3	5	41

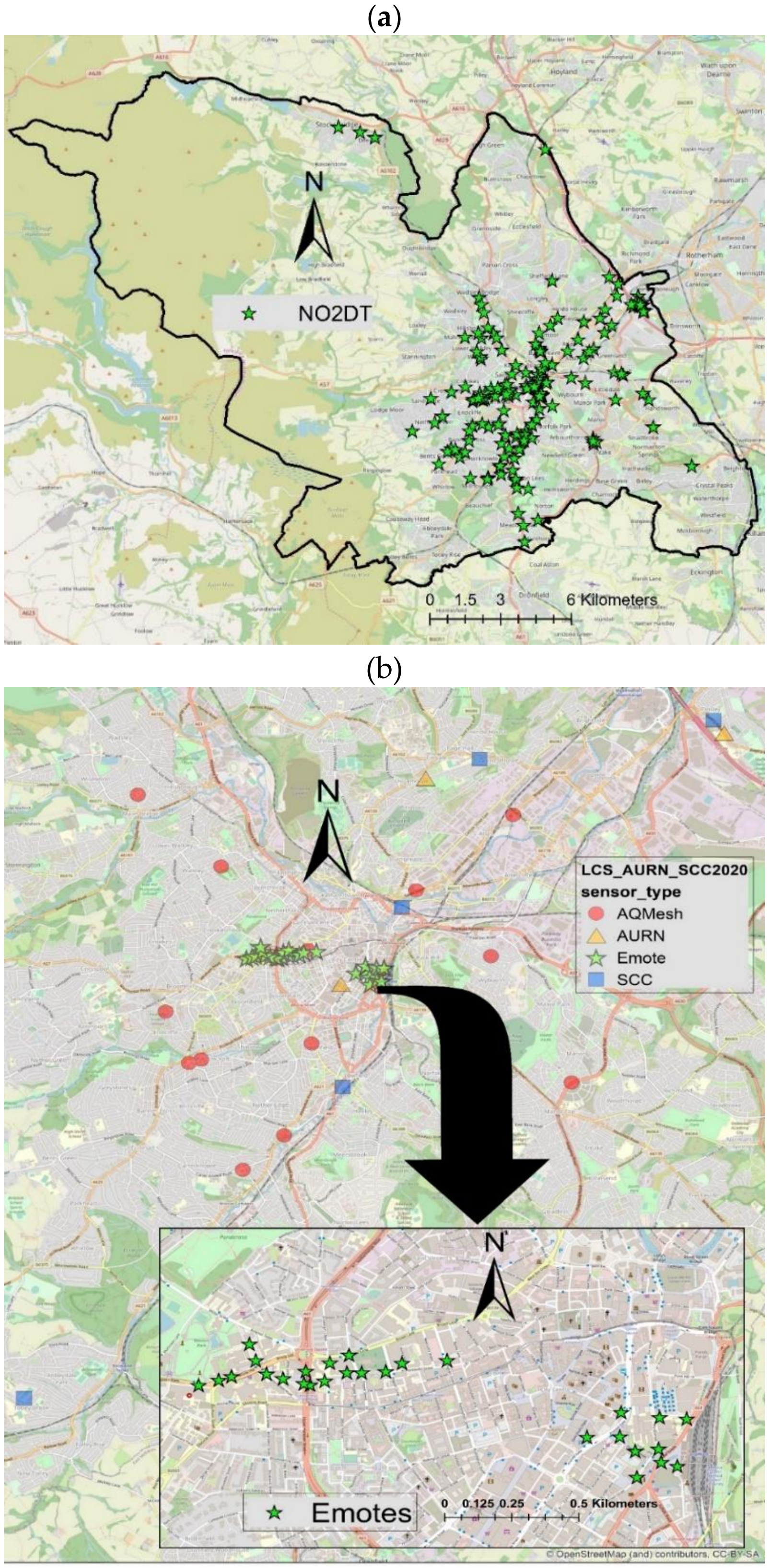


Figure 7.1. Showing the locations of DT, LCS (both AQMesh pods and Envirowatch E-MOTEs), AURN and SCC monitoring sites in Sheffield: (**a**) DT locations; (**b**) AURN, SCC and LCS locations.

7.2.3. NO$_2$ Map Estimated by Airviro

The Airviro dispersion model's estimated map of NO$_2$ concentrations is adopted from Munir et al. [13], who used the Airviro (version 4.01) dispersion model, which is an integrated modelling system for managing the emission database, modelling the dispersion of pollutants, and handling air quality data. Airviro is a state of the art dispersion model used by many researchers, consultants and local authorities globally for air quality modelling. Like other dispersion models, Airviro requires local topography, pollutant emissions and meteorological data to produce maps of estimated air pollutants. The Airviro model, the required inputs and produced outputs are discussed in detail in Munir et al.[13]. Munir et al. [13] estimated the map of NOx concentrations, which were converted to NO$_2$ by analysing the ratio of NO$_2$/NOx. The original map is given in Munir et al. [13] and the adopted map which is further analysed in this paper is shown here in Figure 7.2. The resulting estimated map showed hotspots of NO$_2$ concentrations around the city centre and eastern part of the city towards the Tinsley Roundabout (M1 J34S). NO$_2$ concentrations decrease gradually going away from the city centre, particularly towards the west and northwest of the city.

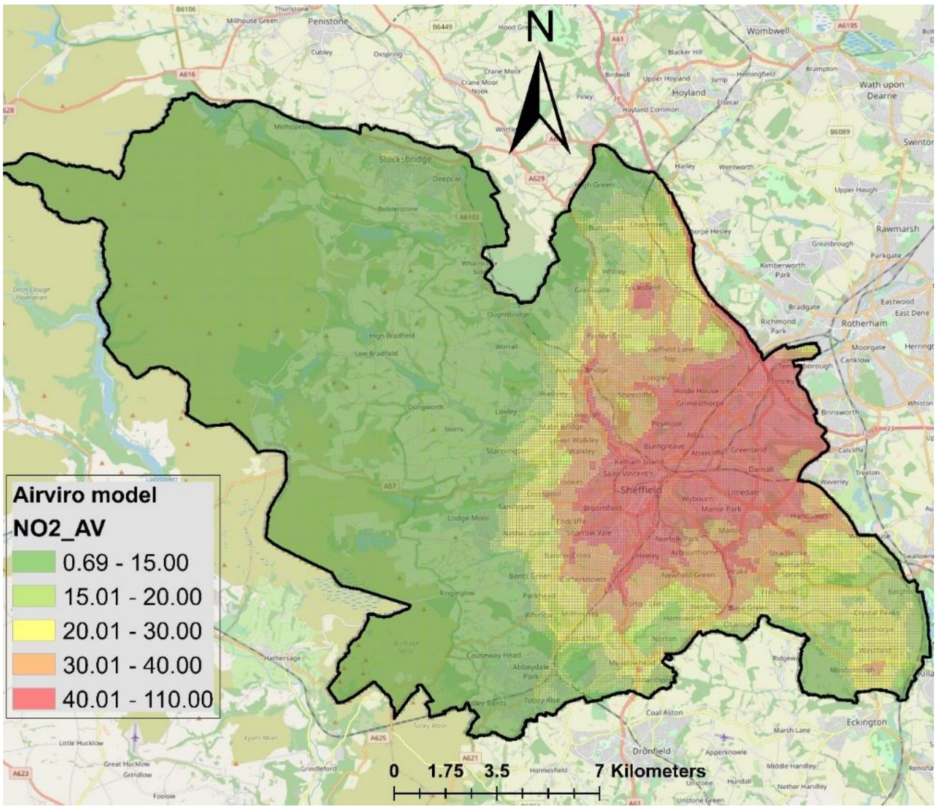


Figure 7.2. Spatial variability of annual mean modelled NO$_2$ in Sheffield. Maps were developed using Airviro model (modified from [13]).

7.2.4. NO$_2$ Map Estimated by LUR

The LUR map used here is adopted from Munir et al. [12]. LUR is a widely employed approach for the estimation of air pollution exposure using geographical information systems (GIS) and statistical analysis to determine the association between geographical features and measured atmospheric pollutant concentrations [15]. The LUR approach was introduced by Briggs et al. [15] and since then has been used in numerous studies around the world [16 - 23]. LUR models associate pollutant concentrations, such as NO$_2$, to site-specific geographical characteristics, e.g., topography, land use, population density and altitude. The use of these variables in the regression model is known to capture small-scale variability on a city scale [24]. For more details, see [12]. The resulting map produced by the

LUR model [12], employing a generalised additive model, a nonlinear regression approach, is shown in Figure 7.3, which is a high-resolution (100 m × 100 m) map. Figure 7.3 demonstrates higher levels of NO_2 in the city centre, eastern side of the city and on major roads. As compared to Figure 7.2, here, the busy roads are successfully highlighted, showing high levels of NO_2 concentrations.

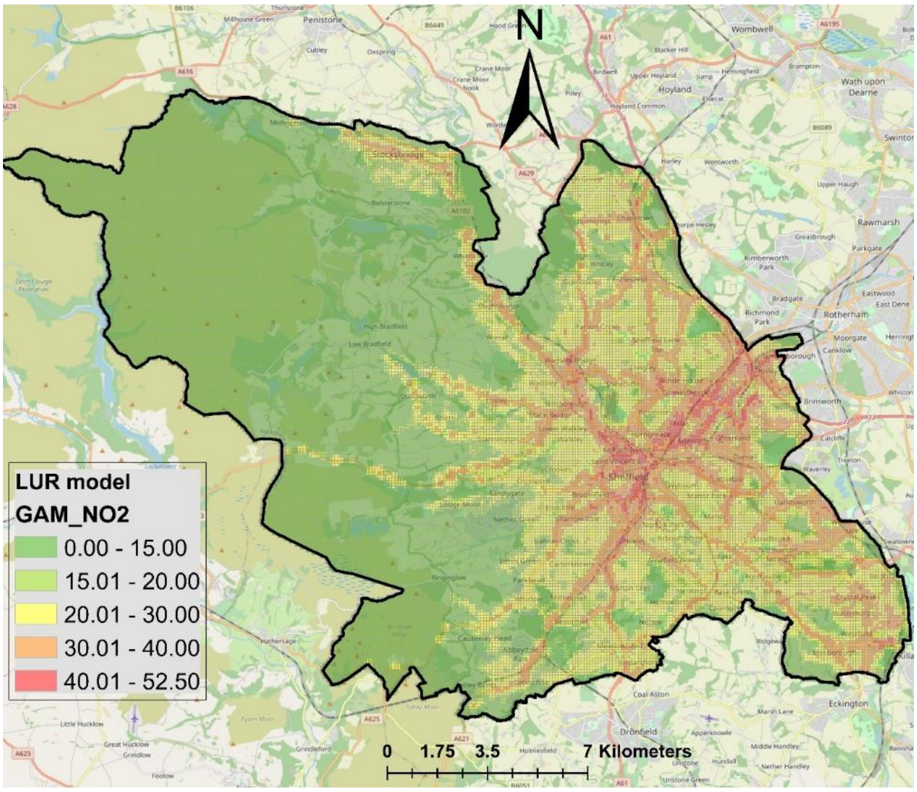


Figure 7.3. Maps of annual mean modelled NO_2 ($\mu g/m^3$) estimated by LUR model in Sheffield. Modified from [12].

7.2.5. Kriging and Universal Kriging

Kriging is a type of geostatistical interpolation technique based on statistical models that include statistical relationships among the measured points (autocorrelation). Kriging is a form of probabilistic and local interpolation technique. Kriging uses Gaussian processes for interpolating between various spatial points.

The kriging formula is expressed as:

$$Z_o(S_o) = \sum_{i=1}^{i=n} \lambda i Zi(Si) \qquad (7.1)$$

where Zi(Si) is the measured value at the ith location, λi is an unknown weight for the measured value at the ith location, $Z_o(S_o)$ is the predicted value at the prediction location.

Simple kriging is used for stationary data if the mean is known; otherwise, ordinary kriging is used. To fuse (integrate) model estimations with measured NO_2 concentrations, here, we employed the universal kriging technique, which is more advanced and capable of handling data with more than one correlated variable. In universal kriging, the additional observations of one or more covariates may lead to increased precision of the predictions. This approach enables the merging of NO_2 observations from a network of sensors, with modelled values providing spatial information from the air quality models. Universal kriging, in contrast to ordinary kriging, allows the overall mean to be non-constant throughout the domain and to be a function of one or more explanatory variables [9, 25]. Universal kriging is similar to kriging, with external drift and mathematically equivalent to regression kriging or residual kriging [8]. In this paper, measured values are fused with LUR and Airviro estimation using "automap" packages [14] in R programming language [26].

7.2.6. Model Validation

Performance of the universal kriging technique was assessed by calculating several statistical metrics. Statistical metrics used for comparing measured and estimated concentrations were factor of two (FAC2), mean bias error (MBE), mean absolute error (MAE), root mean square error (RMSE) and correlation coefficient (r). MAE and RMSE show the size of the average error; however, they do not provide information on the relative magnitude of the difference between predicted and observed values as these are based on absolute values of the difference between estimated and measured values. On the other hand, MBE describes the direction of the error bias. A negative value of MBE shows that predicted values are smaller than the observed values, showing underprediction of the model. FAC2 is the percentage of the predictions within a factor of two of the observed values and the correlation coefficient demonstrates the linear relationship between observed and estimated concentrations. For more details on these metrics see Munir et al. {12, 13].

For comparison purposes, the measured data collected by different sets of sensors were split into training and testing datasets. Training dataset (75% of the data) and testing dataset (25% of the dataset) were randomly selected using "caTools" package of R programming language. After assessing the model performance, the model was retrained using the whole dataset and applied to the rest of the city where measured data were not available.

7.3. Results and Discussion

In this section, firstly, the measured and interpolated NO_2 concentrations (μg/m^3) are analysed (Section 7.3.1), followed by data fusion (Section 7.3.2), wherein model estimations are fused with measured NO_2 concentrations.

7.3.1. Measured and Interpolated NO_2 Concentrations

Figure 7.4 demonstrates both measured and interpolated NO_2 concentrations for DT. Annual mean of measured NO_2 concentrations (μg/m^3) ranged from 13.58 to 91.75. It should be noted that the air quality limit for annual mean NO_2 concentrations is 40 (μg/m^3); therefore, to protect human health, annual mean NO_2 levels should not exceed this level. The observed levels (Figure 7.4) show that the air quality limits are exceeded at several sites in the city, especially in the city centre and on main roads. These readings are only for point locations and there are large gaps between these locations. To create a continuous map, in this paper, we used ordinary kriging for predicting NO_2 concentrations for the locations where measured concentrations were not available. Interpolated concentrations are shown in Figure 7.4b, which demonstrates that air quality standards are violated in some parts of the city centre and in the northeast of the city. However, in most of the city centre and northeast of the city, NO_2 levels range from 35 to 40 (μg/m^3). In the northwest and areas adjacent to the city centre, the NO_2 levels ranged from 30 to 35. Deep green and light green areas mostly in the southwest part of the city show NO_2 levels in the range of 20 to 30 (μg/m^3), which is the area next to the Peak District National Park.

Figure 7.5 shows NO_2 levels measured by LCS, ranging from 8.12 to 136.81 (μg/m^3). Exceedances of air quality standards are mainly shown in the city centre, where 10 Envirowatch E-MOTEs were installed. However, due to the low number of sensors compared to DT, the interpolated maps are not as detailed as those produced by the DT. Envirowatch E-MOTEs installed in the city centre recorded relatively higher concentrations than the other parts of the city. Three Envirowatch E-MOTEs recorded particularly higher NO_2 concentrations: (a) outside Sheffield train station next to the taxi rank, E-MOTE 904 (136.81 μg/m^3), (b) Arundel Gate opposite to Genting Club, E-MOTE 732 (115.55 μg/m^3), and (c) Sheaf Street/Sheaf Square adjacent to the pedestrian crossing for Howard Street, E-MOTE 903 (107.17

μg/m^3). The reason for the higher concentrations is that these sites are very busy in terms of traffic flow and idling vehicles while stationary. The taxi stand (E-MOTE 904) is a good example, showing how idling vehicles contribute to air pollution in the city. E-MOTE 903 is installed next to a pedestrian crossing on a busy road. People coming out of the train station use the pedestrian crossing going towards the high street, Sheffield Hallam University and other parts of the city. The pedestrian crossing is regularly used, interrupting the traffic flow and causing congestion. When the lights turn red, road traffic stops but vehicle engines keep running. As soon as the lights turn green, all vehicles try to accelerate quickly. Therefore, idling of engine and sudden acceleration emits extra pollution, causing this location to be one of the most polluted sites. E-MOTE 732 is installed in a typical street canyon, where the road has tall buildings on both sides, hindering the dispersion of the pollutants emitted by the road traffic and causing the pollution levels to increase. The other sites in the city centre where air quality standards were exceeded are Paternoster Rows, Harmer Lane near Sheaf Street, Harmer Lane near the bus station, Arundel Gate near Surrey Street and Howard Street. Two E-MOTEs at the university campus that exceeded air quality limits were Regent Court and Broad Lane (near St. George's Terrace). In the outskirts of the city, air quality limits were exceeded at three sites where AQMesh pods were installed: Saville Street, Abbeydale Rd and Brightside Lane. At these sites, the main source of emission is road traffic. Therefore, road traffic is the main source of emissions causing violation of AQ standards in Sheffield.

Figure 7.6 combines both DT and LCS, showing a more detailed map of interpolated NO_2 concentrations (μg/m^3). Generally, it demonstrates the same pattern as shown in Figure 7.4. However, in comparison to LUR (Figure 7.3) and Airviro models (Figure 7.2), the interpolated maps are less precise and less detailed. Briggs et al. [10] compared the performance of LUR with spatial interpolation methods (e.g., kriging, Triangular Interpolation Network (TIN) contouring and trend surface analysis) and reported that LUR performed much better than the interpolation techniques. The reason probably was that in urban areas, spatial variability of air pollutants was more controlled by the local emission sources, such as road traffic, rather than a smoothly varying field, which was assumed by the interpolation methods.

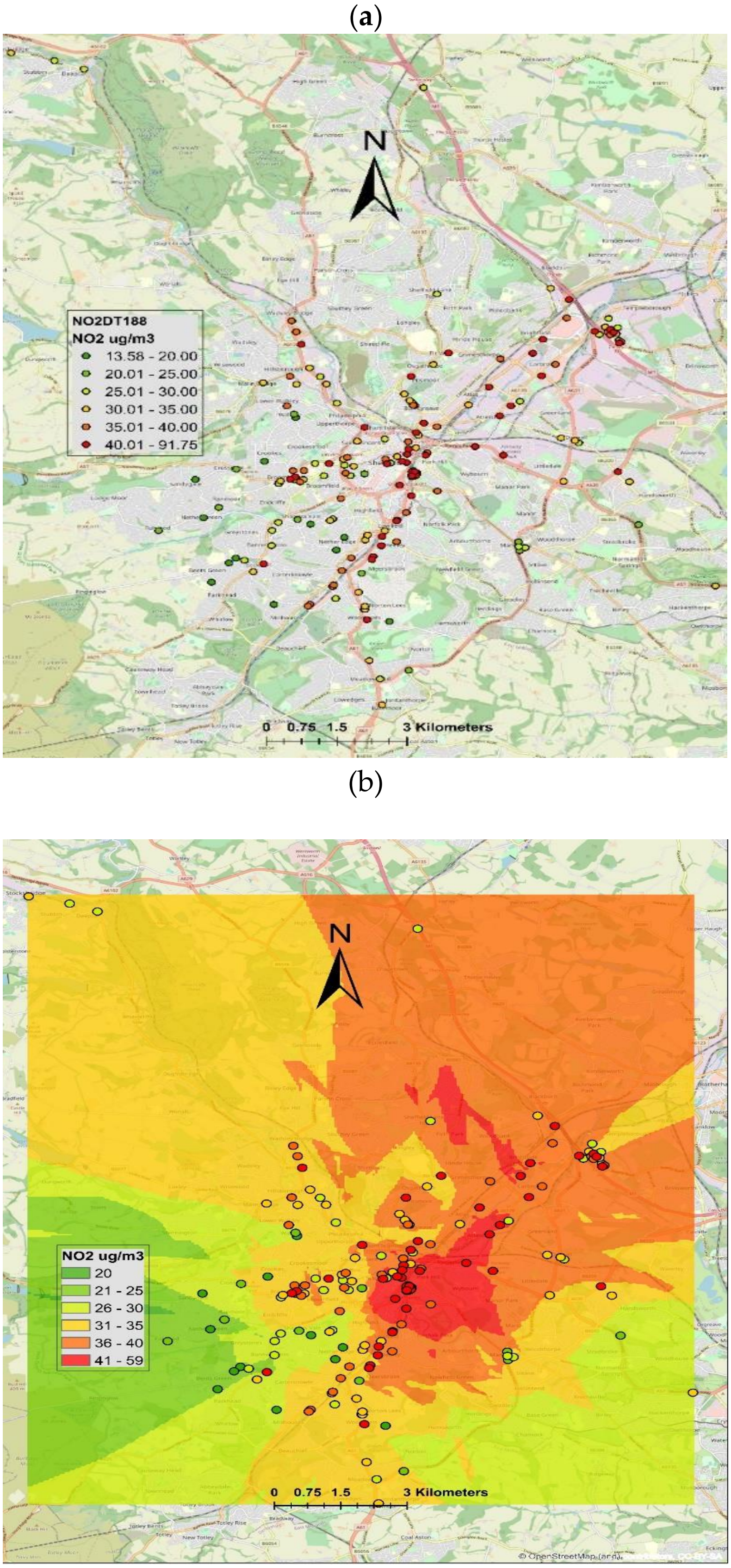
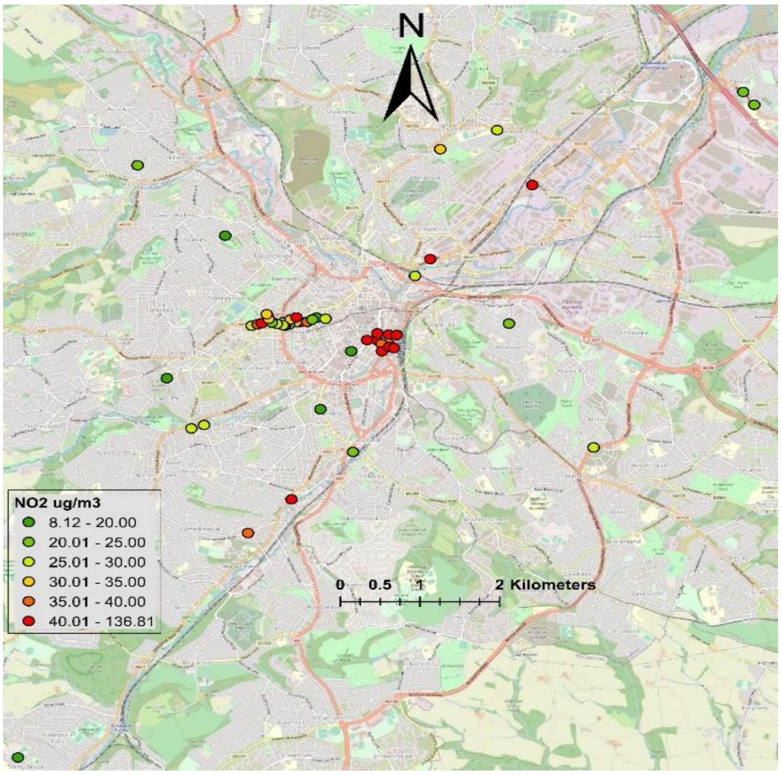


Figure 7.4. Showing point and interpolated NO_2 levels from DT network: (**a**) NO_2 concentrations (μg/m^3) from DT, and (**b**) interpolated NO_2 concentrations (μg/m^3) using ordinary kriging.

(**a**)



(b)

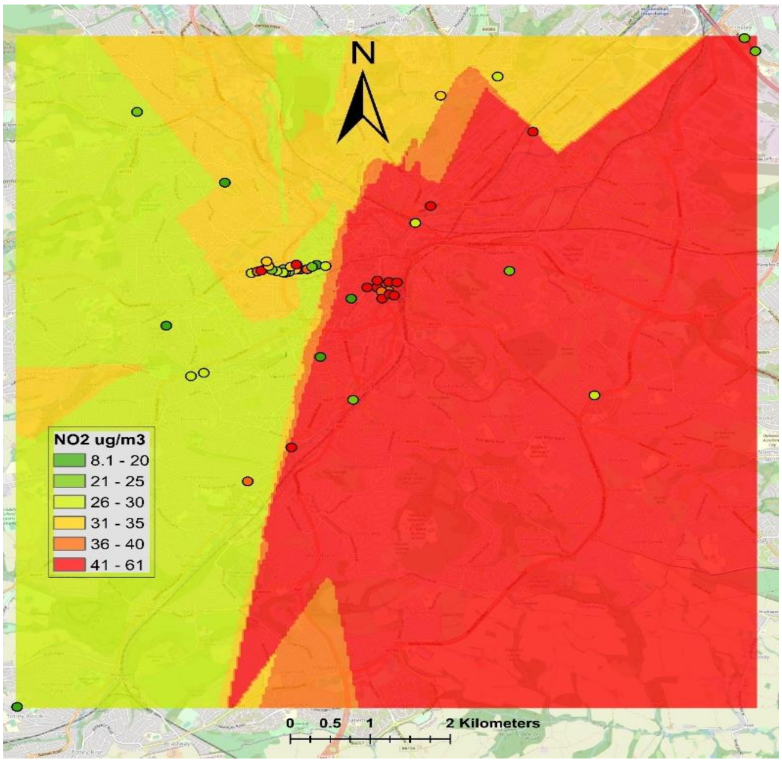


Figure 7.5. Showing point and interpolated NO2 levels from LCS network: (**a**) NO2 concentrations (μg/m^3) from LCS, (**b**) interpolated NO2 concentrations (μg/m^3) using ordinary kriging.

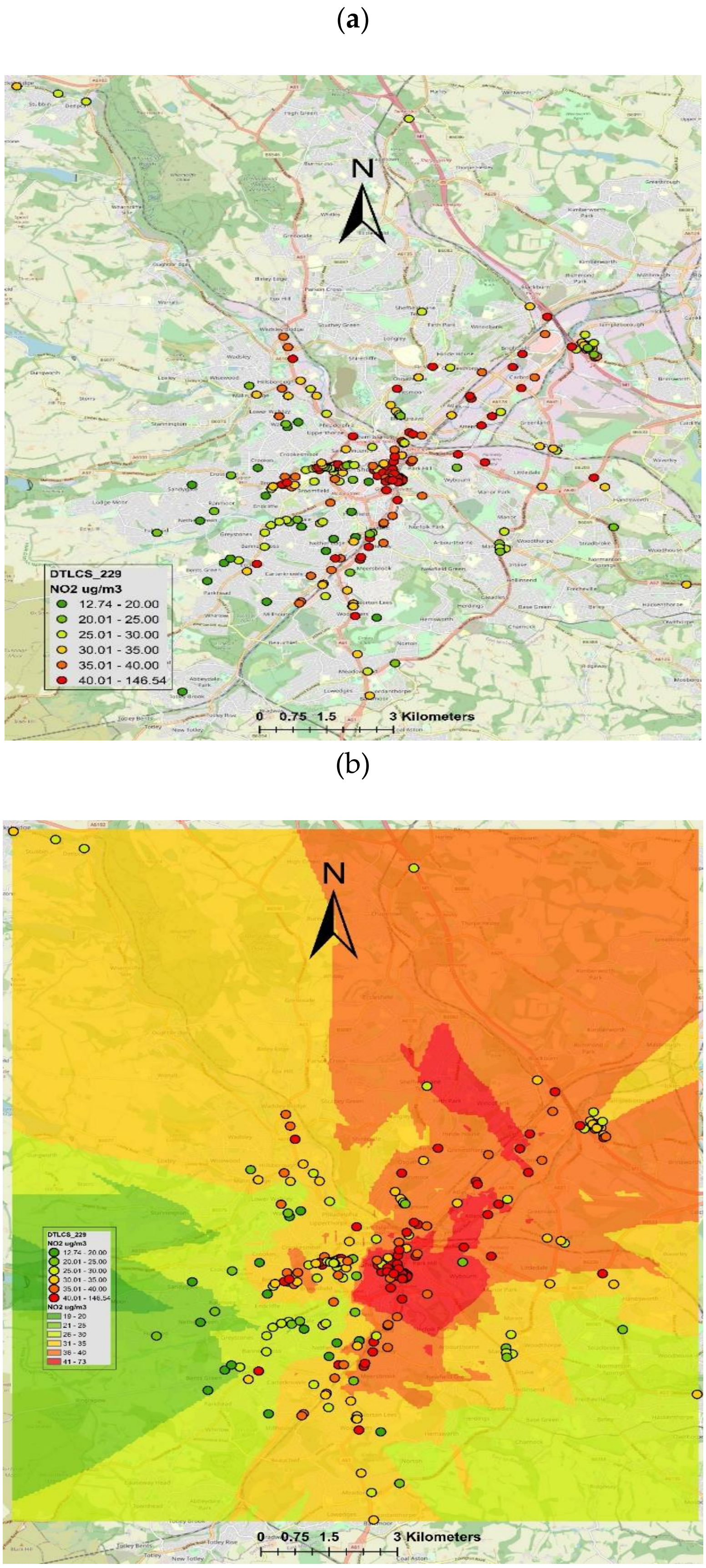


Figure 7.6. Showing point and interpolated NO2 levels from both DT and LCS: (**a**) NO2 concentrations (μg/m^3) from both 188 DT and LCS, and (**b**) interpolated NO2 concentrations (μg/m^3) using ordinary kriging.

7.3.2. Data Fusion—Fusing Model Estimations with Measured Concentrations

In this section, geostatistical universal kriging was used to fuse measured NO_2 concentrations with estimated NO_2 concentrations. NO_2 measured by DT, LCS and DTLCS are expressed as NO_2-DT, NO_2-LCS and NO_2-DTLCS and estimation of LUR and Airviro as NO_2-LUR and NO_2-Airviro, respectively.

7.3.2.1. Fusion of NO_2-LCS with NO_2-Airviro and NO_2-LUR

Universal kriging was used to fuse NO_2-LCS with NO_2-Airviro and NO_2-LUR. Figure 7.7 presents the results of data fusion, where the fusion of NO_2-Airviro with measured NO_2 is presented in Figure 7.7a and the fusion of NO_2-LUR with measured NO_2 is presented in Figure 7.7b. The resulting fused concentrations from the integration between NO_2-Airviro and NO_2-LCS (Airviro-LCS) ranged from 6.14 to 138.93 µg/m^3. The range of these values reflected the measured NO_2-LCS concentrations (ranging from 8.11 to 136.81 µg/m^3), whereas the spatial pattern followed the trend of NO_2-Airviro. On the other hand, the fusion of NO_2-LCS and NO_2-LUR (LUR-LCS) (Figure 7.7b) demonstrated a slightly different pattern. The NO_2 LUR-LCS ranged from 19.45 to 82.98 (µg/m^3), slightly overestimating lower values and underestimating the higher values. For model validation, the fused concentrations were compared with the measured concentrations using the testing dataset (25% randomly selected) applying various statistical metrics (Table 7.2). Overall, NO_2 LUR-LCS showed slightly better correlation (r, 0.96) and less error (RMSE, 9.09) than Airviro-LCS (r, 0.88 and RMSE, 18.16) when compared with measured concentrations. Overall, the fusion of Airviro-LCS slightly underestimated NO_2 concentrations, demonstrated by the negative value of MBE (−4.14).

Table 7.2. Showing the values of different statistical metrics calculated by comparing fused and measured NO_2 concentrations based on randomly selected testing dataset (train/test cross-validation). FAC2, MBE, MAE, RMSE and r stand for factor of two, mean biased error, mean absolute error, root mean squared error, and correlation coefficient.

Metrics	Airviro-LCS	LUR-LCS	Airviro-DT	LUR-DT	Airviro-DTLCS	LUR-DTLCS
FAC2	1	1	1	1	0.98	0.96
MBE	−4.14	1.44	1.56	1.40	1.08	2.24
MAE	12.79	7.99	5.81	5.29	8.20	3.73
RMSE	18.16	9.09	7.18	6.74	10.42	10.43
R	0.88	0.96	0.70	0.70	0.56	0.59

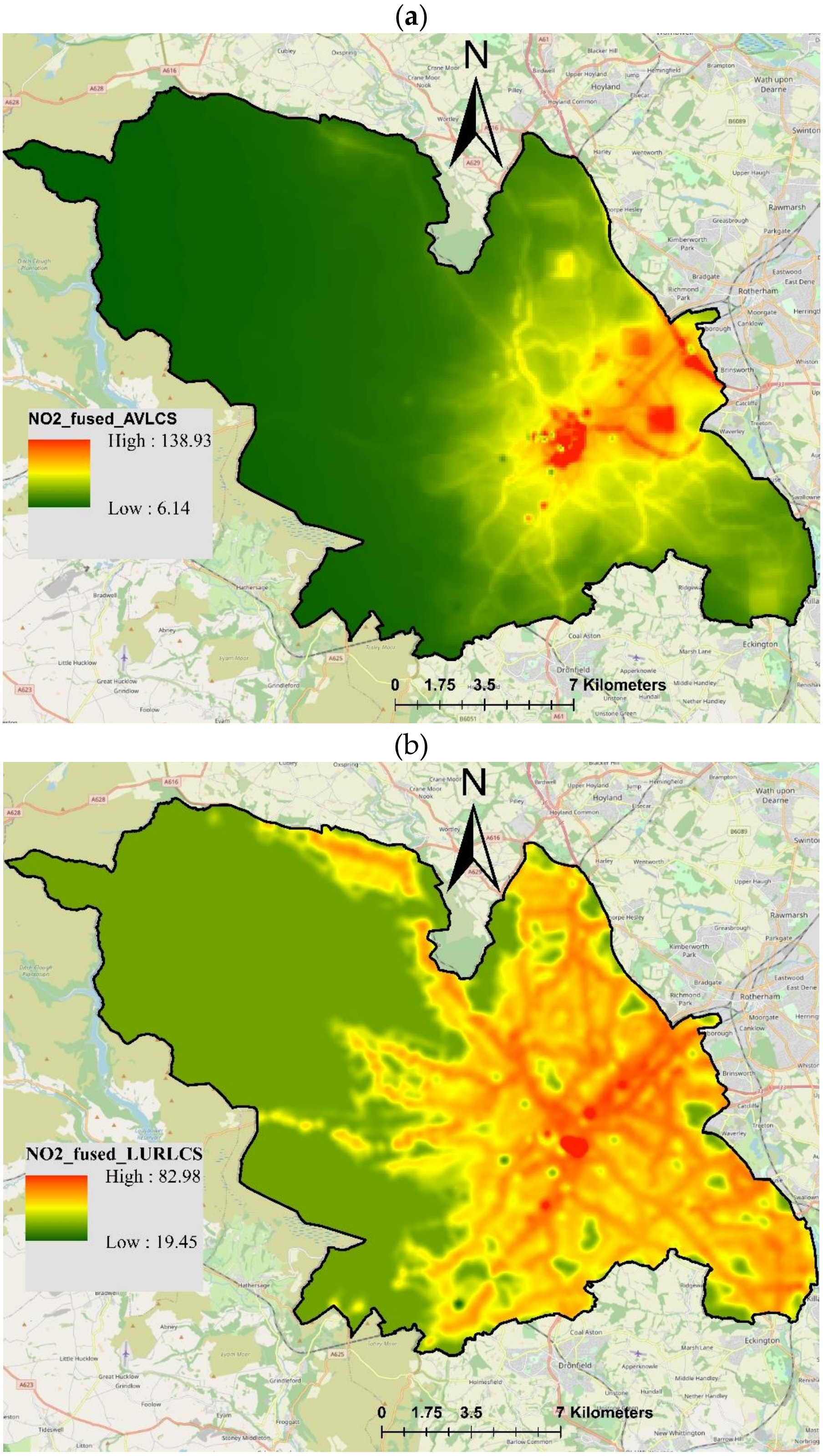


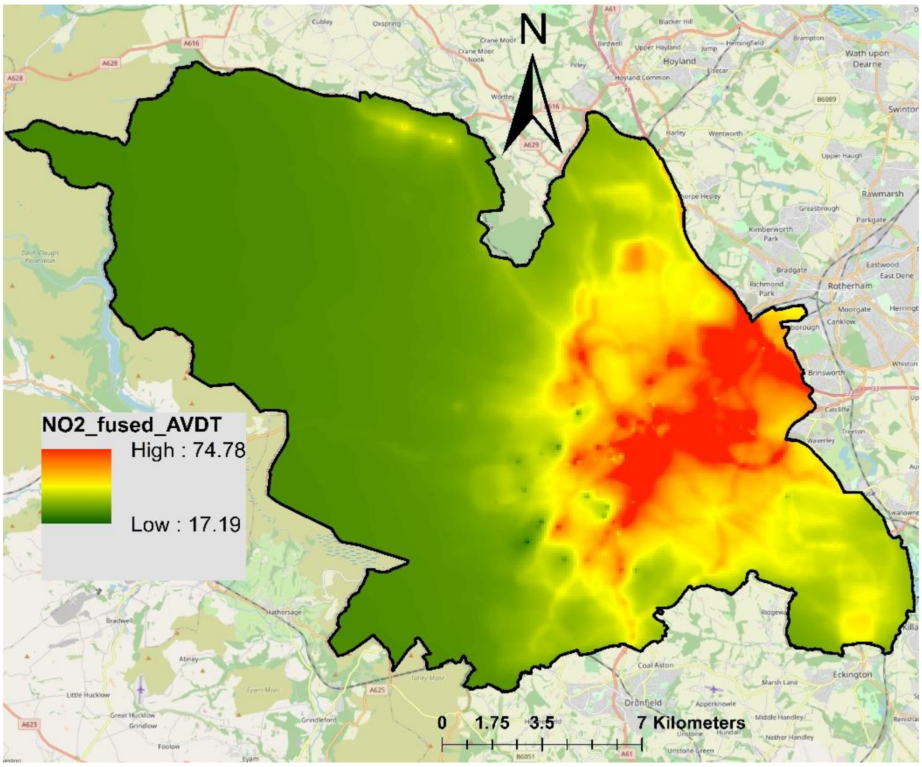
Figure 7.7. Resulting map of fused NO_2 concentrations (μg/m^3) using universal kriging techniques: (**a**) NO_2 Airviro-LCS; (**b**) NO_2 LUR-LCS.

7.3.2.2. Fusion of NO_2-DT with NO_2-Airviro and NO_2-LUR

Maps produced by the fusion of NO_2-DT with NO_2-Airviro and NO_2-LUR are presented in Figure 7.8, where the fusion of Airviro-DT is shown in Figure 7.8a and the fusion of LUR-DT is shown Figure 7.8b. The fused NO_2 Airviro-DT ranged from 17.19 to 74.78 µg/m^3. NO_2-Airviro concentrations ranged from 0.69 to 110, whereas NO_2-DT ranged from 13.58 to 91.75 µg/m^3. Airviro-DT showed higher concentrations in the city centre, east and northeast part of the city and relatively lower concentrations in the outskirts of the city, especially towards the west and northwest part of the city. Figure 7.8b shows NO_2 LUR-DT, which is the fusion between NO_2-LUR and NO_2-DT. Here, higher NO_2 LUR-DT are shown in the city centre, eastern and northeastern parts. However, some busy roads and hotspots are also highlighted in some other parts of the city. NO_2 LUR-DT ranged from 4.13 to 64.49 µg/m^3.

Statistical metrics used to compare measured concentrations with fused concentrations (Table 7.2) showed the same correlation coefficient for both LUR-DT and Airviro-DT (r, 0.70). The values of other metrics, e.g., RMSE and MAE, also demonstrated negligible differences. However, visually, LUR-DT produced more detailed maps and successfully highlighted higher pollution levels on several busy roads.

(**a**)



(b)

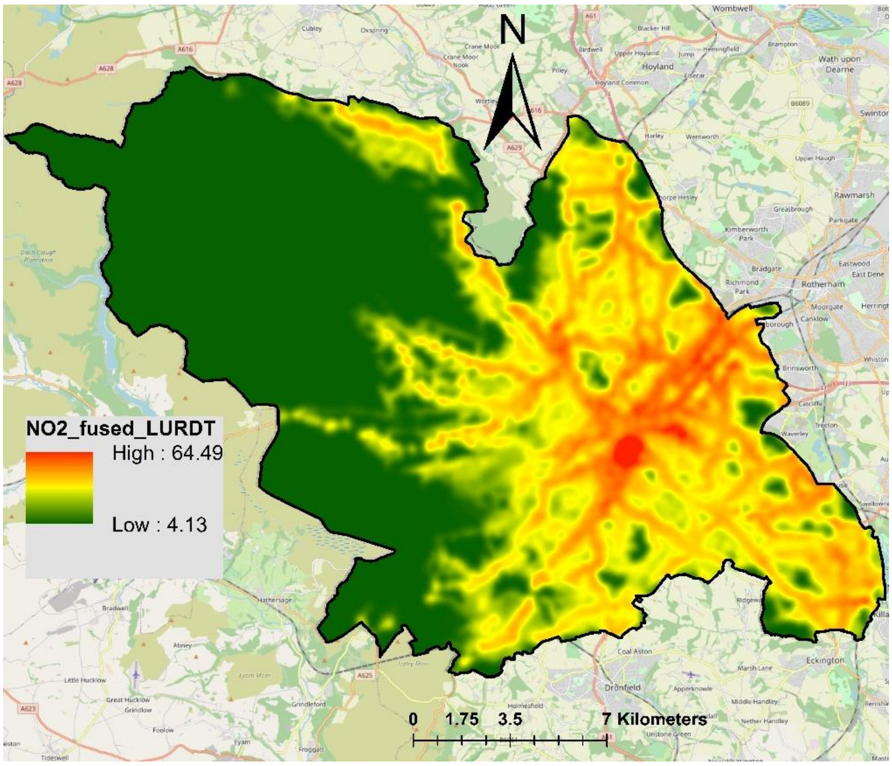


Figure 7.8. Resulting maps of fused NO_2 concentrations (µg/m^3) using universal kriging techniques: (**a**) NO_2 Airviro-DT; (**b**) NO_2 LUR-DT.

7.3.2.3. Fusion of NO_2-DTLCS with NO_2-Airviro and NO_2-LUR

Finally, the measured NO_2-DTLCS were fused with NO_2-Airviro and NO_2-LUR. The fused NO_2 concentrations are presented in Figure 7.9, where the fusion of Airviro-DTLCS is shown in Figure 7.9a and the fusion of LUR-DTLCS is shown in Figure 7.9b. The resulting maps shown in Figure 7.9 look similar to those presented in Figure 7.8 in terms of spatial coverage; however, actual values and their ranges are different. Fused Airviro-DTLCS ranged from 15.11 to 126.88, whereas LUR-DTLCS ranged from 19.45 to 82.98 µg/m^3. The measured NO_2-DTLCS ranged from 8.12 to 136.81 µg/m^3. Combining DT and LCS measurements did not improve the model performance compared to using only NO_2-DT. In contrast, the values of correlation coefficient slightly decreased to 0.56 and 0.59 for Airviro-DTLCS and LUR-DTLCS, respectively. Values of RMSE were 10.42 and 10.43 for Airviro-DTLCS and LUR-DTLCS, respectively (Table 7.2). This shows that increasing the number of sensors will not necessarily improve the output of the data fusion, especially if the sensors are not the same type.

The values of measured NO_2 concentrations are reflected in fused values, whereas the spatial trend is determined by the model values. The fused maps provided better coverage and more realistic concentrations than the interpolated maps using ordinary kriging. Data fusion also improved the NO_2-Airvrio and NO_2-LUR maps, producing more realistic maps based on both measured and estimated concentrations. Overall, the fusion of LUR with measured concentrations produced better results than the Airviro model in terms of correlation coefficient, RMSE and MAE.

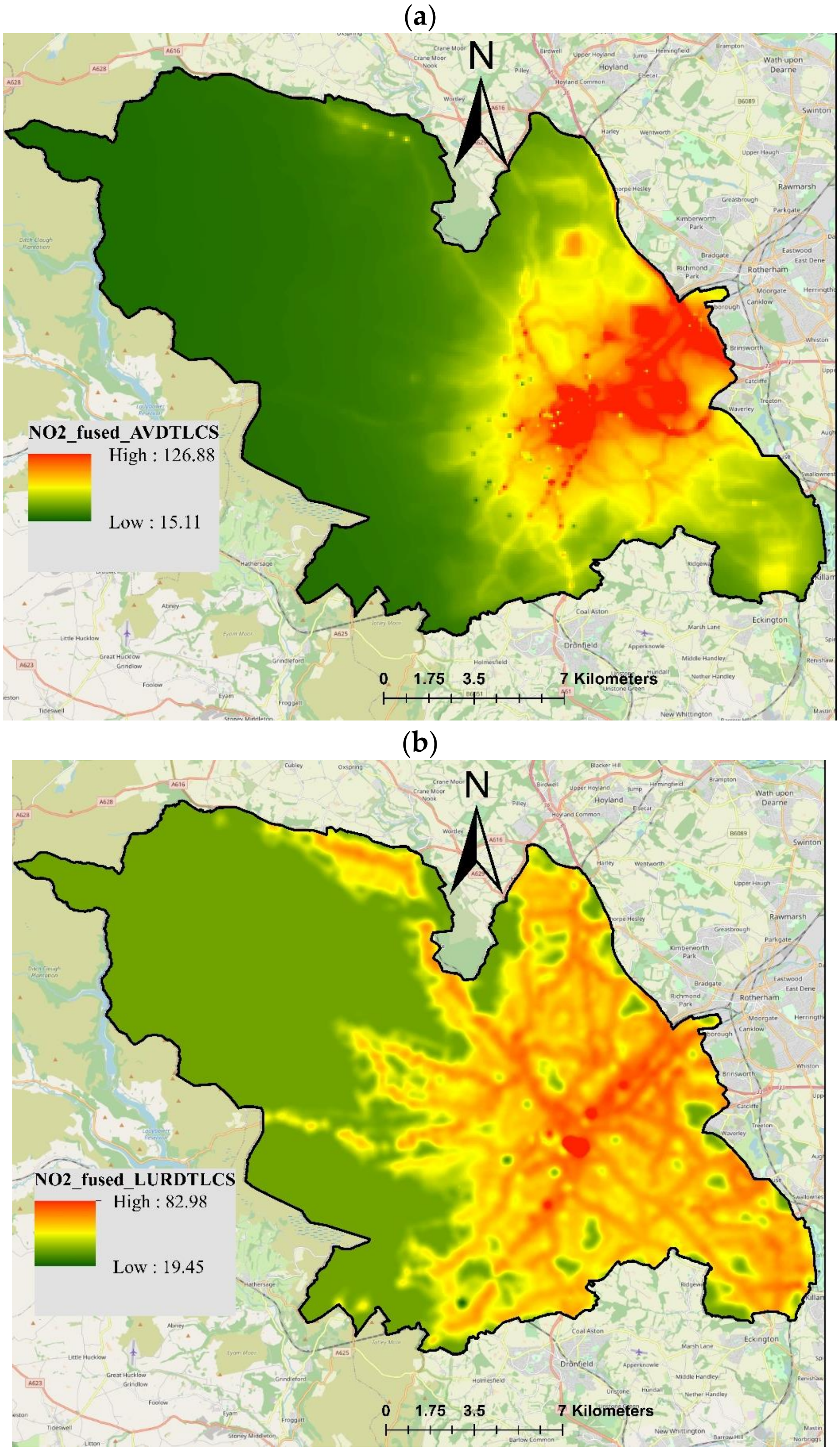


Figure 7.9. Resulting map of fused NO_2 concentrations (μg/m^3) using universal kriging techniques: (**a**) Airviro-DTLCS; (**b**) LUR-DTLCS.

In this study, different monitoring and modelling approaches were used to produce high-resolution spatial maps of NO_2 concentrations in urban areas. Air quality monitoring is the more accurate source of data for assessing air pollutant levels. However, the air quality monitoring network cannot be dense enough to provide measured data for each spatial point in the whole city. Therefore, different modelling approaches (e.g., spatial interpolations, dispersion models and LUR models) are required to predict air quality information between monitoring stations. These models, in addition to measured data, use emission data, meteorology data, traffic characteristics, population data and land use characteristics. One of the issues with air quality modelling is its level of uncertainty, which is significantly higher than the measured data [27]. This is where data fusion techniques play their role by combining measured and modelled data in such a way to improve the spatiotemporal resolution of the measured data and accuracy of the modelled data.

Huang et al. [28] applied data fusion techniques to integrate measured and model estimation over North Carolina, USA. In contrast to this study, Huang et al. [28] analysed the levels of several air pollutants which were $PM_{2.5}$, CO, NOx, NO_2 and five particulate species (organic carbon, elemental carbon, and sulphate, nitrate and ammonium ions). They reported that the application of data fusion reduced biases in the model estimation. The correlation coefficient for the cross-validation test was 0.91 in Huang et al. [28], which was less than this study (r-value 0.96). Gressent et al. [29] analysed air quality data from low-cost sensors and integrated them with the estimation of ADMS-Urban in France. In contrast to this study, which uses annual NO_2 concentrations, Gressent et al. [29] used hourly data and estimated PM_{10} concentrations for 29 November 2018 from 7 a.m. to 7 p.m. They used external drift kriging, which is similar to universal kriging, and reduced the bias from 8% to 2% when considering LCS observations instead of the model alone. Schneider et al. [9] employed geostatistics methodology to fuse NO_2 concentrations obtained from a network of low-cost sensors (AQMesh) and EPISODE dispersion model. EPISODE is a 3-D Eulerian/Lagrangian dispersion model providing atmospheric air pollutant forecasts at urban and regional scales. For more details on the EPISODE model [30]. Schneider et al. [9] evaluated the geostatistics universal kriging methodology for fusing both measured and predicted data of NO_2 in Oslo, Norway during January 2016. Results showed that the fusing method was able to produce realistic hourly NO_2 concentrations, which inherited the spatial trend of the pollutant from the EPISODE model. Furthermore, fused data were compared to measured data from a reference instrument and results showed reasonably good resemblance between measured and fused data, with R^2 value of 0.89 and mean squared error of 14.3 $\mu g/m^3$. Their model showed slightly inferior performance in comparison to this study, with r-value 0.96 and RMSE 9.09. Furthermore, Schneider et al. [9] had used a shorter time-period of one month and fewer sensors as compared to this study, which used NO_2 data for a whole year from more sensors.

Most of the data fusion techniques mentioned above are applied on city scales. However, several researchers have also applied the data fusion approaches to air quality data on a country level [31] or global level [32]. Liang et al. [31] developed a data fusion method and compared the outputs with kriging-with-external-drift (KED) and the chemistry module from WRF-Chem (Weather Research and Forecast Model with Chemistry Module). KED is a type of universal kriging that takes into account the local trend of the variable (e.g., air pollutant concentrations) as well as external drift (a spatial trend) when minimising the variance of estimation [8]. Both KED and WRF-Chem were used to estimate daily $PM_{2.5}$ levels in 10-km grid cells in China during 2013. The estimated concentrations from both KED and WRF-Chem were then fused with measured observations. For fusion, a simple linear regression model was applied between the observed and estimated concentrations for both KED and chemistry models in turns. The regression coefficients obtained from the regression model were used to adjust the predicted concentrations. The performance of the models was evaluated, and KED and regression data fusion methods showed better performance in terms of R^2 value of 0.95 and 0.94, respectively, as compared to the value of 0.51 for WARF-Chem model. Shaddick et al. [32], employing a Bayesian hierarchical modelling framework, estimated that 92% of the world's population lived in areas where air pollution levels exceeded the World Health Organization's air quality guidelines. However, the results of these investigations carried out on a country or global level are not comparable with studies conducted on a city level.

7.4. Conclusions

In this paper, different modelling and data fusion techniques were employed to analyse the spatial variability of NO_2 concentrations (μg/m^3) in the city of Sheffield. NO_2 was monitored by 188 DT, 41 low-cost sensors, 3 AURN and 5 SCC monitoring stations. The main difference between LCS and reference sensors is that LCS are cheaper, compact and their measurements are less reliable, whereas AURN and SCC use reference sensors which are much more expensive and reliable. However, the reference AQMN is sparse, having fewer AQMS, which are not enough for understanding the spatial variability of NO_2 in Sheffield. Therefore, the networks of LCS and DT were used for analysing the spatial variability of NO_2 concentrations and validating the models. Air quality standards were exceeded at several locations in Sheffield, particularly in the city centre and on some busy roads, where annual mean NO_2 levels were higher than 40 μg/m^3. The highest levels of NO_2 were recorded by the Envirowatch E-MOTE installed next to the Sheffield train station taxi rank (136.81 μg/m^3), followed by the Sheaf Street/Sheaf Square pedestrian crossing (115.56 μg/m^3) and Arundel Gate (107.17 μg/m^3).

Three modelling approaches were used to produce maps of NO_2 concentrations in Sheffield: (a) geostatistical kriging interpolation, (b) Airviro dispersion modelling, and (c) LUR based on the generalised additive model. Measured NO_2 concentrations were fused with the estimated concentrations using universal kriging, which is an advanced kriging technique that is employed for data having more than one correlated variable. Six sets of measured and estimated NO_2 data were fused: (i) Fusion of NO_2-Airviro with NO_2-LCS; (ii) Fusion of NO_2-Airviro with NO_2-DT; (iii) Fusion of NO_2-Airviro with NO_2-DTLCS; (iv) Fusion of NO_2-LUR with NO_2-LCS; (v) Fusion of NO_2-LUR with NO_2-DT; and (vi) Fusion of NO_2-LUR with NO_2-DTLCS. Fused NO_2 were compared with measured concentrations in terms of different statistical metrics including FAC2, r, RMSE, MAE and MBE. Maps produced by the fusion of NO_2-LCS and NO_2-LUR produced better results, with an r-value of 0.96, followed by the fusion of NO_2-Airviro with NO_2-LCS, with an r-value of 0.88. Fused maps developed by universal kriging provided better spatial coverage and similarity with measured concentrations than the ordinary kriging interpolation, Airviro and LUR models. Fused maps combining measured and estimated concentrations produced more realistic concentrations and provided better spatial coverage.

This study presents a geostatistical universal kriging approach for fusing measured and estimated NO_2 concentrations to improve our understanding of small-scale spatial variability in NO_2 concentrations in Sheffield. This approach adds value to both measured and estimated values: the measured concentrations are improved by predicting spatiotemporal gaps, whereas the modelled concentrations are improved by constraining them with observed data. The main findings of this study are: (a) The universal kriging approach was capable of estimating realistic NO_2 concentration maps from the fusion of measured and modelled concentrations. The fused NO_2 concentrations inherited spatial patterns of the pollutant from the model estimations and adjusted the modelled values using the measured concentrations. (b) A huge number of sensors are required to provide reasonable spatial coverage at a city level, which is too expensive. Spatial modelling (e.g., dispersion model and LUR model) and data fusion approaches can provide city-level maps by integrating pollutant concentrations measured by the AQMS with modelled concentrations. (c) According to Schneider et al. [9], the accuracy of the data fusion depends on the number of sensors, their spatial distribution, their uncertainty and the accuracy of the estimated values to be fused with measured values. Here, we showed that in addition, the accuracy of the data fusion also depends on the uniformity of the sensors, meaning that using the same type of sensors can result in better accuracy. Increasing the number of sensors will not necessarily improve the model outputs, especially if sensors are not of the same type. For example, maps produced by the fusion of NO_2-LCS with NO_2-LUR (r-value 0.96) and NO_2-Airviro (r-value 0.88) produced better results than when both NO_2-DT and NO_2-LCS were used together and fused with NO_2-LUR or NO_2-Airviro, when the r-values decreased to 0.59 and 0.56, respectively.

The uniqueness of this study is that it uses estimated NO_2 concentrations from two sources—a dispersion model and a land use regression model—and measured NO_2 concentrations from three sources: diffusion tubes, reference sensors and low-cost sensors. The study applies geostatistical universal kriging for data fusion and produces high-resolution maps of NO_2 in Sheffield.

Limitations/weaknesses: (a) NO_2 diffusion tube data were downloaded from the Sheffield City Council website; we do not know which correction factors were applied before they were published online. (b) Low-cost sensors have their limitations and are not as accurate as reference sensors. The data from the low-cost sensors are postprocessed by the manufacturers with the aim of correcting cross-interferences as well as the effects of temperature and relative humidity. The AQMesh pods include an O_3-filtered NO_2 sensor from Alphasense, which is designed to reject O_3 and hence eliminate cross-sensitivity issues. However, none of these sensors was collocated with the reference sensor in the field; therefore, we could not develop and apply correction factors.

In the future, we aim to apply this approach to high-resolution spatiotemporal data (e.g., hourly data on city scale) to further improve spatiotemporal estimation of NO_2 concentrations. Furthermore, the data fusion work would be done in as near to real time as possible in the form of an app, so as to provide public health advice.

Author Contributions:

Conceptualization, SM; Formal analysis, SM; Funding acquisition, MM and DC; Methodology, SM; Supervision, MM and DC; Visualization,SM; Writing – original draft, SM. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Engineering and Physical Sciences Research Council (EPSRC) (grant number—EP/R512175/1) and Siemens plc. The APC was paid by UKRI/EPSRC.

Institutional Review Board Statement:

Not applicable

Informed Consent Statement:

Not applicable

Data Availability Statement: Data used this paper mostly come from the two papers already published and cited in the reference list [12, 13].

Acknowledgments: We would like to express our thanks to the Engineering and Physical Sciences Research Council (EPSRC), grant number—EP/R512175/1 and Siemens plc for funding this project. We are also thankful to Sheffield City Council for providing NO_2 diffusion tubes data.

Conflicts of Interest: All the authors declare no conflict of interest.

7.5. References

1. Public Health England. Guidance, Air pollution: Applying All Our Health. 2020. Available online: https://www.gov.uk/government/publications/air-pollution-applying-all-our-health/air-pollution-applying-all-our-health (accessed on 19 January 2021).
2. Fan, Z.; Pun, V.C.; Chen, X.C.; Hong, Q.; Tian, L.; Ho, S.S.H.; Lee, S.C.; Tse, L.A.; Ho, K.F. Personal exposure to fine particles ($PM_{2.5}$) and respiratory inflammation of common residents in Hong Kong. *Environ. Res.* **2018**, *164*, 24–31.
3. WHO. Review of Evidence on Health Aspects of Air Pollution-REVIHAAP Project: Final Technical Report. World Health Organziation Regional Office for Europe. 2013. Available online: https://www.euro.who.int/__data/assets/pdf_file/0004/193108/REVIHAAP-Final-technical-report-final-version.pdf (accessed on 28 January 2021).
4. Landrigan, P.J. Air pollution and health. *Lancet Public Health* **2016**, *2*, 4–5, doi:10.1016/S2468-2667(16)30023-8.
5. DEFRA. Improving Air Quality in the UK Tackling Nitrogen Dioxide in Our Towns and Cities, UK Overview Document, December 2015. Available online: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/486636/aq-plan-2015-overview-document.pdf (accessed on 9 April 2020).

6. Xie, X.; Semanjski, I.; Gautama, S.; Tsiligianni, E.; Deligiannis, N.; Rajan, R.T.; Pasveer, F.; Philips, W. A Review of Urban Air Pollution Monitoring and Exposure Assessment Methods. *ISPRS Int. J. Geo. inf.* **2017**, *6*, 389, doi:10.3390/ijgi6120389.
7. Briggs, D.J. The Role of GIS: Coping With Space (And Time) in Air Pollution Exposure Assessment. *J. Toxicol. Environ. Health* **2005**, *68*, 1243–1261.
8. Hengl, T.; Heuvelink, G.; Stein, A. *Comparison of Kriging with External Drift and Regression-Kriging*; ITC Technical note; International Institute for Geo-Information Science and Earth Observation (ITC): Enschede, The Netherlands, 2003.
9. Schneider, P.; Castell, N.; Vogt, M.; Dauge, F.R.; Lahoz, W.A.; Bartonova, A. Mapping urban air quality in near real-time using observations from lowcost sensors and model information. *Environ. Int.* **2017**, *106*, 234–247.
10. Briggs, D.J.; de Hough, C.; Gulliver, J.; Wills, J.; Elliott, P.; Kingham, S.; Smallbone, K. A regression-based method for mapping traffic-related air pollution: Application and testing in four contrasting urban environments. *Sci. Total Environ.* **2000**, *253*, 151–167.
11. Castanedo, F. A Review of Data Fusion Techniques. *Sci. World J.* **2013**, *2013*, 1–19, doi:10.1155/2013/704504.
12. Munir, S.; Mayfield, M.; Coca, D.; Mihaylova, L.S. A nonlinear land-use regression approach for modelling NO_2 concentrations in urban areas—Using data from low-cost sensors and diffusion tubes. *Atmosphere* **2020**,*11*, 736, doi:10.3390/atmos11070736.
13. Munir, S.; Mayfield, M.; Coca, D.; Mihaylova, L.S.; Osammor, O. Analysis of air pollution in urban areas with Airviro dispersion model—A Case Study in the City of Sheffield, United Kingdom. *Atmosphere* **2020**, *11*, 285, doi:10.3390/atmos11030285
14. Hiemstra, P. Automatic Interpolation Package. "Automap", Version 1.0-14, a Package for R Programming Language. 2015. Available online: https://cran.r-project.org/web/packages/automap/automap.pdf (accessed on: 28/01/2021).
15. Briggs, D.J.; Collins, S.; Elliott, P.; Fischer, P.; Kingham, S.; Lebret, E.; Pryl, K.; van Reeuwijk, H.; Smallbone, K.; van der Veen, A. Mapping urban air pollution using GIS: A regression-based approach. *Int. J. Geogr. Inf. Sci.* **1997**, *11*, 699–718.
16. Beelen, R.; Hoek, G.; Fischer, P.; van den Brandt, P.A.; Brunekreef, B. Estimated long-term outdoor air pollution concentrations in a cohort study. *Atmos. Environ.* **2007**, *41*, 1343–1358.
17. Eeftens, M.; Beelen, R.; de Hoogh, K. Development of land use regression models for $PM_{2.5}$, $PM_{2.5}$ absorbance, PM_{10} and PM coarse in 20 European Study areas; results of the ESCAPE project. *Environ. Sci. Technol.* **2012**, *46*, 11195–11205.
18. Hoek, G.; Beelen, R.; De Hoogh, K.; Vienneau, D.; Gulliver, J.; Fischer, P. A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmos. Environ.* **2008**, *42*, 7561–7578.
19. Lee, J.-H.; Wu, C.-F.; Hoek, G.; De Hoogh, K.; Beelen, R.; Brunekreef, B.; Chan, C.-C. Land use regression models for estimating individual NO_x and NO_2 exposures in a metropolis with a high density of traffic roads and population. *Sci. Total. Environ.* **2014**, *472*, 1163–1171.
20. Muttoo, S.; Ramsay, L.; Brunekreef, B.; Beelen, R.; Meliefste, K.; Naidoo, R.N. Land use regression modelling estimating nitrogen oxides exposure in industrial south Durban, South Africa. *Sci. Total Environ.* **2017**, *610–611*, 1439–1447.
21. Rahman, M.M.; Yeganeh, B.; Clifford, S.; Knibbs, L.D.; Morawska, L. Development of a land use regression model for daily NO_2 and NOx concentrations in the Brisbane metropolitan area, Australia. *Environ. Modell. Softw.* **2017**, *95*, 168–179.
22. Stedman, J.; Vincent, K.; Campbell, G.; Goodwin, J.; Downing, C. New high resolution maps of estimated background ambient NOx and NO_2 concentrations in the U.K. *Atmos. Environ.* **1997**, *31*, 3591–3602.
23. Ryan, P.H.; LeMasters, G.K.; Biswas, P.; Levin, L.; Hu, S.; Lindsey, M.; Bernstein, D.I.; Lockey, J.; Villareal, M.; Hershey, G.K.K.; et al. A comparison of proximity and land use regression traffic exposure models and wheezing in infants. *Environ. Health Perspect.* **2007**, *115*, 278–284.
24. Ryan, P.H.; LeMasters, G.K. A review of land-use regressionmodels for characterizing intraurban air pollution exposure. *Inhal. Toxicol.* **2007**, *19* (Suppl. 1), 127–133.
25. Goovaerts, P. Geostatistics for Natural Resources Evaluation. Applied Geostatistics Series, 1997, xiv, 483 pp. New York, Oxford: Oxford University Press. ISBN 0 19 511538 4.
26. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2019. Available online: https://www.R-project.org/ (accessed on 28 January 2021).
27. Denby, B.; Garcia, V.; Holland, D.M.; Hogrege, C. Integration of air quality modeling and monitoring data for enhanced health exposure assessment. *Air Waste Manag. Assoc.* **2009**, 46–49.
28. Huang, R.; Zhai, X.; Ivey, C.E.; Friberg, M.D.; Hu, X. Using Air Quality Model-Data Fusion Methods for Developing Air Pollutant Exposure Fields and Comparison with Satellite AOD-Derived Fields: Application over North Carolina, USA. In *Air Pollution Modeling and Its Application XXV*; Mensink, C., Kallos, G., Eds.; Springer Proceedings in Complexity book series; Springer: Cham, Switzerland, 2018; doi:10.1007/978-3-319-57645-9_33.
29. Gressent, A.; Malherbe, L.; Colette, A.; Rollin, H.; Scimia, R. Data fusion for air quality mapping using low-cost sensor observations: Feasibility and added-value, *Environ. Int.* **2020**, *143*, 105965, doi:10.1016/j.envint.2020.105965.
30. Slordal, L.H.; Walker, S.E.; Solberg, S.S. *The Urban Air Dispersion Model EPISODE Applied in AirQUIS 2003—Technical Description*; NILU—Norwegian Institute for Air Research: Kjeller, Norway, 2003

31. Liang, F.; Gao, M.; Xiao, Q.; Carmichael, G.R.; Pan, X.; Liu, Y. Evaluation of a data fusion approach to estimate daily PM2.5 levels in North China. *Environ. Res.* **2017**, *158*, 54–60.
32. Shaddick, G.; Thomas, M.L.; Green, A.; Brauer, M.; van-Donkelaar, A.; Burnett, R. Data integration model for air quality: A hierarchical approach to the global estimation of exposures to ambient air pollution. *J. R. Stat. Soc. Ser. Appl. Stat*. **2018**, *67*, 231–253.

CHAPTER 8: APPLICATION OF DENSITY PLOTS AND TIME SERIES MODELLING TO THE ANALYSIS OF NITROGEN DIOXIDES MEASURED BY LOW-COST AND REFERENCE SENSORS IN URBAN AREAS

Said Munir[a,]*, Martin Mayfield[a],

[a]Department of Civil and Structural Engineering, the University of Sheffield, Sheffield, S1 3JD, UK
*corresponding author (smunir2@sheffield.ac.uk), Mob: +447986001328, Fax: +44 (0) 114 222 5700

Published: Nitrogen 2021, 2(2), 167-195
https://doi.org/10.3390/nitrogen2020012
Received: 5 February 2021; Revised: 26 March 2021; Accepted: 8 April 2021; Published: 13 April 2021.

Abstract

Temporal variability of NO_2 concentrations measured by 28 Envirowatch E-MOTEs, 13 AQMesh pods, and eight reference sensors (five run by Sheffield City Council and three run by the Department for Environment, Food and Rural Affairs (DEFRA)) was analysed at different time scales (e.g., annual, weekly and diurnal cycles). Density plots and time variation plots were used to compare the distributions and temporal variability of NO_2 concentrations. Long-term trends, both adjusted and non-adjusted, showed significant reductions in NO_2 concentrations. At the Tinsley site, the non-adjusted trend was −0.94 (−1.12, −0.78) µgm^{-3}/year, whereas the adjusted trend was −0.95 (−1.04, −0.86) µgm^{-3}/year. At Devonshire Green, the non-adjusted trend was −1.21 (−1.91, −0.41) µgm^{-3}/year and the adjusted trend was −1.26 (−1.57, −0.83) µgm^{-3}/year. Furthermore, NO_2 concentrations were analysed employing univariate linear and nonlinear time series models and their performance was compared with a more advanced time series model using two exogenous variables (NO and O_3). For this purpose, time series data of NO, O_3 and NO_2 were obtained from a reference site in Sheffield, which were more accurate than the measurements from low-cost sensors and, therefore, more suitable for training and testing the model. In this article, the three main steps used for model development are discussed: (i) model specification for choosing appropriate values for p, d and q, (ii) model fitting (parameters estimation), and (iii) model diagnostic (testing the goodness of fit). The linear auto-regressive integrated moving average (ARIMA) performed better than the nonlinear counterpart; however, its performance in predicting NO_2 concentration was inferior to ARIMA with exogenous variables (ARIMAX). Using cross-validation ARIMAX demonstrated strong association with the measured concentrations, with a correlation coefficient of 0.84 and RMSE of 9.90. ARIMAX can be used as an early warning tool for predicting potential pollution episodes in order to be proactive in adopting precautionary measures.

Keywords: ARIMA; ARIMAX; air quality modelling; low-cost sensors; air pollution; nonlinear modelling; time series analysis; autoregressive; moving average; Sheffield.

8.1. Introduction

Air pollution is one of the most serious environmental threats to health, killing 6.4 million people in 2015 worldwide both in developed and less-wealthy nations [1]. Out of these, 2.8 million deaths were caused by indoor air pollution and 4.2 million deaths by outdoor air pollution. Air pollution is causing various health problems including respiratory problems, cardiovascular diseases, lung cancer and asthma [2]. Walters and Ayres [3] have also reported that air pollution, especially particulate matter and nitrogen dioxide (NO_2) pollution, may cause premature deaths and hospital admissions for conditions such as cardiovascular problems, allergic reactions and lung cancer. It is reported that exposure to air pollution is particularly harmful for children, people with existing health problems and the elderly [4–6]. Evidence suggests that the negative impacts of air pollution are dependent on the levels of air pollutants and length of exposure, higher levels and longer exposure resulting in more severe adverse effects [3,6]. Furthermore, air pollution may reduce visibility, damage historical buildings and monuments, affect vegetation and reduce crop yield and quality [4–7].

Air quality modelling is carried out for several purposes, including air quality prediction, quantifying the impacts of air pollution, modelling the impacts of various factors on air pollution, modelling pollution processes and transport, running and testing emission scenarios, modelling the dispersion of air pollutants in the atmosphere, quantifying the emissions of air pollutants from various emission sources, determining long-term trends in air pollutant concentrations and producing high-resolution spatiotemporal maps of air pollution [6–15]. In this study, the main purposes were to compare the performance of different time series models and to develop a time series model for predicting future NO_2 concentrations in Sheffield, UK. Time series models are one of the popular tools for predicting the future by understanding the past changes in air pollution concentrations [16].

Time series modelling is a useful tool for data analysis and has several benefits, which include data cleaning, data understanding and forecasting [17]. Time series modelling can help us filter out the noise and reveal the true signal in a dataset. Once a dataset is cleaned and the time series is divided into its different components, it helps us understand the true nature of the dataset. Finally, like other modelling approaches, time series modelling helps us predict future levels with the help of present and past levels of the time series (here NO_2 concentrations) [17]. To build a time series model the time series data must be stationary. For a time series to be stationary, the mean, variance and covariance of the time series should not be a function of time [18]. If the time series is not stationary, it should be stationarised first before fitting a model to it [18]. To stationarise a time series, it should be detrended and deseasonalised using differencing and power transformations [16]. A time series model can be linear or nonlinear depending on the relationship between current and past observations [16]. The Autoregressive (AR) and moving average (MA) models are the two widely used linear time-series models [19]. The AR and MA models can be combined to form the autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) models [20]. To model and predict a seasonal time series, the seasonal autoregressive integrated moving average (SARIMA) model is used, which is a variation of ARIMA [19]. ARIMA along with its various variations are also known as the Box–Jenkins models because they are based on the Box–Jenkins principle [19]. Classification of time series models is shown in Figure 8.1 [21–23].

Although air pollutant levels have decreased in the UK and Europe during the last decade or so, some air pollutant levels still exceed air quality standards in many urban areas [24] and

620 air quality management areas (AQMAs) have been declared across the UK [25]. Out of these areas, five cities (Birmingham, Leeds, Southampton, Derby and Nottingham) will not achieve the targets until 2025 and London will not comply until 2030 if additional control measures are not taken [24]. Most of the exceedances are due to the high levels of NO_2 and PM_{10} [26], emphasising that these air pollutants are a serious problem in urban areas. Further actions are required to cut emissions and understand the main drivers of high levels of NO_2 and PM_{10} in urban areas. Air pollution modelling is a part of these actions.

In first part of this article, NO_2 concentrations measured by different grades of sensors are compared graphically, whereas, in the second part, a comparison of time series modelling approaches is made using measured time series data of NO_2 collected at an urban monitoring site in Sheffield. A comparison is made between linear and nonlinear time series models and the better performing model is then compared with the ARIMAX model, which, in addition to persistence, uses exogenous variables. The aim is to assess the suitability of time series models for NO_2 prediction and choose the best time series model for air quality modelling in the urban environment

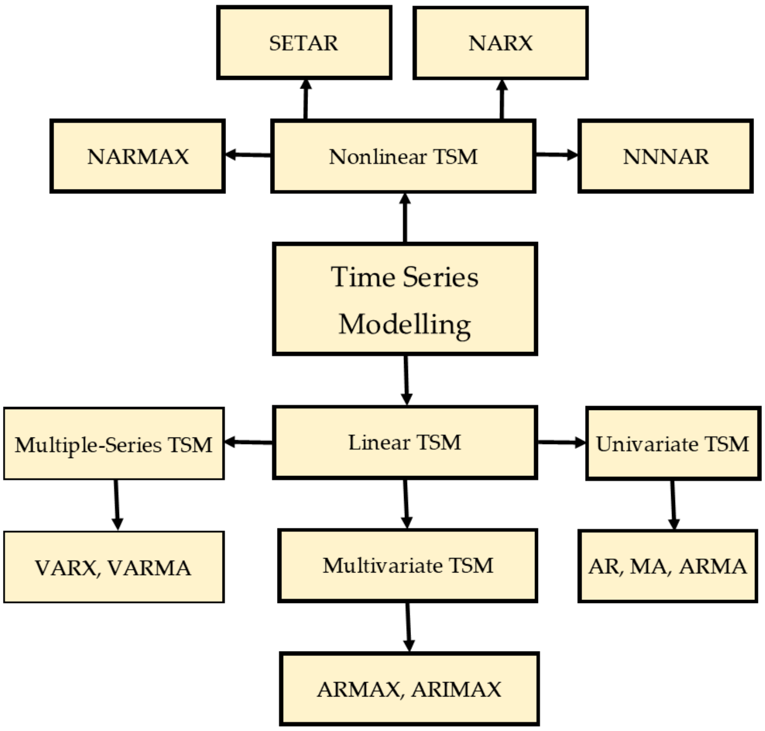


Figure 8.1. Classification of time series modelling techniques.

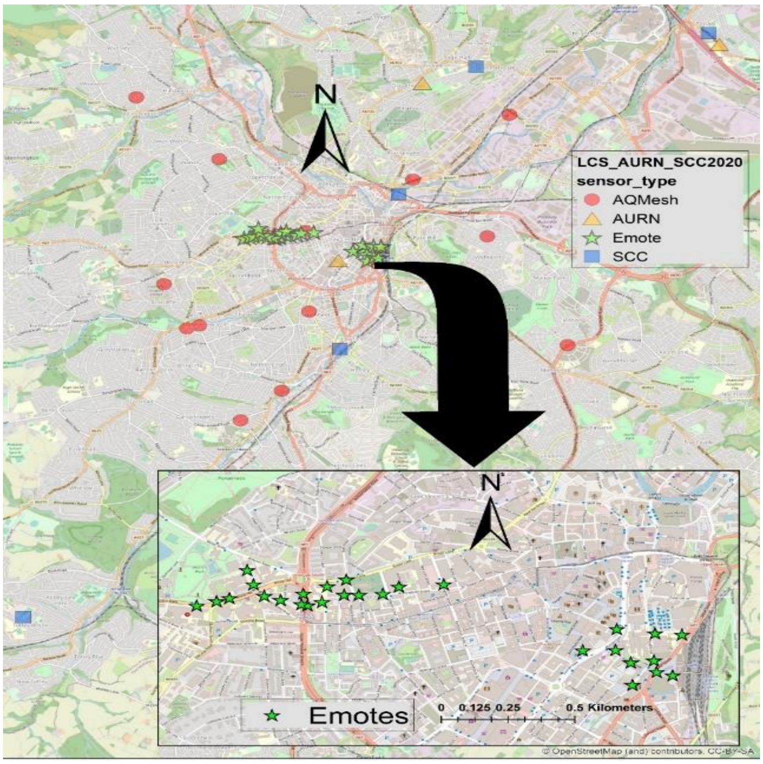
8.2. Methodology

8.2.1. A brief description of the monitoring network

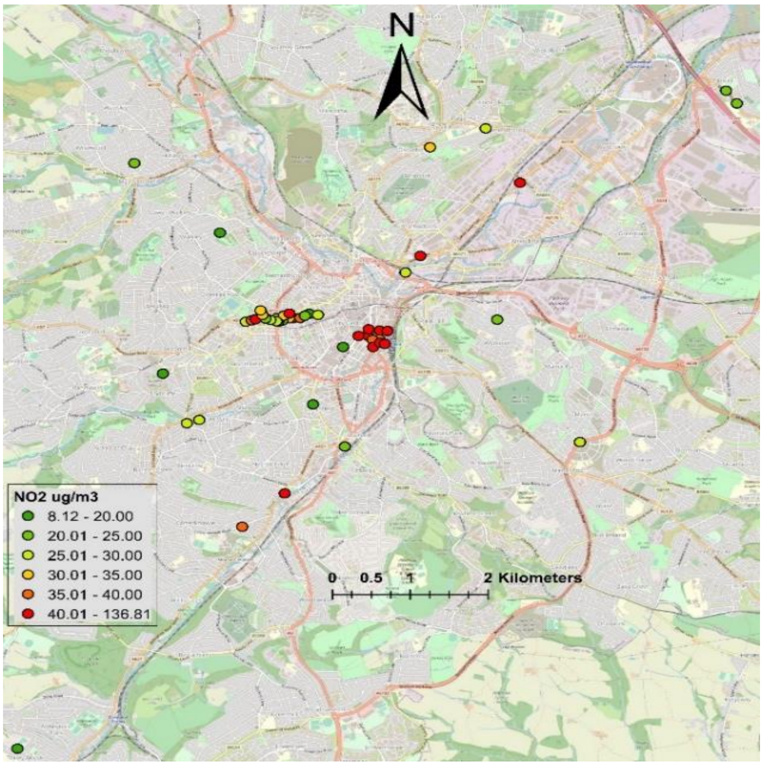
In this study, NO_2 concentrations ($\mu g/m^3$) from a network of low-cost sensors (LCSs) and reference sensors were analysed in Sheffield, United Kingdom. A network of LCSs was made by the Urban Flows Observatory, University of Sheffield, UK, consisting of 28 Envirowatch E-MOTEs and 13 AQMesh pods. These are all urban traffic sites. In addition, five reference air quality monitoring stations are operated by Sheffield City Council (SCC) and three are operated by the UK Department for Environment, Food and Rural Affairs (DEFRA), which are part of the Automatic Urban and Rural Network (AURN), the largest air quality monitoring network in the UK. The locations of the air quality monitoring stations (AQMSs) are shown in Figure 8.2 and their names, coordinates and annual mean NO_2 concentrations are given in Table 8.1. Data from these sensors were analysed from August 2019 to September 2020.

LCS are compact, portable and use less power when compared to reference instruments. LCS range in price from a couple of thousand to several thousand pounds (for a relatively sophisticated multi-pollutant and meteorological sensor with communication capabilities). Reference sensors are expensive, both to purchase and maintain, and bulky but are the most accurate units, recommended for use by EU and UK government bodies for AQ monitoring and comply with standards such as MCERTS in the UK. A single reference unit costs in the region of twenty thousand pounds to monitor a single gas or gaseous species or particle pollutants. LCS and reference sensors employ different techniques of air pollutant measurement, which include optical particle counters, light scattering, metal oxide semiconductor sensors, electrochemical sensors, nondispersive infrared sensors, ultraviolet fluorescence, chemiluminescence, infrared photometry and photo-ionisation detection sensors. For more detail see Borrego et al. (2016) and Mead et al. (2013). The LCS used in this project are either Envirowatch E-MOTEs or AQMesh pods. Envirowatch E-MOTEs are deployed in a local mesh deployed in a cluster, providing data via ZigBee, within a certain area for high resolution monitoring, no more than 100 m from each other, with a gateway providing uplink capability. AQMesh sensors are independent and can be deployed at both high and low spatial resolution. In this case each sensor independently sends data to a cloud server using GPRS.

(a)


N
LCS_AURN_SCC2020
sensor_type
AQMesh
AURN
Emote
SCC
N
Emotes
0 0.125 0.25 0.5 Kilometers
© OpenStreetMap (and) contributors, CC-BY-SA

(b)



Figur 8.2. (a) The locations of air quality monitoring stations (AQMSs) in Sheffield; (b) annual mean NO_2 levels (µg/m^3) measured by low-cost sensors (LCSs) and Automatic Urban

and Rural Network (AURN) and Sheffield City Council (SCC) AQMSs in Sheffield from August 2019 to September 2020.

Table 8.1. Showing the names, ID and annual mean NO_2 concentration (µg/m^3) measured by Low-Cost Sensors (AQMesh and Envirowatch E-MOTEs), AURN and Sheffield City Council (SCC) Sites from August 2019 to September 2020.

Site name	Sensor type	Sensor ID	NO_2 (µg/m^3)
Brightside Lane	AQMesh	2003150	41.3
Saville Street	AQMesh	2005150	46.9
Cundy Street	AQMesh	2007150	19.8
off Endcliffe Crescent	AQMesh	2008150	12.7
Sharrow Vale Rd	AQMesh	2009150	25.1
Abbeydale Rd	AQMesh	2450206	43.7
London Rd	AQMesh	2001150	14.6
Prince of Wales RD	AQMesh	2006150	28.2
Maltravers Rd	AQMesh	2004150	20.8
Hunter's Bar School	AQMesh	2450204	29.0
Malin Bridge PS	AQMesh	1999150	22.6
Broad Lane	AQMesh	1998150	19.2
Carter Knowle Bridge	AQMesh	2450205	37.6
Tinsley	AURN	SHE	22.9
Devonshire Green	AURN	SHDG	19.1
Barnsley Road	AURN	SHBR	32.1
Regent Court, E-camp	E_MOTE	711	41.7
Leavygreave Road, E-camp	E_MOTE	712	34.2
Gell Street, E-camp	E_MOTE	701	21.1
Upper Hanover/Henderson’s building	E_MOTE	702	28.7
Behind Jessop West	E_MOTE	703	23.7
Diamond/Bio-incubator	E_MOTE	704	34.2
Broad Lane/St George’s Terrace	E_MOTE	713	44.9
Portobello Street, Mappin Street	E_MOTE	714	37.6
28 Portobello Street, EC	E_MOTE	705	20.8
Howard Street, CC	E_MOTE	731	41.2
Arundel Gate/Genting Club	E_MOTE	732	115.6
Arundel Gate/Surrey Street	E_MOTE	733	43.5
Harmer Lane/Pond Street	E_MOTE	901	44.2
Harmer Lane/Sheaf Street	E_MOTE	902	49.9
Pond Street/Sheaf Building	E_MOTE	736	38.3
Howard Street/Science Park	E_MOTE	734	39.0
Paternoster Rows	E_MOTE	735	42.4
Sheaf Street/Sheaf Square	E_MOTE	903	107.2
Railway Station Taxi rank	E_MOTE	904	136.8
Upper Hanover St/Info. Commons	E_MOTE	707	26.6
Leavygreave Road/Favell Road	E_MOTE	708	25.7
Hounsfield Rd/Hicks Building	E_MOTE	709	24.9
Sheffield Children’s Hospital	E_MOTE	710	28.2
Robert Hadfield Building	E_MOTE	706	25.2
Brook Hill/Firth Court1	E_MOTE	737	39.3
Brook Hill/Firth Court2	E_MOTE	738	42.9
Arts Tower Concourse	E_MOTE	739	34.3
Arts Tower Concourse/Library	E_MOTE	740	33.2
Firvale	SSC	GH1	25.0
Tinsley	SSC	GH2	24.1
Lowfield	SSC	GH3	24.6
Wicker	SSC	GH4	25.9
King Ecgbert	SSC	GH5	8.1

LCSs are compact, portable and use less power when compared to reference instruments. LCSs range in price from a couple of thousand to several thousand pounds (for a relatively sophisticated multi-pollutant and meteorological sensor with communication capabilities). Reference sensors are expensive, both to purchase and maintain, and bulky, but are the most accurate units, recommended for use by EU and UK government bodies for air quality (AQ) monitoring and complying with standards such as MCERTS in the UK. A single reference unit costs in the region of 20,000 pounds to monitor a single gaseous or particle pollutant. LCSs and reference sensors employ different techniques for air pollutant measurement, which include optical particle counters, light scattering, metal oxide semiconductor sensors, electrochemical sensors, nondispersive infrared sensors, ultraviolet fluorescence, chemiluminescence, infrared photometry and photo-ionisation detection sensors. For more detail, see [27,28]. The LCSs used in this project were either Envirowatch E-MOTEs [29] or AQMesh pods [30]. The Envirowatch E-MOTEs are deployed in a local mesh in a cluster, providing data via ZigBee, within a certain area for high-resolution monitoring, no more than 100 m from each other, with a gateway providing an uplink capability. AQMesh sensors are independent and can be deployed at both high and low spatial resolutions. Both Envirowatch E-MOTEs and AQMesh pods are electrochemical sensors, whereas the sensors used by DEFRA and SCC are reference sensors which use chemiluminescent analysers for NO_2 and an ultraviolet (UV) absorption analyser for O3 measurements. The reference sensors are more reliable and accurate than the electrochemical sensors. The lowest detection limit for reference sensors and electrochemical sensor is < 2 μg/m^3. Electrochemical sensors are smaller and cheaper to purchase and maintain. Envirowatch E-MOTEs in a cluster communicate with a gateway by means of the Zigbee protocol within a specific area for high-resolution monitoring. The use of this protocol allows the individual units to communicate with each other and pass data from sensors that are not in range or without line-of-sight of the gateway. Using GPRS, the gateway device communicates the collected data over an internet connection to a cloud server operated by Envirowatch. Each AQMesh pod independently sends data to a cloud server using GPRS. For the reference sensors, the data logger is connected to the central management and coordination unit (CMCU) central computer, which collects the data using a GPRS mobile phone connection or wireless broadband.

8.2.2. Comparing NO_2 measured at different sites

In this study, we compared NO_2 concentrations measured at different AQMSs using LCSs and reference sensors from August 2019 to September 2020. For inter-site comparison, density plots and time variation plots were used, which were developed in R programming language [31] using the “openair” package [32]. Density plots, also known as kernel density estimation (KDE) plots, showed the distribution of NO_2 concentrations and are smoothed versions of histograms. The peaks of density plots show where values of the variables are concentrated. Furthermore, time variation plots were developed to show the temporal variability in NO_2 concentrations over different temporal scales; especially, they depict the diurnal, weekly and annual cycles of NO_2 concentrations.

8.2.3. Long-term temporal trend analysis

Quantification of temporal trends in air pollutant concentrations serves to assess the effects of emission control strategies over a given period of time. In this study, the temporal trend of

NO_2 concentrations was determined over the last 20 years (2000–2019) at the Tinsley AQMS, which is part of the UK AURN. This is an urban background site located at the Sheffield Tinsley Community Centre, approximately 200 metres east of the M1 motorway. The temporal trend was also determined at the Devonshire Green AQMS from 2014 to 2019, an urban background site, installed in Devonshire Green Park in Sheffield within a self-contained air-conditioned unit, surrounded mainly by open land and vegetation. To calculate long-term trend in NO_2 concentrations, here we employed the TheilSen function of the ”openair”' package [32]. The TheilSen function is a non-parametric approach and uses bootstrap simulations. This technique estimates all the regression parameters through bootstrap resampling. The technique is not affected by outliers as it is based on the median (not mean) and can be applied to a non-normal distribution.

8.2.4. Time series model

To carry out time series modelling in this study, NO_2, NO and O_3 data were used from the Devonshire Green AQMS in Sheffield. The site is part of the AURN network and has been monitoring various air pollutants, including NO_2, NO, NOx and O_3, since 2013. Figure 8.3 shows the monthly average of pollutants, which exhibits a general trend of higher NO, NO_2 and NOx concentrations in winter and lower concentrations in summer. In contrast, O_3 concentrations were higher in spring and summer and lower in winter. In winter the atmosphere is relatively static and the atmospheric boundary layer is shallower, which is a hindrance in pollutant dispersions; whereas, in summer, the atmosphere is more turbulent and both horizontal and vertical dispersion processes are active in dispersing locally emitted pollutants. This is probably the main reason that the concentrations of NO, NO_2 and NOx were lower in summer and higher in winter. On the other hand, ground level O_3 is a secondary pollutant and is produced by the photochemical reactions of NOx and volatile organic compounds (VOCs) in the presence of solar radiation. In summer, high temperature and solar radiation lead to higher levels of O_3 production than in winter, when the weather in the UK is cold, not conducive to O_3 formation. O_3 concentration was highest in May, when in addition to the weather conditions, O_3 precursors were abundant, in contrast to June and July when the precursors were relatively lower [33].

Time series data mainly have three building blocks: seasonality, trend and residuals [34]. Deconstruction (decomposition) of a time series is helpful in understanding the behaviour of a time series. The seasonality (seasonal component) of the time series refers to the fluctuations in the levels of pollutants related to seasonal factors. Seasonality is always of a fixed and known period, normally twelve months. The trend component is the overall long-term pattern of the time series and indicates if a pollutant concentration is increasing or decreasing over time. The residual or error component of the time series is the remainder part that cannot be attributed to seasonality and trend components. The three components of the time series are shown below in Equation (8.1):

$$Y = Tc + Sc + Rc \quad (8.1)$$

In Equation (8.1), a time series (Y) is divided into three components, which are the trend component (Tc), seasonal component (Sc) and residual or remainder component (Rc). For dividing the time series into three components, seasonal-trend decomposition based on loess (STL) was used, which was initially proposed by Cleveland et al. [35]. The process of extracting these components from the time series is referred to as decomposition. The three components of the NO_2 time series are shown in Figure 8.4.

In this study, air pollution data were analysed using four time series models: ARIMA, ARIMAX, the self-exciting threshold autoregressive (SETAR) and neural network nonlinear autoregressive (NNNAR). For more details on these models, see [36–38]. Time series model development and application consist of three main steps [39]:

- model specification;
- model fitting (parameters estimation); and
- model diagnostic (testing the goodness of fit of the fitted model).

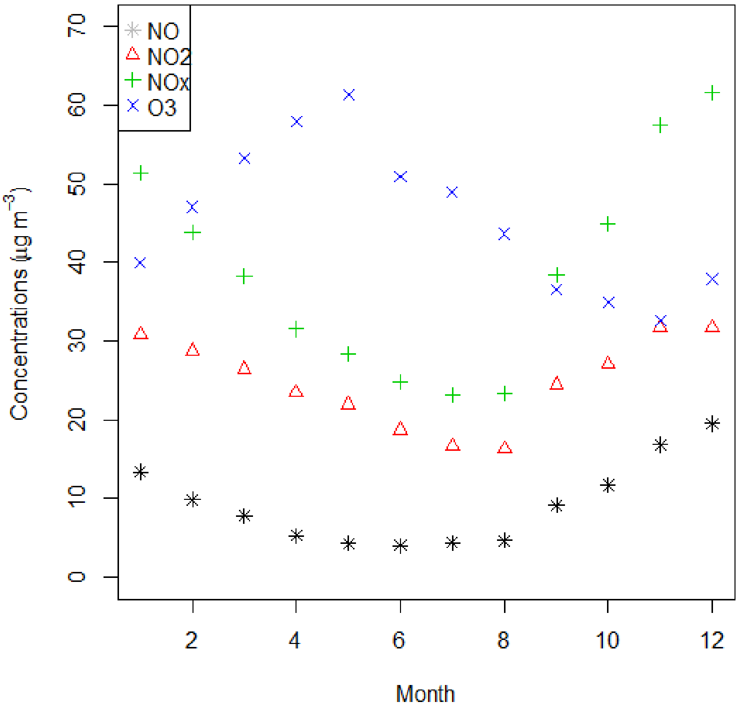


Figure 8.3 Monthly average concentrations of NO, NO_2, NOx and O_3 in different months of the year from Devonshire Green AQMS in Sheffield, 2015 - 2019.

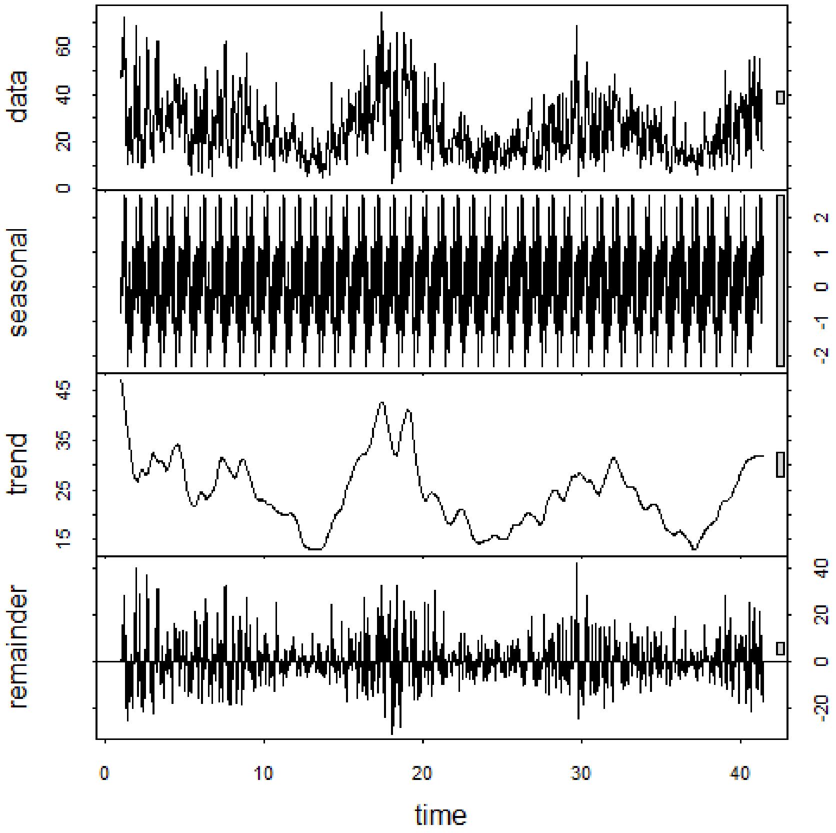


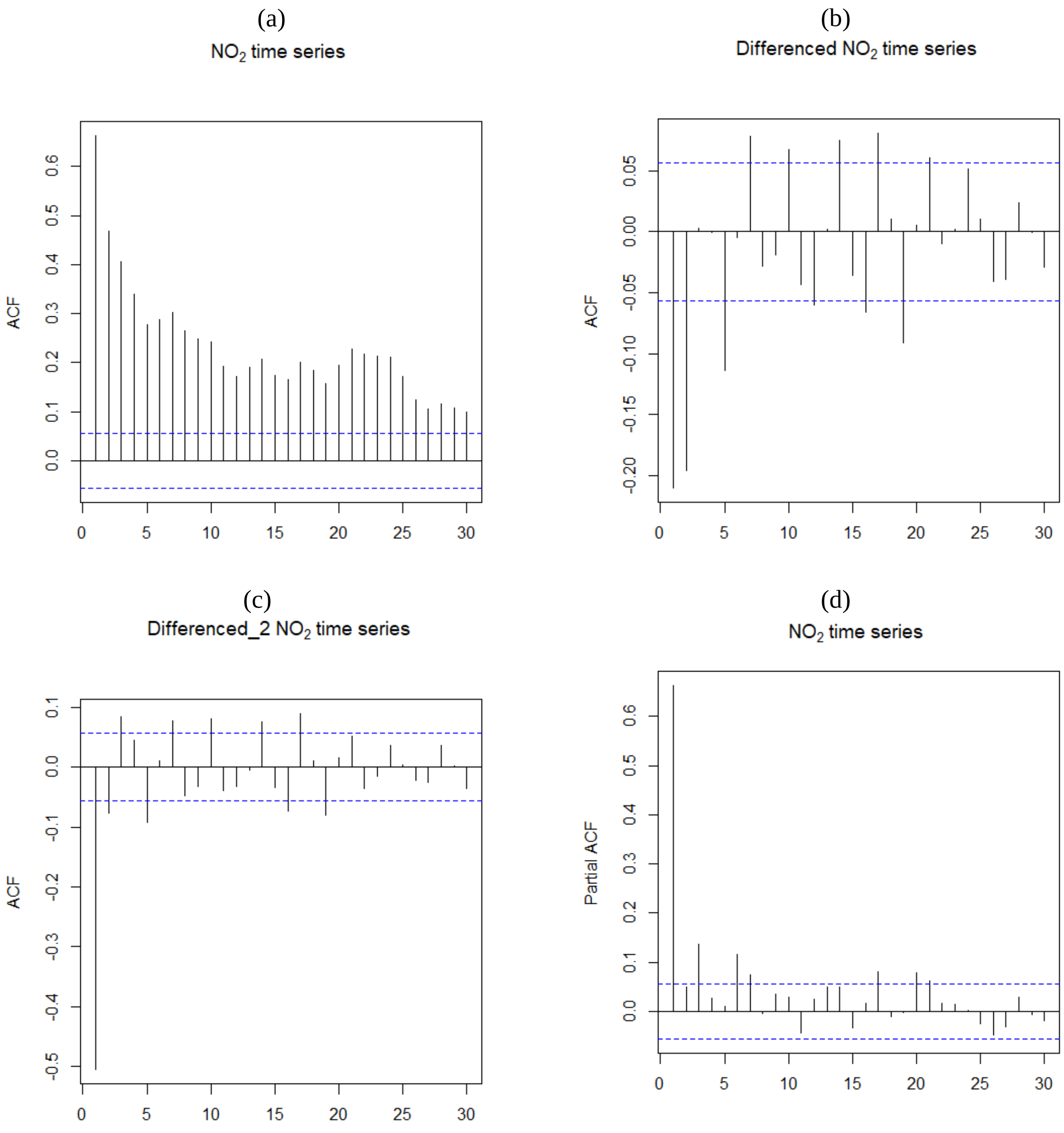
Figure 8.4. Three additive components obtained from STL (Seasonal-Trend Decomposition based on Loess) decomposition of NO_2 concentration (µg/m^3) collected at Devonshire Green AQMS, Sheffield.

8.2.4.1. Model specification

Model specification means choosing appropriate values for p, d and q for a given time series, where p is the lag orders (i.e., the number of lagged values that NO_2 is regressed on), referred to as the autoregressive component, d is the degree of differencing (integration) and q is the order of moving average. Differencing a series involves simply subtracting its current and previous values' d times. Often, differencing is used to stabilize a series when the stationarity assumption is not met. For model specification (to specify the values of p, d and q), firstly, the auto-correlation function (ACF) and partial auto-correlation function (PACF) plots of the NO_2 time series were used. The ACF plot (Figure 8.5a) displays correlation between a series and its lags (previous values) and helps determine the order of differencing (d). Furthermore, the ACF plot can help in determining the order of the MA component. An MA component represents the error of the model as a combination of previous error terms [39]. The order q determines the number of terms to be included in the model. The ACF plot of the NO_2 series (Figure 8.5a) shows the properties of a nonstationary series, as the ACF fails to die out rapidly with increasing lags. All ACF values are significant (significantly greater than zero) and the only pattern is perhaps a linear decrease with increasing time lag. This shows that NO_2 series is nonstationary. Before a time-series model (e.g., the ARIMA model) is applied, the time-series must be stationarised, which means the variance, mean and auto-covariance of the time series need to be time invariant. The ACF plots of the NO_2 series differenced once and twice are presented in Figure 8.5b,c, respectively. After differencing once, the pattern emerged much more clearly, suggesting that the ARIMA model with difference 1 (d = 1) was probably a suitable model.

Table 8.2. ARIMA model specification and corresponding AIC values of NO_2 time series, where ‘p’ represents order of autoregressive, ‘d’ represents difference and ‘q’ represents MA.

AIC	p	d	q
8910.88	1	1	1
9104.28	1	2	1
8910.53	2	1	1
8912.42	2	1	2
8944.8	1	0	0
8941.53	1	0	1
8904.29	3	1	1
8905.83	4	1	1


(a)
(b)
(c)
(d)

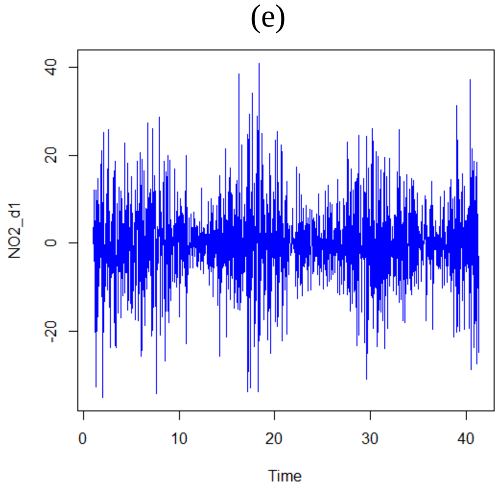


Figure 8.5. Auto-correlation function (ACF) plot of NO_2 time series (not differenced) (a), differenced once (b), differenced twice (c), partial ACF plot (d) and differenced once NO_2 plot of oscillating pattern around zero with no visible strong trend (e) showing that the series is stationary.

The PACF plot (Figure 8.5d) displays the correlation between a variable and its lags that is not explained by previous lags. PACF plots are useful when determining the order p of the AR component, which specifies the number of lags used in the model; for example, AR (2) or ARIMA (2,0,0) showed that the order of AR was 2. The PACF plot showed that an AR (1) model should be considered (Figure 8.5d). The plot of differenced NO_2 is shown in Figure 8.5e, wherein the oscillating pattern around the zero shows that the series was stationary after the series was differenced once. This suggested that differencing of order 1 terms is sufficient and should be included in the model. Sometimes, if the series is not stationarised after differencing once, we need to difference it again. In addition to the ACF and PACF plots, we also considered model selection criteria based on the value of the Akaike's information criterion (AIC). The time series model with a lower value for the AIC showed a better fit. Based on the AIC values (Table 8.2), ARIMA (1,1,1) was selected. Differencing, autoregressive and moving average components make up an ARIMA model. ARIMA (1,1,1) showed that the model incorporated differencing of degree 1 and used an AR term of first lag and an MA of order 1. Although some of the other models (Table 8.2) had slightly lower AIC values, the difference was not significant and the model was more complicated, which is against the principle of parsimony, meaning that the selected model should use the lowest number of parameters providing adequate representation of the time series. When the ARIMA (1,1,1) model was applied to the square root and log of the NO_2 time series in contrast to observed NO_2 concentrations, the AIC values dropped to 3230.6 and 1117.5, respectively, which are much lower than the values presented in Table 8.2. Therefore, log-NO_2 was adopted in this study instead.

8.2.4.2. Model fitting

Model fitting is basically parameter estimation. This study used the maximum likelihood estimation approach, which produces more accurate estimation of parameters in comparison to least square estimation [39]. The main advantage of the maximum likelihood estimation is that it uses all of the information in the data rather than just using the mean and variance of the data, as is the case with least square estimation [39]. Using daily NO_2 concentrations (2015 to 2019), the dataset was divided into training (70%) and testing (30%) datasets, which were selected randomly. Time series linear models were fitted (trained) with the TSA-package [40] and forecast-package [41] in R programming language [31]. The neural network nonlinear autoregressive and self-exciting threshold autoregressive (SETAR) nonlinear models were implemented in R-package ”tsDyn” [38]. The model performance was assessed using both training and testing datasets. The specified models were run to estimate the model parameters.

8.2.4.3. Model diagnostic and forecasting

Residuals of a model are defined as the difference between the actual (observed) and predicted (modelled) values as shown in equation 8.2:

$$\text{residual} = \text{actual values} - \text{predicted values} \qquad (8.2)$$

Residual analysis of the model was performed by graphical presentation using a plot of standardised residuals, a quantile–quantile (Q–Q) plot and a histogram of the residuals. For an adequate model, a plot of standardised residuals shows a rectangular scatter around a zero horizontal level with no trends, whereas a Q–Q plot and histogram of the residuals show the normality of the error terms.

The model performance was assessed by comparing predicted with observed NO_2 concentrations both graphically, using time plots and scatter plots, and using several statistical metrics, including correlation coefficients (r), root mean square error (RMSE), factor of two (FAC2), mean biased error (MBE) and mean absolute error (MAE) [42,43]. The Pearson correlation coefficient measures the strength of the linear relationship between the measured and modelled concentrations. FAC2 is the fraction of predicted concentrations within a factor of two of the measured concentrations. RMSE, MAE and MBE provide an estimation of the error of the model prediction. However, RMSE and MAE do not define the direction of the error, as they provide absolute error, whereas MBE, in addition to the size, defines the direction of the error; i.e., negative MBE indicates under-prediction whereas positive MBE indicates over-prediction of a model.

The fitted model was used to predicted NO_2 concentrations, which were crossvalidated with the testing dataset. If the predicted value (forecast) is denoted by $\hat{Y}_t$ (l), where l is the lead time for forecast and time t is the forecast origin, then forecasting one time-unit into the future (one step ahead) (Yt (1)) can be achieved as in Equation (8.3) [39]:

$$\hat{Y}_t(1) = \mu + \phi(Y_t - \mu) \qquad (8.3)$$

Where μ is the intercept and φ is the estimated parameter of the AR component. Equation (8.3) shows that, to forecast the next step, a proportion φ of the current deviation from the process mean is added to the process mean. To consider a general lead time l, we replace time t by t + l in Equation (8.3) and take the conditional expectation of both sides, which produces Equation (8.4) [39]:

$$\hat{Y}_t(l) = \mu + \phi\,[\hat{Y}_t(l - 1) - \mu] \text{ for } l > 1 \qquad (8.4)$$

Equation (8.4) is recursive in the lead time l and shows how the forecast for any lead time l can be built up from the forecast for a shorter lead time l by starting with the initial forecast as given by Equation (8.3). The forecast $\hat{Y}_t$ (2) is then obtained from $\hat{Y}_t$ (2) = μ + φ [$\hat{Y}_t$ (1) − μ], then $\hat{Y}_t$ (3) from $\hat{Y}_t$ (2) and so on. Equation (8.4) is also known as the ”difference equation form” of the forecast [39]. An explicit expression for the forecast in terms of the observed history of the time series for Equation (8.4) can also be given in the form of Equation (8.5) [39].

$$\hat{Y}_t = \mu + \varphi\ l\ (Yt - \mu)\ (8.5)$$

Equation (8.5) shows that the current deviation from the mean is discounted by a factor φl , whose magnitude decreases with increasing lead time l. The discounted deviation is then added to the process mean to produce the lead l forecast. It should be noted that, in forecast, if an exogenous variable is used, then the number of steps ahead are ignored and the number of forecast periods is set to the number of rows of exogenous variables in the new data, which here is equal to the number of rows of the testing dataset.

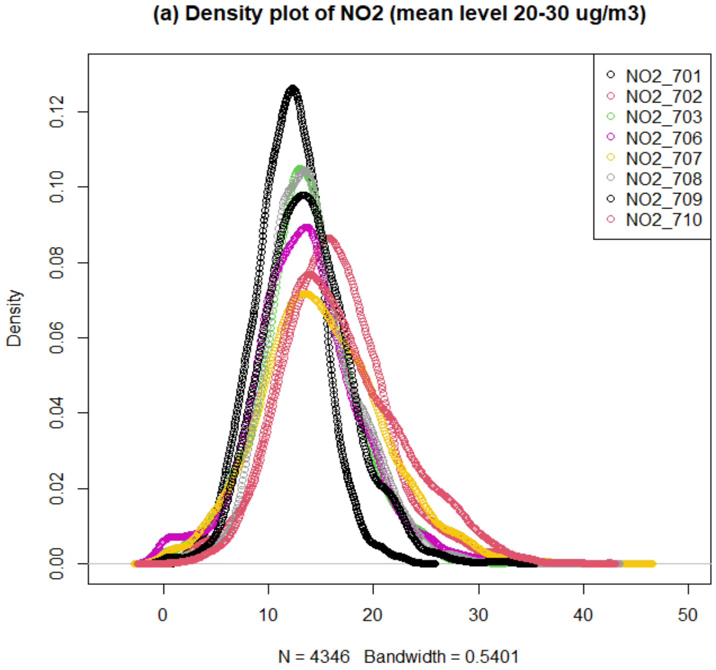
8.3. Results and discussion

8.3.1. Comparing NO_2 measured at different sites

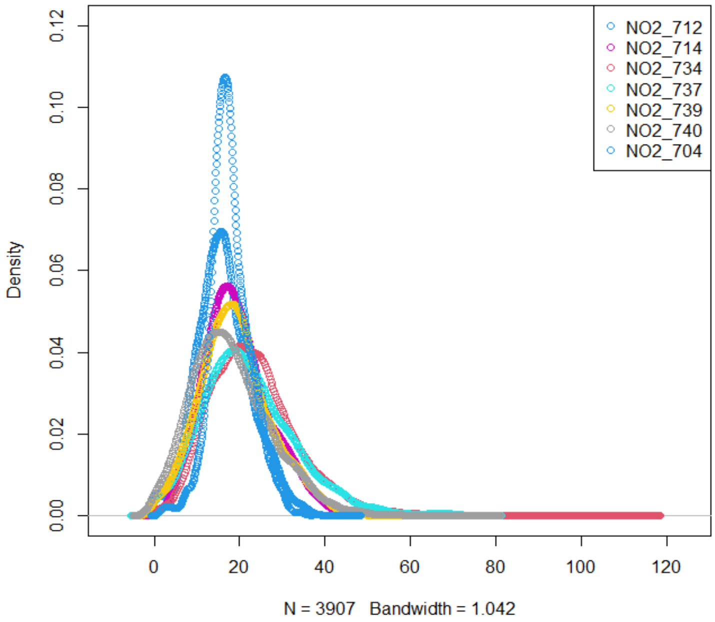
8.3.1.1. Density plots

NO_2 concentrations (µg/m^3) measured by different grades of sensors at different monitoring stations in Sheffield were compared employing densities plots. Density plots offer an easy way to compare the distribution of NO_2 concentrations measured at different monitoring stations. Figure 8.6 shows NO_2 concentrations measured by Envirowatch E-MOTEs. The performance of Envirowatch E-MOTEs was assessed by Munir et al. [44] by comparing their concentrations with a nearby reference sensor in Sheffield. The sensors were divided into four sub-groups based on the mean NO_2 concentrations. In first group of sensors (Figure 8.6a), the range of NO_2 concentrations was 0 to 50 µg/m^3 and the highest density (mode) occurred at 10–15 µg/m^3. This represents the group of sensors with the lowest concentrations, which are deployed outside the city centre at the University of Sheffield campus. The second group (Figure 8.6b) had mean values of 31–40 µg/m^3 and the highest density was recorded at 20–25 µg/m^3. Most of the sensors in this group had a range of 0 to 60 µg/m^3, however some sensors recorded NO_2 concentrations as high as 120 µg/m^3. The third group (Figure 8.6c) had mean concentrations of 41 to 50 µg/m^3 and the concentrations ranged from 0 to 140 µg/m^3. These sensors recorded mean values over 40 µg/m^3 , exceeding annual air quality standard. The fourth group (Figure 8.6d) recorded the highest NO_2 concentrations, where mean concentration was greater than 100 (NO_2 > 100 µg/m^3). The highest density was shown at 70 to 80 µg/m^3. In total, 11 sensors violated AQ standards (NO_2 > 40 µg/m^3). Three sites with mean NO_2 levels higher than 100 (µg/m^3) were: (i) the Taxi Rank at the Sheffield Railway Station (NO_2_904, 136.81 µg/m^3), (ii) Arundel Gate opposite to the Genting Club (NO_2_732, 115.56 µg/m^3) and (iii) the pedestrian crossing at Sheaf Street/Sheaf Square (NO_2_903, 107.17 µg/m^3). The reason for recording higher NO_2 concentrations is that these sensors are installed next to busy locations in terms of road traffic flows and engine idling while stationary. The taxi stand (E-MOTE 904) is a good example of how engine idling can result in worse air pollution in urban areas. E-MOTE 903 is installed next to a pedestrian crossing on a busy road. People coming out of the train station use the pedestrian crossing going towards the high street, Sheffield Hallam University, and other parts of the city. The traffic light turns red and green regularly, which cause congestions. When the lights are red, road traffics stop but vehicle engines keep running. When traffic lights turn green, all vehicles try to accelerate quickly. Therefore, idling of engine and sudden acceleration emit extra pollution. E-MOTE 732 is installed in a typical street canyon, where the road has tall buildings on both sides, hindering the dispersion of the pollutants emitted by the road traffics, causing the pollution levels to go up. The other sites in the city centre where air quality standards were violated were Paternoster Rows, Harmer Lane near Sheaf Street, Harmer Lane near the bus station, Arundel Gate near Surrey Street, and Howard Street. Two E-MOTEs at the University campus that exceeded air quality limits were Regent Court and Broad Lane (near St. George's Terrace).

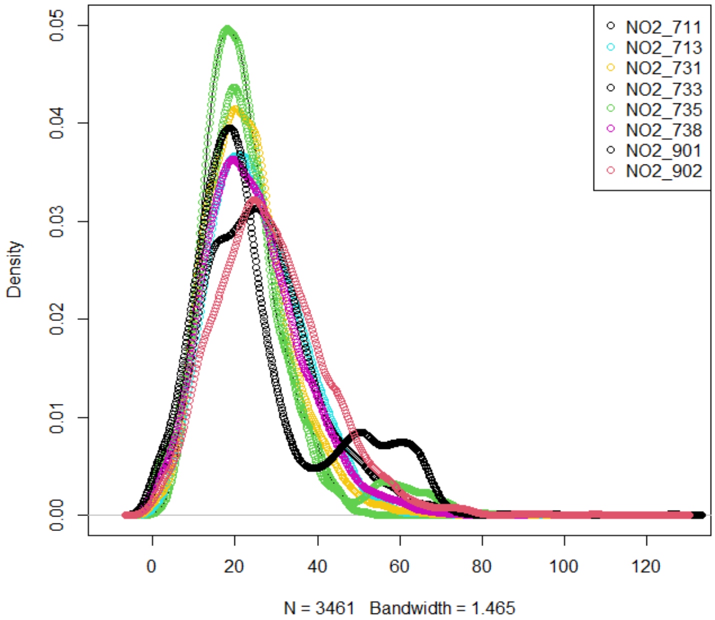
The distributions of NO_2 concentrations measured by different AQMesh pods are compared using density plots in Figure 8.7. Most of the AQMesh pods were deployed in the outskirts of the city (Figure 8.2). The distribution is mostly skewed right with heavy right tails. Only three AQMesh pods exceeded AQ standards, which were NO_2_206 (Abbeydale Rd, 43.67 μg/m^3), NO_2_2003 (Brightside Lane, 41.34 μg/m^3) and NO_2_2005 (Savile Street, 46.91 μg/m^3). The measurements seem realistic as these sensors are deployed on very busy roadsides.



(b) Density plot of NO2 (mean level 31-40 ug/m3)



(c) Density plot of NO2 (mean level 41-50 ug/m3)



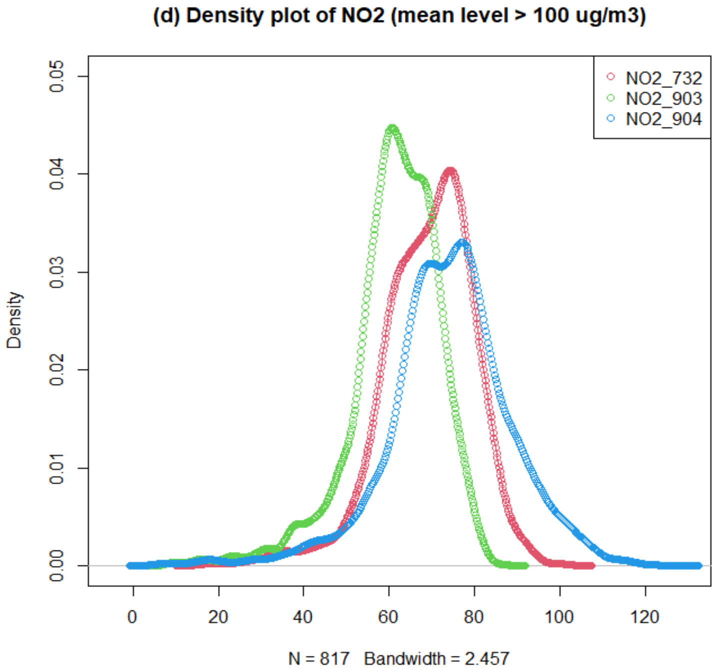
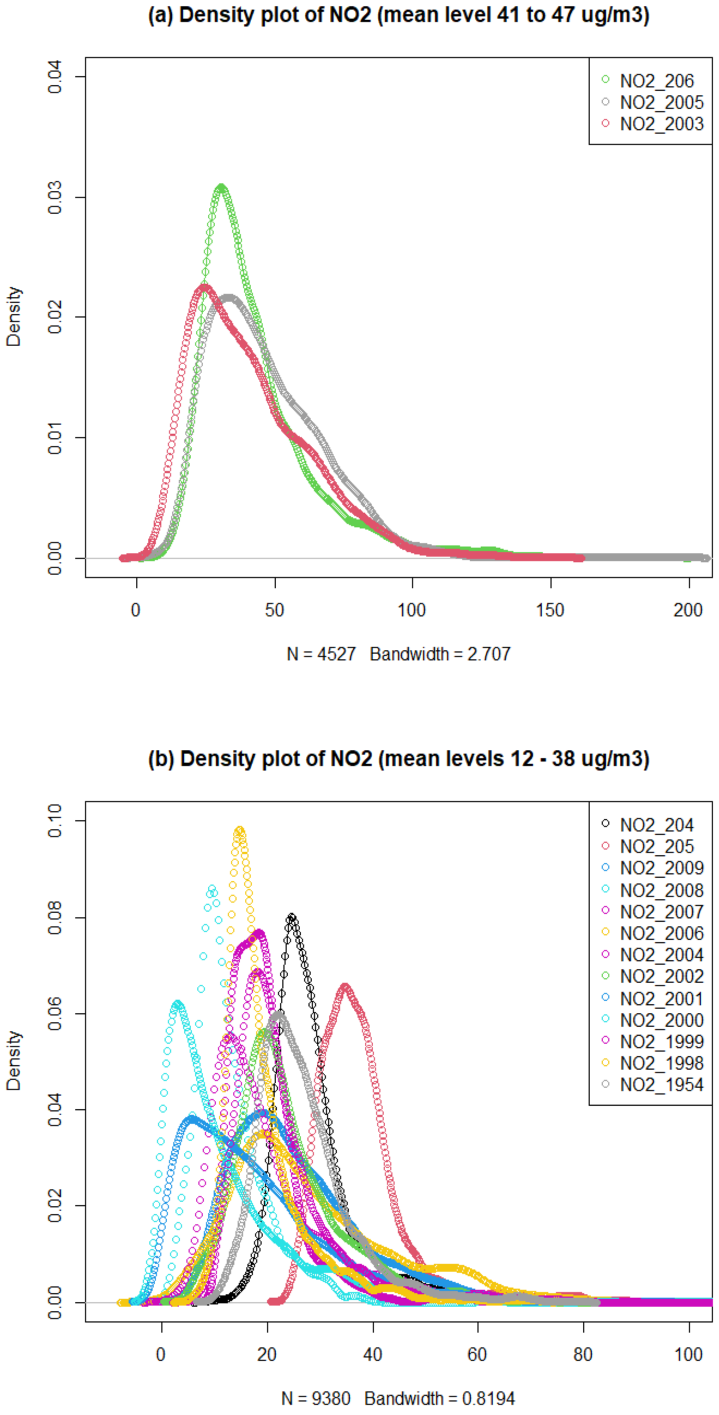


Figure 8.6. Different density plots of hourly NO_2 concentrations (µg/m^3) measured by Envirowatch E-MOTEs, August 2019–September 2020 in Sheffield.

Density plots of NO_2 concentrations measured by AURN sites and SSC sites are shown in Figure 8.8a and 8.8b, respectively. The distributions are skewed right. AURN sites had mean NO_2 concentrations of 23, 19 and 32 µg/m^3 at the Sheffield Tinsley, Devonshire Green and Barnsley road AQMSs, respectively. The SSC sites Firvale, Tinsley, Lowfield, Wicker and King Ecgbert had mean concentrations of 25, 24, 25, 26 and 8 µg/m^3, respectively. The concentrations measured at both AURN and SCC sites were well below the AQ standards. The lowest concentrations were recorded at the King Ecgbert site, which is located in a background location well outside the city centre. In this section, we considered the distribution and annual mean of NO_2 concentrations measured at different AQMSs of different grades. Density plots are a useful tool for understanding the distribution of NO_2 concentrations and comparing the distributions of different monitoring sites. However, density plots provide no information on the temporal variability of NO_2 concentrations, which are analysed in the coming section.



(a) Density plot of NO2 (mean level 41 to 47 ug/m3)

(b) Density plot of NO2 (mean levels 12 - 38 ug/m3)

Figure 8.7. Density plots of hourly NO_2 concentrations (µg/m^3) measured by AQMesh pods from August 2019 to September 2020 in Sheffield.

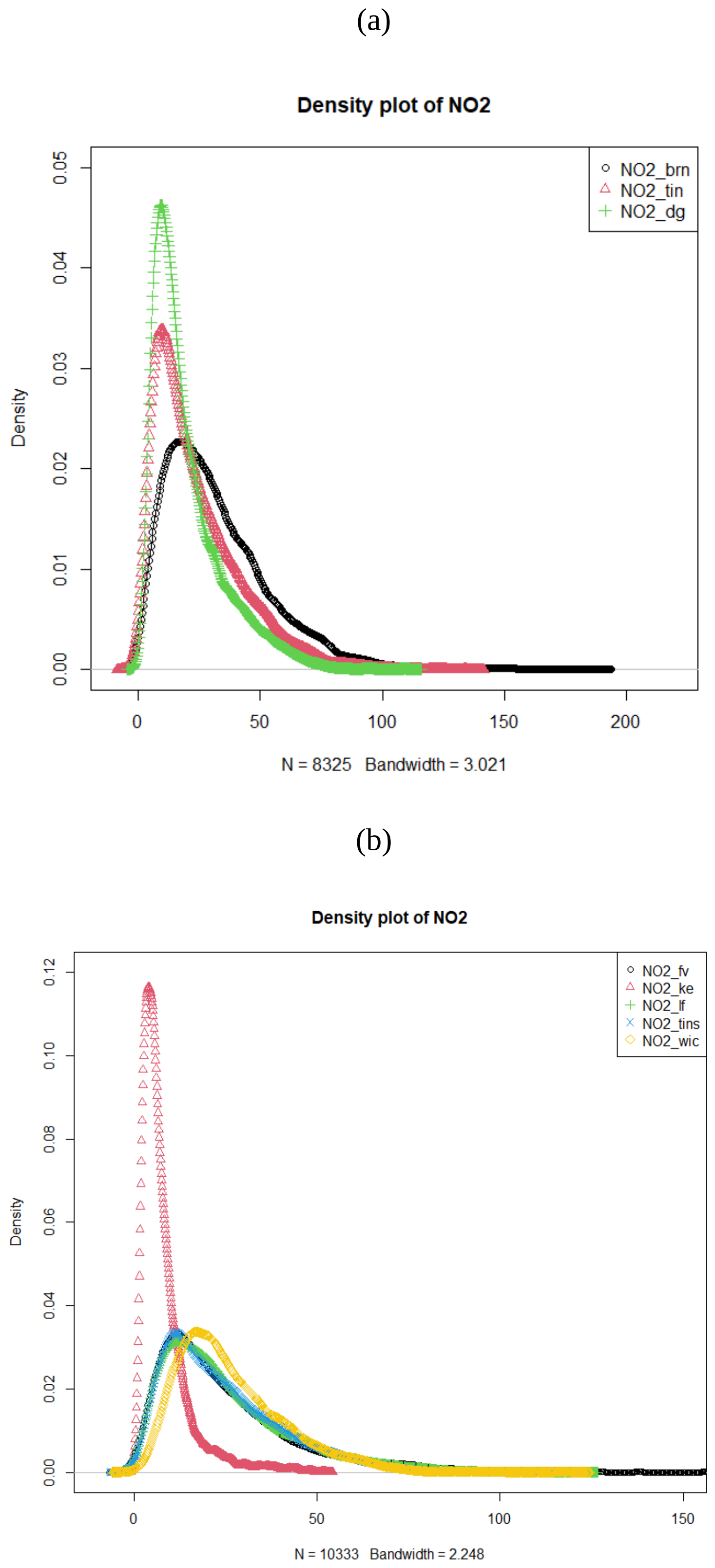


Figure 8.8. Density plots of NO_2 concentrations (μg/m^3) measured by: (a) AURN sites-Barnsley road (brn), Tinsley (tin) and Devonshire Green (dg); and (b) SCC sites-Firvale (fv), King Ecgbert (ke), Lowfield (lf), Tinsley (tins) and Wicker (wic), from August 2019–September 2020.

8.3.1.2. Time variation plots

It is important to know how NO_2 concentrations vary at different time scales. Time variation plots show variations in NO_2 concentrations on diurnal, weekly and annual cycles. Such information is useful in understanding the emission sources of air pollutants. In this section, temporal variabilities of NO_2 concentrations are analysed by employing time variation plots, using data from AURN, SSC, Envirowatch E-MOTEs and AQMesh pods. NO_2 concentrations measured by AURN and SCC sites are analysed first to set a benchmark of temporal variation for the LCS. Barnsley is a roadside (urban traffic) AQMS, whereas Tinsley and Devonshire Green are urban background sites. Therefore, NO_2 concentrations measured at Barnsley site (NO_2_brn) were higher than at the other two sites (Figure 8.9). Figure 8.9 shows that, on diurnal cycles, NO_2 concentrations were higher in the busy morning hours (07:00–09:00 h) and busy afternoon hours (17:00–19:00 h) and lower at night and midday. Weekly cycles showed that NO_2 concentrations were significantly lower on weekends (Sunday being the lowest) than on weekdays. Annual cycles of NO_2 showed higher concentrations in colder months (e.g., January, February, November and December) than the warmer months (e.g., June, July and August). All three AURN sites demonstrated the same temporal trends on diurnal, weekly and annual cycles. Diurnal and weekly cycles of NO_2 concentrations are controlled by road traffic flow, whereas annual cycles in addition to emissions are controlled by meteorological parameters. In winter, the temperature is low and the atmosphere is stagnant, which hinders the dispersion of air pollutants emitted locally. In contrast, in summer the atmosphere is more turbulent, encouraging both vertical and horizontal dispersion of pollutants.Figure 8.9 shows temporal variability of NO_2 concentrations using data from five SSC sites: NO_2_fv (Firvale), NO_2_ke (King Ecgbert), NO_2_lf (Lowfield), NO_2_tins (Tinsley) and NO_2_wic (Wicker). SCC AQMS demonstrated a temporal trend similar to AURN sites. In addition to some minor differences between different sites, NO_2 concentrations were significantly lower at the King Ecgbert site, which is a background site located outside the city centre in a quite location.

Figure 8.10 shows the temporal variability of NO_2 concentrations using data from five SSC sites: NO_2_fv (Firvale), NO_2_ke (King Ecgbert), NO_2_lf (Lowfield), NO_2_tins (Tinsley) and NO_2_wic (Wicker). SCC AQMSs demonstrated a temporal trend similar to AURN sites. In addition to some minor differences between different sites, NO_2 concentrations were significantly lower at the King Ecgbert site, which is a background site located outside the city in a quite location.

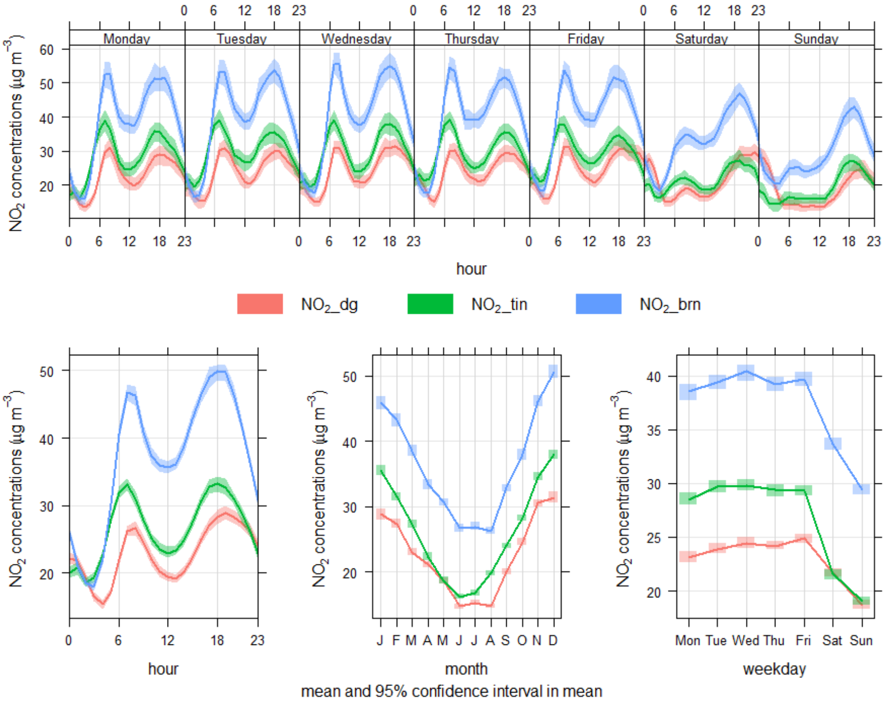


Figure 8.9. Time variation plots of NO_2 concentrations (µg/m^3) in Sheffield at the three AURN AQMS: Barnsley, Tinsley and Devonshire Green.

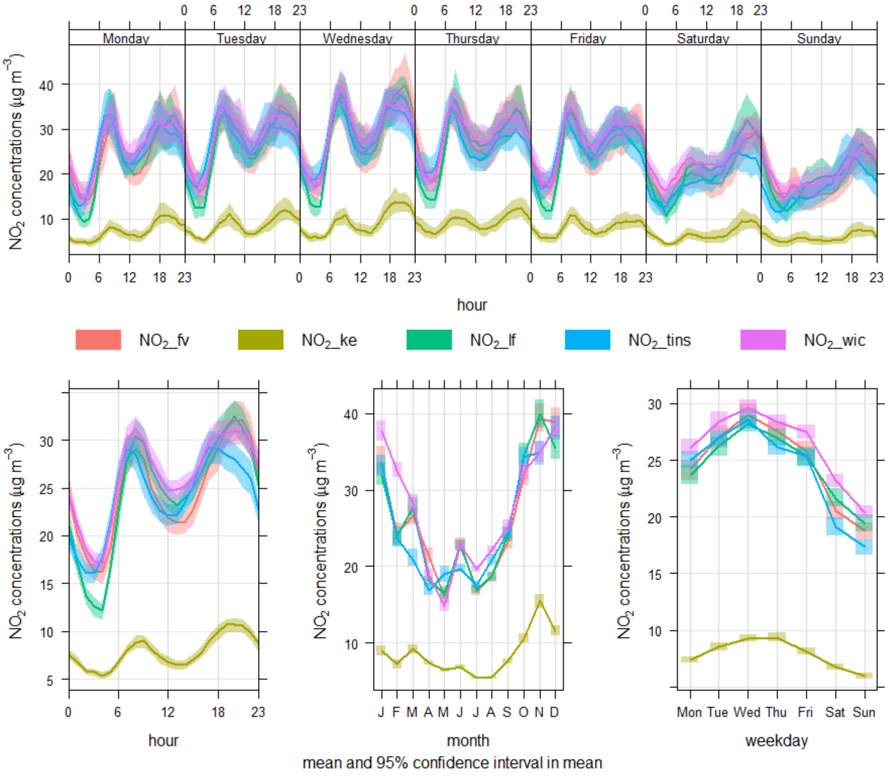


Figure 8.10. Time variation plots of NO_2 concentrations (µg/m^3) measured by SCC sites: Firvale, King Ecgbert, Lowfield, Tinsley and Wicker, from August 2019 – September 2020.

To demonstrate temporal variability in NO_2 concentrations (Figure 8.11) measured by Envirowatch E-MOTEs, four E-MOTEs were selected: NO_2_904 (Railway Station), NO_2_902 (Harmer Lane/Sheaf Street), NO_2_731 (Howard Street) and NO_2_738 (Brook Hill, near Firth Court). These four sites were selected because NO_2_904 presented an unusual time profile, whereas NO_2_902, NO_2_738 and NO_2_731 showed a typical time variations similar to many other sensors. The diurnal, weekly and annual cycles followed almost similar pattern as those shown by AURN and SSC sites. However, there were some minor differences between various E-MOTEs installed at different locations within the city. Also, differences in winter vs. summer and weekend vs. weekday concentrations were much prominent in AURN sites than in E-MOTEs. NO_2_904 demonstrated higher concentrations at all times, even at nights and weekend, as the taxi stand next to the train station remains busy at all times waiting for train passengers (probably with vehicles idle while waiting).

Figure 8.12 shows the diurnal, weekly and annual cycles of NO_2 concentrations measured by four AQMesh pods. The four AQMesh sites were: NO_2_1999 (Malin Bridge School), NO_2_2001 (London Rd), NO_2_2003 (Brightside Lane) and NO_2_204 (Hunter's Bar School). Here the diurnal and weekly cycles showed similar trends; however, the annual cycles were different from each other and from the ones shown by AURN and SSC AQMSs. The NO_2_2001 annual cycle was particularly different as it showed higher concentrations in summer months than in winter months. However, weekly cycles presented normal trends, as expected and shown by AURN sites.

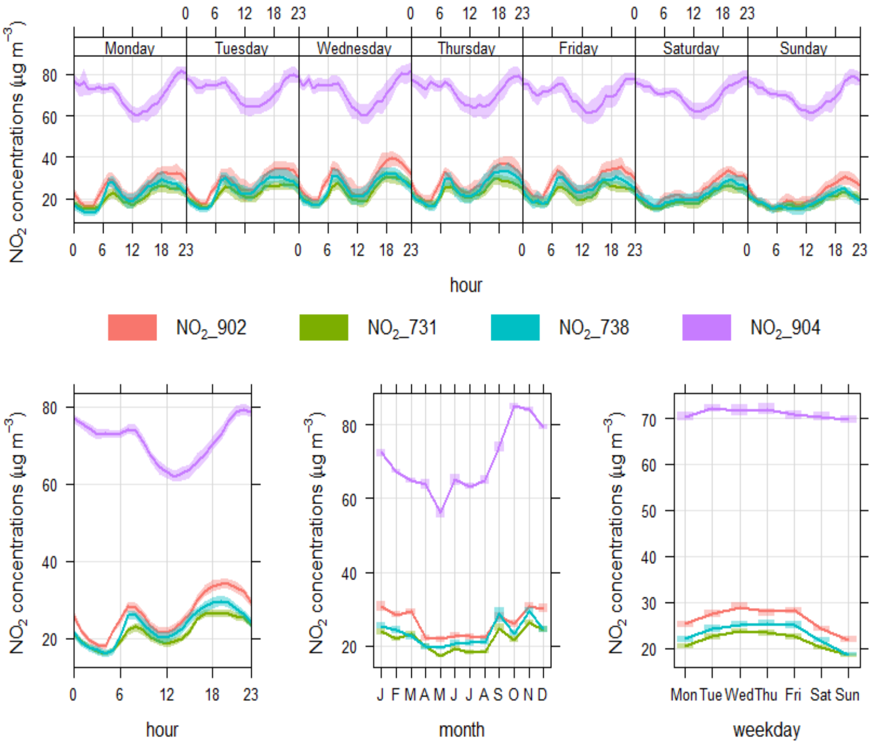


Figure 8.11. Time variation plots of NO_2 concentrations (µg/m^3) measured by Envirowatch E-MOTEs.

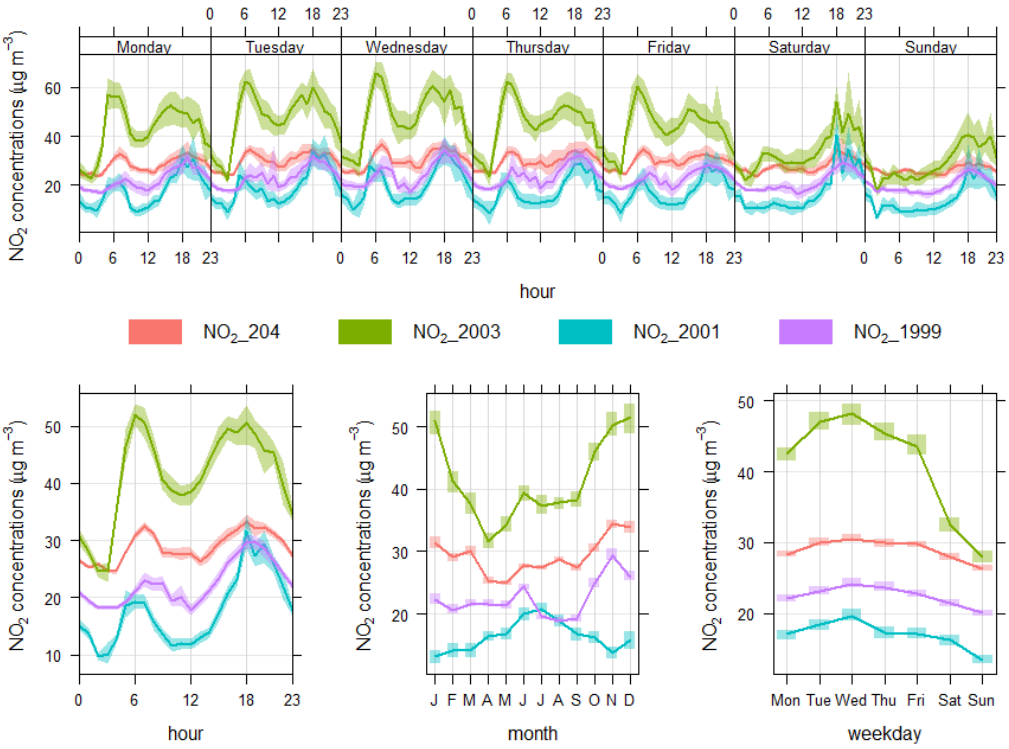


Figure 8.12. Time variation plots of NO_2 concentrations (µg/m^3) measured by AQMesh pods.

8.3.2. Long term trends in NO_2 concentrations

The temporal trend in NO_2 concentrations was determined using the TheilSen function for the last 20 years (2000–2019) at Tinsley and for the last six years (2014–2019) at the Devonshire Green AQMS. Both of these sites are part of the AURN. Trends were expressed in µgm^{-3}/year. Both non-adjusted (non-deseasonalised) and adjusted (deseasonalised) trends were calculated. Sheffield Tinsley demonstrated negative trends during the study period. Both adjusted and non-adjusted trends were negative and highly significant. The non-adjusted trend was −0.94 (−1.12, −0.78) µgm^{-3}/year, whereas the adjusted trend was −0.95 (−1.04, −0.86) µgm^{-3}/year (Figure 8.13). The temporal trend at the Devonshire Green site was also negative and highly significant. The non-adjusted trend was −1.21 (−1.91, −0.41) µgm^{-3}/year and the adjusted trend was −1.26 (−1.57, −0.83) µgm^{-3}/year (Figure 8.14).

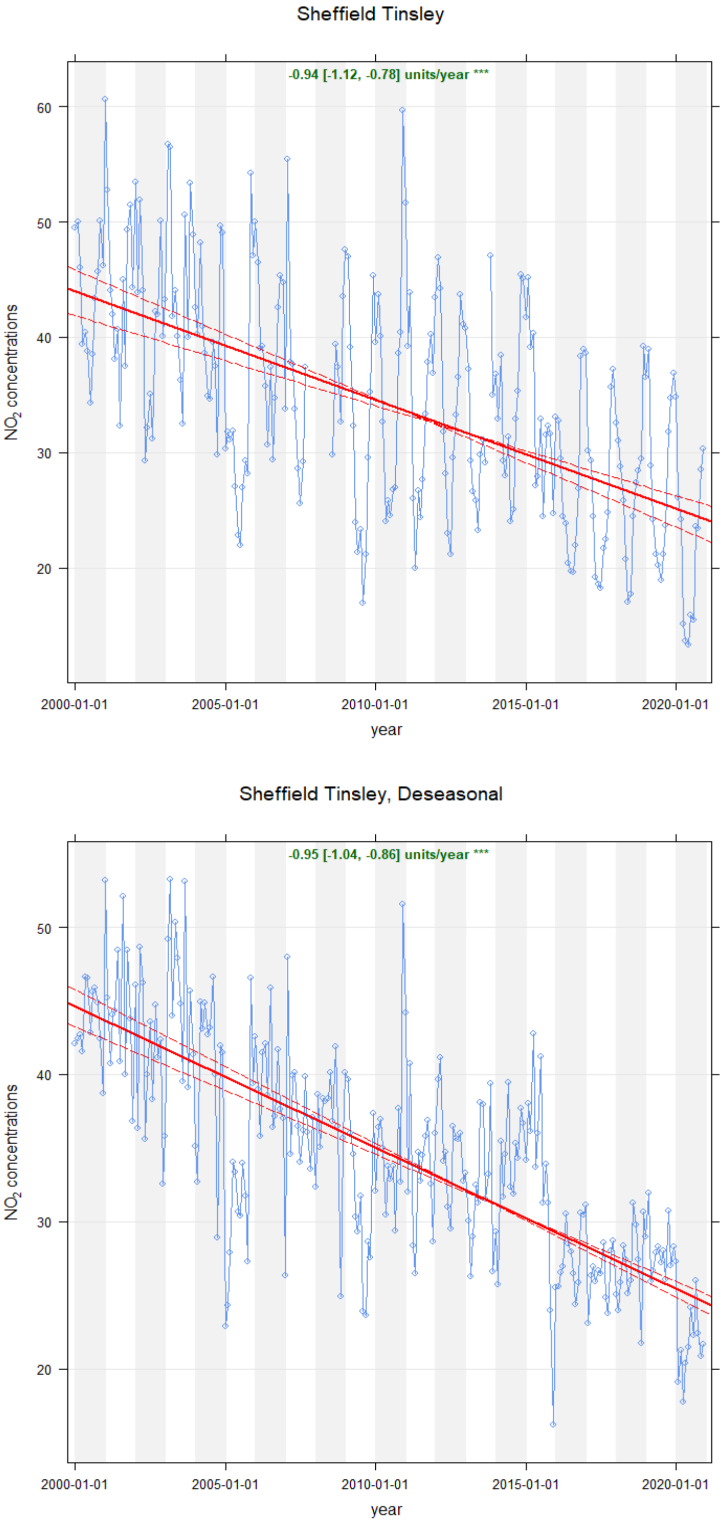


Figure 8.13. Long-term temporal trend for NO_2 concentrations (µg/m^3) (2000–2019) at the Sheffield Tinsley site, one of the AURN sites. *** shows that the trend is highly significant.

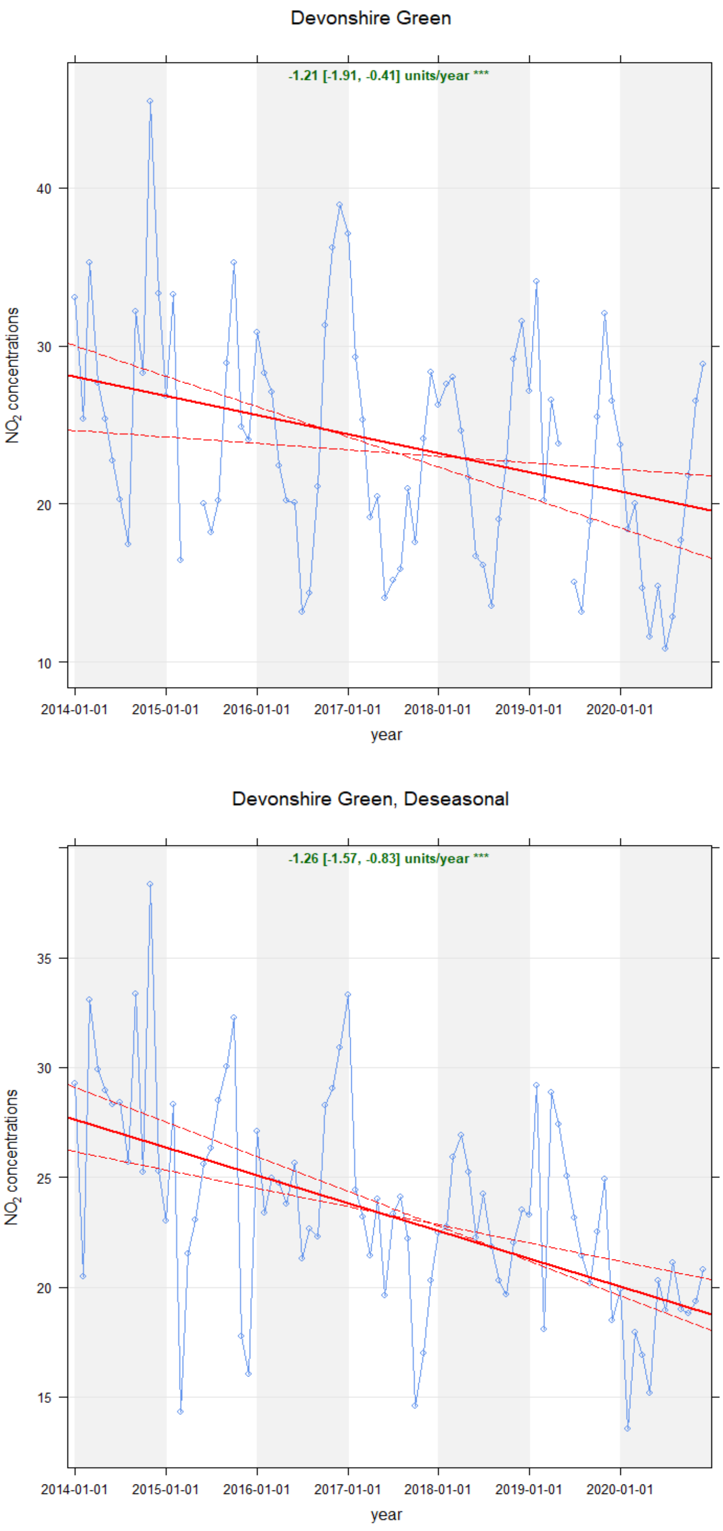


Figure 8.14. Long-term temporal trend for NO_2 concentrations (µg/m^3) (2014–2019) at the Sheffield Devonshire Green site. *** shows that the trend is highly significant.

At both AQMSs, NO_2 levels decreased during the study period. The reduction in air pollution levels could have been caused by reductions in pollutant emissions or changes in climatic conditions. However, when the trends were adjusted for the effect of changes in climatic conditions, the trends were still negative and slightly greater than the non-adjusted trends (Figures 8.13 and 8.14). This probably proves that reduction in NO_2 levels were due to reductions in emissions, showing that, despite the fact that the number of vehicles on roads has gone up, the amount of exhaust emissions has decreased. The reduction in exhaust emissions was probably caused by the stringent emission policies of the UK government. Although generally pollution levels have decreased, the reduction is not uniform spatially and temporally. For example, the trend at the Tinsley site was lower (−0.94 µgm−3/year) than at the Devonshire Green site (−1.21 µgm−3/year). However, it should be noted that at Devonshire Green the trend was determined for a shorter period. When a shorter period was also considered for Tinsley, i.e., from 2014 to 2019, the trend was −1.65 µgm−3/year. This probably shows that reductions in pollutant concentrations have been greater more recently. However, the reductions are not enough and NO_2 levels still exceed air quality standards in several parts of the city. Therefore, more actions are required to further improve air quality in Sheffield to comply with air quality guidelines.

8.3.3. Time series modelling

8.3.3.1. Comparison of linear and nonlinear time series

The performances of two univariate nonlinear persistent models, NNNAR and SETAR, were compared to that of the linear ARIMA model. Several statistical metrics were calculated to assess the performances of these models (Table 8.3). Comparing their performances in Table 8.3, it can be observed that the ARIMA model performed better than the two nonlinear models. Correlation coefficients (r-values) for ARIMA, NNNAR and SETAR were 0.59, 0.45 and 0.44 and RMSE values were 8.61, 10.45 and 10.56, respectively, which show that there is probably no need for the use of a more complicated nonlinear model for air quality prediction, as ARIMA in this particular example performed better than the nonlinear counterparts. The nonlinear models are not discussed further. For more details on SETAR and NNNAR, readers are referred to Fırat [36] and Waheeb et al. [37].

Table 8.3. Comparing the performances of different models, including both linear and nonlinear models, using the testing dataset (cross validation). MBE, MAE and RMSE are expressed in µg/m^3. r is the value of correlation coefficient.

Model	FAC2	MBE	MAE	RMSE	r
SETAR	0.90	-0.13	8.28	10.56	0.44
NNET	0.89	-0.29	8.12	10.45	0.45
ARIMA	0.91	-0.26	6.46	8.61	0.59

Here, we first discuss the univariate ARIMA model and then compare its performance with ARIMAX. The estimated parameters of the fitted ARIMA (1,1,1) model are shown in Table 8.4. The estimated parameters were used to predict NO_2 concentrations. A comparison of predicted and observed NO_2 concentrations for the testing (held-out) dataset is shown in Figure 8.15. As expected, the ARIMA model performed better when model performance was assessed based on the training dataset with r-value 0.68, MBE 0.07 and FAC2 0.93 than when the model performance was assessed using testing dataset having r-value 0.59, MBE −0.26 and FAC2 0.91.

Table 8.4. Estimating the parameters of ARIMA (1,1,1) model for NO_2 concentrations training dataset

ARIMA model applied to log_NO_2 with p, d and q order of 1, 1, 1		
Coefficients	AR1 (ϕ)	MA1(θ)
	0.5362	-0.9511
S.E.	0.0296	0.0117
Sigma square ($\boldsymbol{\sigma}^2$) estimated as 0.148: log likelihood = -556.75, AIC = 1117.5		

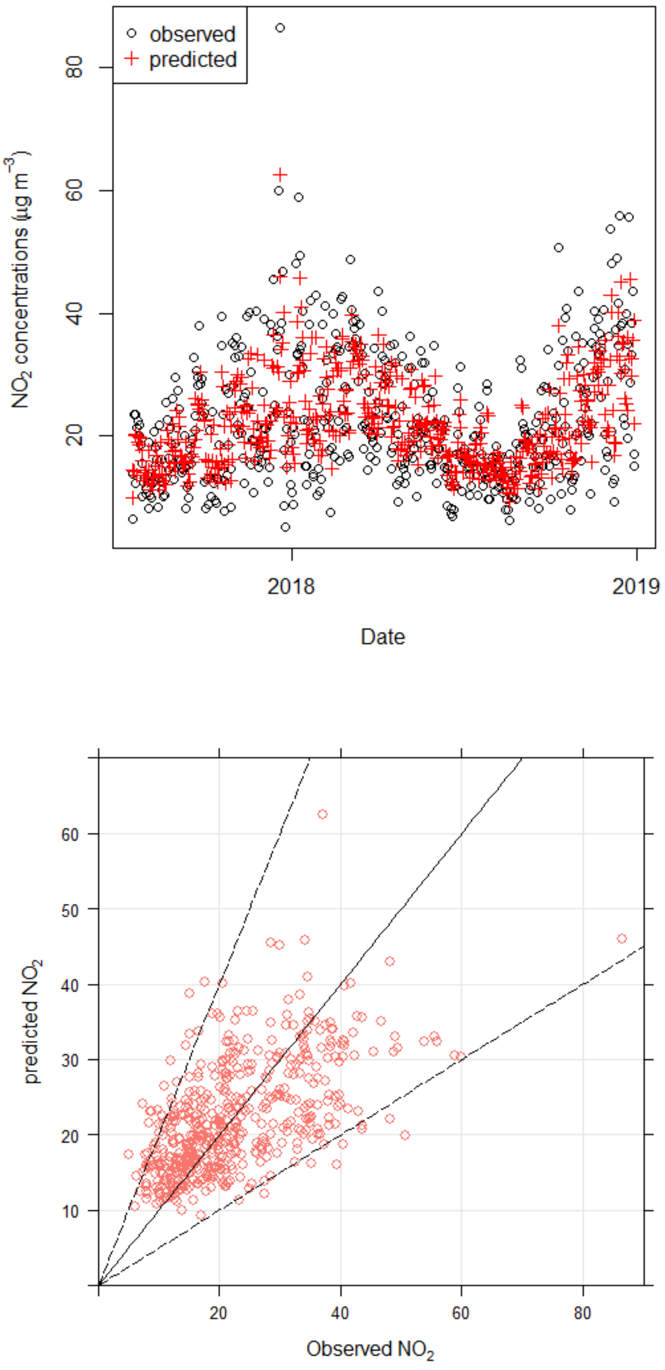


Figure 8.15. Predicted vs. observed daily average NO_2 concentrations (μg/m^3) using the ARIMA (1,1,1) model. Model performance was assessed against held-out testing data, which

was not used for model fitting. No exogenous variable was used in the model. The solid line represents the 1:1 relationship, whereas the dashed lines represent the 1:0.5 and 1:2 relationships, between observed and predicted concentrations. The dashed lines show the points that are within a factor of two (FAC2).

Residual analysis was performed for the model. A plot of standardised residuals showed a rectangular scatter around a zero horizontal level with no trends, which showed the adequacy of the model. A quantile–quantile plot of the residuals showed that the data points followed the straight line closely, which led us to accept the normality of the error term. This was also confirmed by the histogram of the residuals, which appeared closed to normal (Figure not shown for brevity).

8.3.3.2. Outputs of ARIMAX model

In addition to the autoregressive, moving average and differencing components used by univariate ARIMA, the more advanced ARIMAX model incorporates external variables, also known as exogenous regressors, into time series modelling. The exogenous variables incorporated into the ARIMAX model should be strongly correlated with the modelled variable. Here, in the first instance we used a single exogenous variable (NO concentrations) for modelling NO_2 concentrations and assessed how it improved the model performance compared to the ARIMA model developed in the previous section, which did not use any external variable. In the next step, NO and O_3 were used as exogenous variables and their effect was assessed on the model goodness of fit. Ideally, meteorological parameters such as temperature, relative humidity and wind speed should have been used in the model as well. However, due to data availability problems, they were not used as regressors in the model.

The estimated parameters of the ARIMAX (1,1,1) model are shown in Table 8.5. Various statistical metrics calculated for the fitted model and cross validation were significantly improved as compared to ARIMA model. The values of r, RMSE, MBE, MAE and FAC2 were 0.85, 7.11, 0.07, 6.53 and 0.96, respectively, for the fitted model. However, when the model performance was assessed using the testing dataset the values of r, RMSE, MBE, MAE and FAC2 were 0.70, 10.15, 7.81, − 6.84 and 0.88, respectively (Table **8.**6). For further analysis, model diagnostics were carried out by analysing the residuals of the model. The residual plots, Q–Q plots and histogram of error terms showed that the residuals were normally distributed and that the model performance was adequate (Figures are not shown for brevity).

Table 8.5. Estimating the parameters of the ARIMAX (1,1,1) model for the NO_2 concentrations ($\mu g/m^3$) training dataset, with NO as exogenous variable.

ARIMAX model applied to log_NO_2 with p, d and q order of 1, 1, 1 and xreg as NO

Coefficients	AR1 (ϕ)	MA1(θ)	XREG
	0.2533	-0.9689	0.4133

Sigma square ($\boldsymbol{\sigma}^2$) estimated as 0.056: log likelihood = 20.82, AIC =-31.64

Table 8.6. Model statistics showing the value of several metrics to assess the model performance by comparing observed and predicted NO_2 concentrations (µg/m^3) for testing and training data using NO as exogenous variables. MBE, MAE and RMSE have the unit of µg/m^3.

Statistics	FAC2	MBE	MAE	RMSE	r
Training data	0.96	0.07	5.53	7.11	0.85
Testing data	0.88	-6.84	7.81	10.15	0.70

Finally, we added two exogenous regressors (concentrations of NO and O_3) to ARIMAX model for predicting NO_2 concentrations. The estimated parameters of the ARIMAX model along with AIC value are given in Table 8.7. The values of different statistical metrics calculated for assessing the model performance are provided in Table 8.8, where r-values were 0.90 and 0.84 and RMSE values were 11.75 and 9.90 for training and testing dataset, respectively. Graphical presentation showed strong association between predicted and observed concentrations (Figure not shown for brevity). This showed that the exogenous variables helped improve the model performance significantly compared to univariate persistent models.

Table 8.7. Estimating the parameters of ARIMAX (1,1,1) model for NO_2 concentrations (µg/m^3) training dataset with NO and O_3 as exogenous variables.

ARIMAX model applied to log_NO_2 with p, d and q order of 1, 1, 1 and xreg as NO and O_3				
Coefficients	AR1 (ϕ)	MA1(θ)	XREG2 (NOx)	XREG3 (O_3)
	0.21	-0.985	0.839	-0.108
Sigma square (**σ**^{2}) estimated as 0.014: log likelihood = 866.86, AIC =-1721.73				

Table 8.8. Statistical metrics assessing the model performance by comparing observed and predicted concentrations for both training and testing dataset using NO and O_3 as exogenous variables. . MBE, MAE and RMSE have the unit of µg/m^3.

Statistics	FAC2	MBE	MAE	RMSE	r
Training data	0.65	-10.25	10.45	11.75	0.90
Testing data	0.73	-7.94	9.34	9.90	0.84

The above analysis showed that traditional time series persistent models, e.g., ARMA or ARIMA are useful tools for air pollution analysis and prediction, however they only depend on the behaviour of the past data without taking into account the effect of other pollutants and meteorological parameters that interact with the modelled pollutant. Therefore, these approaches fail to predict future levels accurately. The more advanced versions of these models like ARIMAX are able to analyse the effect of environmental factors and other pollutants and can result in better prediction by reducing the model error and strengthening correlation between modelled and observed concentrations.

The association of NO_2 concentration with NO and O_3 concentrations is shown in the form of scatter plot (Figure 8.16). Chemistry of NO_2 with O_3 and NO is shown in equation 8.6 and 8.7.

$$NO + O_3 \rightarrow NO_2 + O_2 \qquad (8.6)$$

$$NO_2 + h\nu\ (+O_2) \rightarrow NO + O_3 \qquad (8.7)$$

Equation 6 and 7 define the chemistry of NO, NO_2 and O_3. In equation 6 one molecule of O_3 is consumed, and one molecule of NO_2 is produced. In contrast, in equation 8.7 one molecule of NO_2 is consumed and one molecule of O_3 is produced, so no net chemistry occurs. NO_2 is positively correlated with NO (r = +0.75) and NOx (r = +0.89) and negatively correlated with O_3 (r = -0.68). The negative association between NO_2 and O_3 is well known (e.g., Jenkin, 2004; Clapp and Jenkin, 2001; Munir, 2012). It shows that NO_2 concentration is strongly correlated with O_3 and other NOx species. Therefore, adding O_3 and NO as exogenous regressors in the model can explain a significant proportion of NO_2 and improves the model performance.

Catalano et al. [47] developed an ARIMAX model and compared its performance to Artificial Neural Network (ANN). They modelled NO_2 concentrations at Marylebone road London using several exogenous variables, which were traffic volume, wind speed, wind direction, temperature and lagged NO_2 concentrations. According to Catalano et al. [47] ARIMAX model performed better than ANN. The mean ratio of the predicted to the measured concentrations was 0.89 for ARIAMX and 0.86 for ANN. Both models had the same correlation coefficient 0.91, however, mean absolute percentage error (MAPE) values were 18.62 and 16.53 for ARIMAX and ANN, respectively. This shows that the prediction of ARIMAX model is comparable with neural network and can be used for air pollution forecast in urban areas to provide timely warning of high air pollution levels to the public.

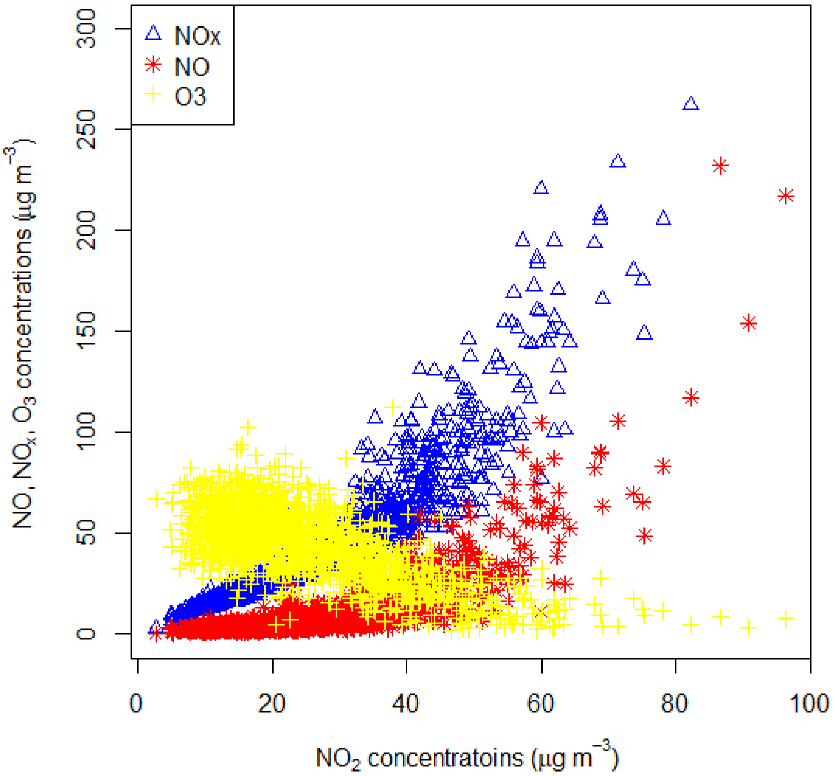


Figure 8.16. Scatter plot showing the association of NO_2 with NO, NOx and O_3 concentrations (µg/m^3) at the Devonshire Green AQMS in Sheffield.

8.4. Conclusion

We have a network of AQMSs in Sheffield consisting of low-cost and reference sensors. Here, NO_2 concentrations measured by Envirowatch E-MOTEs, AQMesh pods and AURN and SCC AQMSs were analysed to characterise the temporal variability of NO_2 concentrations. Density plots are a useful tool for comparing and analysing the distributions of NO_2 concentrations measured by various sensors. Time variation plots were employed to characterise and compare the temporal variability of NO_2 concentrations measured at different monitoring sites. Time variation plots visualise how NO_2 concentrations vary during different time periods (e.g., diurnal, weekly and annual cycles) and help us understand their emission sources. Long-term data for NO_2 concentrations showed negative trends at both the Sheffield Tinsley and Devonshire Green sites, indicating that pollutant emissions decreased because of stringent emission policies. However, the reductions in pollution varied both spatially and temporally. Moreover, further smart interventions are required to cut emissions and improve air quality to comply with air quality guidelines.

NO_2 concentrations were modelled from the Devonshire Green AQMS in Sheffield using univariate linear and nonlinear and multivariate time series models with exogenous variables. The addition of exogenous variables to the ARIMAX model significantly improved the model performance. Model specification, fitting and diagnostics were discussed. Model performance was assessed by calculating several statistical metrics, including r, RMSE, MBE, FAC2 and MAE. The ARIMA model showed better performance than nonlinear persistent models, showing that linear models were not only easy to apply and interpret but also showed better performance than the more complicated models such as SETAR and NNNAR. The best model fit was achieved when NO_2 concentration was modelled using ARIMAX with two exogenous variables (NO and O_3). These variables were strongly correlated with NO_2. NO was positively correlated, whereas O_3 was negatively correlated with NO_2.

Time series models with exogenous regressors can be successfully used to predict future pollution concentrations, providing early warning for the public so that timely precautionary measures can be adopted. The main weakness of the study was that we did not provide full details of the low-cost sensors calibration. Ideally, several of these low-cost sensors should have been collocated with reference sensors for detailed comparison and calibration. However, due to several practical problems (e.g., planning permission), this was not possible and the only calibration carried out was described in Munir et al. [44]. They compared the measurements of Envirowatch E-MOTEs with the measurements of an AURN site (Sheffield Devonshire Green). Future work will include installation of low-cost sensors, collocated with reference sensors for over a year, and a comparison of their measurements during different time periods. This study provides a detailed methodology for time series modelling, which can be used as an early warning tool for air pollution episodes, and compares different linear and nonlinear modelling approaches.

Author Contributions: Conceptualization, S.M.; methodology, S.M.; validation, S.M.; formal analysis, S.M.; investigation, S.M.; resources, S.M. and M.M.; data curation, S.M.; writing-original draft preparation, S.M.; writing-review and editing, S.M. and M.M.; visualization, S.M.; supervision, M.M. project administration, S.M. and M.M.; funding acquisition, M.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Engineering and Physical Sciences Research Council (EPSRC) (grant number - EP/R512175/1) and Siemens plc. The APC was paid by UKRI/EPSRC.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in reference number [44,48].

Acknowledgments: We would like to express our thanks to the Engineering and Physical Sciences Research Council (EPSRC), grant number—EP/R512175/1, and Siemens plc for funding this project.

Conflicts of Interest: All the authors declare no conflicts of interest.

8.5. References

1. Landrigan, P.J. Air pollution and health. Lancet Public Health 2017, 2, E4–E5. [CrossRef]
2. WHO. World Health Organization, Review of Evidence on Health Aspects of Air Pollution; REVIHAAP; WHO: Copenhagen, Denmark, 2013.

3. Walters, S.; Ayres, J. The health effects of air pollution. In Pollution: Causes, Effects and Control, 4th ed.; Chapter 11; Harrison, R.M.,Ed.; Royal Society of Chemistry: Cambridge, UK, 2001; p. 275. ISBN 0-85404-621-6.
4. Azam, G.A.; Zanjani, R.B.; Mood, B.M. Effects of air pollution on human health and practical measures for prevention in Iran. J.Res. Med. Sci. 2016, 21, 65.
5. Manisalidis, I.; Stavropoulou, E.; Stavropoulos, A.; Bezirtzoglou, E. Environmental and Health Impacts of Air Pollution: A Review. Front. Public Health 2020, 8, 14. [CrossRef]
6. Khallaf, M. (Ed.) The Impact of Air Pollution on Health, Economy, Environment and Agricultural Sources; InTech: London, UK, 2011. [CrossRef]
7. Ivaskova, M.; Kotes, P.; Brodnan, M. Air Pollution as an Important Factor in Construction Materials Deterioration in Slovak Republic. Procedia Eng. 2015, 108, 131–138. [CrossRef]
8. Aldrin, M.; Haff, I. Generalised additive modelling of air pollution, traffic volume and meteorology. Atmos. Environ. 2005, 39, 2145–2155. [CrossRef]
9. Andersen, S.B.; Weatherhead, E.C.; Stevermer, A.; Austin, J.; Brühl, C.; Fleming, E.L.; de Grandpré, J.; Grewe, V.; Isaksen, I.; Pitari, G.; et al. Comparison of recent modeled and observed trends in total column ozone. J. Geophys. Res. Space Phys. 2006, 111. [CrossRef]
10. Arnold, S.R.; Chipperfield, M.P.; Blitz, M.A. A three-dimensional model study of the effect of new temperature-dependent quantum yields for acetone photolysis. J. Geophys. Res. Space Phys. 2005, 110. [CrossRef]
11. Baur, D.G.; Saisana, M.; Schulze, N. Modelling the effects of meteorological variables on ozone concentration—a quantile regression approach. Atmos. Environ. 2004, 38, 4689–4699. [CrossRef]
12. Brasseur, G.P.; Hauglustaine, D.A.; Walters, S.; Rasch, P.J.; Müller, J.-F.; Granier, C.; Tie, X.X. MOZART, a global chemical transport model for ozone and related chemical tracers: 1. Model description. J. Geophys. Res. Space Phys. 1998, 103, 28265–28289. [CrossRef]
13. Munir, S.; Chen, H.; Ropkins, K. Modelling the impact of road traffic on ground level ozone concentration using a quantile regression approach. Atmos. Environ. 2012, 60, 283–291. [CrossRef]
14. Westmoreland, E.J.; Carslaw, N.; Carslaw, D.C.; Gillah, A.; Bates, E. Analysis of air quality within a street canyon using statistical and dispersion modelling techniques. Atmos. Environ. 2007, 41, 9195–9205. [CrossRef]
15. Wilkening, I.H.; Baraldi, D. CFD modelling of accidental hydrogen release from pipelines. Int. J. Hydrog. Energy 2007, 32, 2206–2215. [CrossRef]
16. Adhikari, R.; Agrawal, K.R. An Introductory Study on Time Series Modeling and Forecasting; LAP Lambert Academic Publishing: Saarbrücken, Germany, 2013. Available online: https://arxiv.org/abs/1302.6613 (accessed on 28 May 2019).
17. Bush, T. Time Series Analysis: Definition, Benefits, Models. 2020. Available online: https://pestleanalysis.com/time-seriesanalysis/ (accessed on 21 March 2021).
18. Srivastava, T.A. Complete Tutorial on Time Series Modeling in R. 2015. Available online: https://www.scribd.com/document/343194193/A-Complete-Tutorial-on-Time-Series-Modeling-in-R (accessed on 19 May 2019).
19. Hipel, W.K.; McLeod, I.A. Time Series Modelling of Water Resources and Environmental Systems; Elsevier: Amsterdam, The Netherlands, 1994.
20. Cochrane, J.H. Time Series for Macroeconomics and Finance; Graduate School of Business, University of Chicago: Chicago, IL, USA, 1997. Available online:

https://static1.squarespace.com/static/5e6033a4ea02d801f37e15bb/t/5ed92dcb7665261af1aa23f2/1591291342389/time_series_book.pdf (accessed on 10 April 2021).

21. Kadiyala, A.; Kumar, A. Multivariate Time Series Models for Prediction of Air Quality Inside a Public Transportation Bus Using Available Software. Environ. Prog. Sustain. Energy 2012, 33, 337–341. [CrossRef]

22. Leontaritis, I.; Billings, S. Input output parametric models for nonlinear systems. Int. J. Control. 1985, 41, 303–344. [CrossRef]

23. Billings, S.A.; Chen, S. Extended model set, global data and threshold model identification for severely nonlinear systems. Int. J. Control. 1989, 50, 1897–1923. [CrossRef]

24. DEFRA. Improving Air Quality in the UK Tackling Nitrogen Dioxide in Our Towns and Cities. UK Overview Document. 2015. Available online: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/486636/aq-plan-2015-overview-document.pdf (accessed on 9 October 2020).

25. Brunt, H.; Barnes, J.; Longhurst, J.; Scally, G.; Hayes, E. Local Air Quality Management policy and practice in the UK: The case for greater Public Health integration and engagement. Environ. Sci. Policy 2016, 58, 52–60. [CrossRef]

26. DEFRA. Improving Air Quality in the UK: Tackling Nitrogen Dioxide in Our Towns and Cities. In Draft UK Air Quality Plan for Tackling Nitrogen Dioxide.; 2017. Available online: https://consult.defra.gov.uk/airquality/air-quality-plan-for-tacklingnitrogendioxide/supporting_documents/Draft%20Revised%20AQ%20Plan.pdf (accessed on 5 October 2017).

27. Borrego, C.; Costa, A.M.; Ginja, J.; Amorim, M.; Coutinho, M.; Karatzas, K.; Sioumis, T.; Katsifarakis, N.; Konstantinidis, K.; deVito, S.; et al. Assessment of air quality microsensors versus reference methods: The EuNetAir joint exercise. Atmos. Environ. 2016, 147, 246–263. [CrossRef]

28. Mead, M.I.; Popoola, O.A.M.; Stewart, G.B.; Landshoff, P.; Calleja, M.; Hayes, M.; Baldovi, J.J.; McLeod, M.W.; Hodgson, T.F.;Dicks, J.; et al. The use of electrochemical sensors for monitoring urban air quality in low-cost, highdensity networks. Atmos.Environ. 2013, 70, 186–203. [CrossRef]

29. The E-MOTE–Air Quality and Noise Pollution Monitoring. Available online: http://www.envirowatch.ltd.uk/e-mote/ (accessed on 10 April 2021).

30. AQMesh|The Best Small Sensor Air Quality Monitoring System. Available online: https://www.aqmesh.com/ (accessed on 10 April 2021).

31. R Core Team. R: A Language and Environment for Statistical Computing; Version 3.5.2; R Foundation for Statistical Computing: Vienna, Austria, 2019. Available online: https://www.R-project.org/ (accessed on 23 October 2019)Version 3.5.2.

32. Carslaw, D.C.; Ropkins, R. Openair: An R package for air quality data analysis. Environ. Model. Soft. 2012, 27, 52–61. [CrossRef]

33. Munir, S.; Chen, H.; Ropkins, K. Characterising the temporal variations of ground level ozone and its relationship with trafficrelated air pollutants in the UK: A quantile regression approach. Int. J. Sustain. Dev. Plan. 2014, 9, 29–41. [CrossRef]

34. Gerbing, D. Time Series Components. School of Business Administration Portland State University. 2016. Available online: http://web.pdx.edu/~{}gerbing/515/Resources/ts.pdf (accessed on 22 March 2021).

35. Cleveland, B.R.; Cleveland, S.W.; McRae, E.J.; Terpenning, I. STL: A Seasonal-Trend Decomposition Procedure Based on Loess. J. Off. Stat. 1990, 6, 3–33.

36. Firat, E.H. SETAR (Self-exciting Threshold Autoregressive) Non-linear Currency Modelling in EUR/USD, EUR/TRY and USD/TRY Parities. Math. Stat. 2017, 5, 33–55. [CrossRef]

37. Waheeb, W.; Ghazali, R.; Shah, H. Nonlinear Autoregressive Moving-average (NARMA) Time Series Forecasting Using Neural Networks. In Proceedings of the 2019 International Conference on Computer and Information Sciences (ICCIS), Aljouf, Saudi Arabia, 3–4 April 2019. [CrossRef]

38. Di-Narzo, F.A.; Aznarte, L.J.; Stigler, M. tsDyn: Nonlinear Time Series Models with Regime Switching. Version 0.9-48.1. 2019. Available online: https://cran.r-project.org/web/packages/tsDyn/tsDyn.pdf (accessed on 4 September 2019).

39. Cryer, D.J.; Chan, S.K. Time Series Analysis with Applications in R; Springer Texts in Statistics; Springer: New York, NY, USA, 2008;ISBN 978-0-387-75958-6.

40. Chan, S.K.; Ripley, B. TSA: Time Series Analysis. R Package Version 1.2. 2018. Available online: https://CRAN.R-project.org/package=TSA (accessed on 2 May 2019).

41. Hyndman, R.; Athanasopoulos, G.; Bergmeir, C.; Caceres, G.; Chhay, L.; O'Hara-Wild, M.; Petropoulos, F.; Razbash, S.; Wang, E.;Yasmeen, F. Forecast: Forecasting Functions for Time Series and Linear Models. R Package Version 8.5. 2019. Available online: http://pkg.robjhyndman.com/forecast (accessed on 10 September 2019).

42. Carslaw, D. Defra Regional and Transboundary Model Evaluation Analysis–Phase 1. Version: 15 April 2011. Available online: https://uk-air.defra.gov.uk/assets/documents/reports/cat20/1105091514_RegionalFinal.pdf (accessed on 12 April 2021).

43. Sayegh, A.S.; Munir, S.; Habeebullah, T.M. Comparing the Performance of Statistical Models for Predicting PM10 Concentrations. Aerosol Air Qual. Res. 2014, 14, 653–665. [CrossRef]

44. Munir, S.; Mayfield, M.; Coca, D.; Jubb, S.A.; Osammor, O. Analysing the performance of low-cost air quality sensors, their drivers, relative benefits and calibration in cities—a case study in Sheffield. Environ. Monit. Assess. 2019, 191, 1–22. [CrossRef]

45. Clapp, J.L.; Jenkin, E.M. Analysis of the relationship between ambient levels of O3, NO_2 and NO as a function of NOx in the UK. Atmos. Environ. 2001, 35, 6391–6405. [CrossRef]

46. Jenkin, M.E. Analysis of sources and partitioning of oxidant in the UK—Part 1: The NOX-dependence of annual mean concentrations of nitrogen dioxide and ozone. Atmos. Environ. 2004, 38, 5117–5129. [CrossRef]

47. Catalano, M.; Galatioto, F.; Bell, M.; Namdeo, A.; Bergantino, A.S. Improving the prediction of air pollution peak episodes generated by urban transport networks. Environ. Sci. Policy 2016, 60, 69–83. [CrossRef]

48. Munir, S.; Mayfield, M.; Coca, D.; Mihaylova, L.S. A nonlinear land use regression approach for modelling NO_2 concentrations in urban areas—using data from low-cost sensors and diffusion tubes. Atmosphere 2020, 11, 736. [CrossRef]

CHAPTER 9: CONCLUSIONS AND FUTURE WORK

9.1. Introduction

Poor air quality (AQ) is a growing environmental issue especially in large urban areas causing various negative impacts on our natural environment, buildings, vegetation and human health. Air pollutant levels have declined in the UK for the last couple of decades, however the reductions are not sufficient to comply with air quality guidelines and therefore air pollutant levels especially NO_2 and PM_{10} still exceed AQ standards in large urban areas in the UK. In the UK Air Quality Management Areas have been declared by many local authorities due to the high levels of traffic related air pollutants, for example NO_2. Therefore, to reduce air pollution levels in urban areas to comply with AQ guidelines, further actions are required including real time AQ monitoring, modelling and introducing different interventions for cutting emissions. This project focused on air quality monitoring and modelling to analyse spatial and temporal variability of NO_2 concentrations in urban areas as a case study in Sheffield.

Firstly, a detailed literature review was conducted to understand the state of the art of air quality monitoring sensors, spatial and temporal modelling of air quality in urban areas, and data fusions approaches. In light of the literature review, research gaps were identified and objectives of the project were defined. The whole idea of the project revolves around the types of low-cost sensors, calibration of low-cost sensors, structuring a dense multipurpose air quality monitoring network, using data from the air quality monitoring network develop and validate models for studying the spatial variability of NO_2 in Sheffield. Furthermore, data fusion techniques are employed to integrate model estimations with measured NO_2 concentrations to improve data quality and produce high-resolution NO_2 maps. Finally using data from both low-cost and reference sensors temporal variability of NO_2 is analysed.

This PhD thesis is composed in the alternative format ‘publication format thesis’ incorporating a collection of papers that are already published in peer review journals as open access. Therefore, it is important to clearly present the structure of the thesis first and show how different chapters of the thesis are linked together.

9.2. Alternative format thesis

This PhD thesis is written in publication format, referred to as ‘alternative format thesis’, in which each published paper makes a separate chapter. The papers presented in this thesis are already published and are available online. However, in addition to the published papers, two chapters are added in the beginning of the thesis as an introduction and one chapter at the end to summarise the thesis. Chapter 1 introduces the motivation of the project, defines the main aims and objectives and shows how the objectives are addressed. Chapter 2 consists a detailed literature review to understand state of the art of the air quality monitoring, modelling and data fusion approaches. Chapters 3, 4, 5, 6, 7, and 8 are made of the published papers. Chapter 9 (this chapter) summarises the whole thesis and describes how the objectives are addressed in this thesis. Table 9.1 summarises how the research gaps identified in literature review are addressed in this thesis.

Table 9.1. Showing how different research gaps identified in the literature review have been addressed in the thesis.

Research gap	How are the research gaps addressed?
Literature review demonstrated that most of the purpose-designed air quality networks have been structured on ado basis, without using a formal approach for selecting sites for the sensors deployment. In addition, the criteria of structuring the network are currently defined by single questions rather than attempting to create a network to serve multiple functions. There is a need for developing a formal approach based on multi-criteria for selecting the sites for the sensors installation.	In this thesis, a formal approached is proposed for selecting the sites for sensors installation. In this thesis, we proposed a methodology supported by numerical, conceptual and GIS frameworks for structuring AQMN using social, environmental and economic indicators as a case study in Sheffield. The main factors used for the selection of air quality monitoring stations were population-weighted pollution concentration (PWPC) and weighted spatial variability (WSV) incorporating population density (social indicator), pollution levels and spatial variability of air pollutant concentrations (environmental indicator). Total number of sensors was decided on the basis of budget (economic indicator), whereas the number of sensors deployed in each output area was proportional to weighted spatial variability.
Low-cost sensors are a useful tool for air quality monitoring. However, they require outfield calibration. Furthermore, it was found that previously most of the sensor calibration studies whether in the laboratories or outdoor fields were carried out for a short period. Therefore, further research was required to compare the performance of low-cost sensors to each other and to reference sensor in outdoor fields for a longer period of time at least for a year.	In this thesis, ten low-cost sensors known as Envirowatch E-MOTEs were deployed for a year for monitoring several air pollutants and meteorological parameters. The air quality measurements of these sensors were compared to each other and to a reference sensor installed nearby. The low-cost sensors were able to successfully capture the temporal variability such as diurnal, weekly and annual cycles in air pollutant concentrations and demonstrated significant similarity with reference instruments. However, NO_2 and CO concentrations measured by low-cost sensors showed stronger positive correlation with each other than NO concentrations. To further improve the quality of measurements made by low-cost sensors, several calibrations models were employed, including both linear (multiple linear regression) and nonlinear (generalised additive) models were developed to calibrate the E-MOTE data and reproduce NO and NO_2 concentrations measured by the reference instruments. The nonlinear model demonstrated significantly better performance than linear model by capturing the nonlinear association between the response and explanatory variables. The best model developed for reproducing NO_2 concentrations returned values of 3.91

	and 0.81 for root mean squared error (RMSE) and coefficient of determination, respectively.
Literature review revealed that previously several approaches have been employed for modelling the spatial variability of air pollutants in urban area. However, little was done on comparing different modelling approaches for analysing the spatial variability of pollutants in urban areas.	In this PhD thesis, we employed two approaches to model NO_2 concentrations in Sheffield and compared the models performance. The modelling approaches were Airviro dispersion modelling system and Land-Use Regression (LUR) model. The performance of these models were compared for predicting NO_2 concentrations. Several LUR models were developed using different modelling techniques and different datasets. The Generalised Additive Model (GAM) that was trained and validated by using data from NO_2 diffusion tubes showed better performance than the other LUR counterparts using linear regression. The GAM-LUR model also outperformed the Airviro model. Correlation coefficient and RMSE values were 0.70 and 8.69 for the GAM-LUR model, and 0.48 and 44.01 for the Airviro model, respectively. Comparison of the two models was not possible for PM_{10}, as no LUR model was developed for PM_{10} due to the lack of measured PM_{10} concentrations to train the LUR model. Therefore, it is concluded that GAM-LUR model performed better at urban scale and should be preferred over commercially available Airviro dispersion modelling system for modelling NO_2 concentrations in urban areas.
Literature review revealed that the application of data fusion techniques improved the quality of data; however, little is done to demonstrate how these techniques are helpful in improving the quality of estimated air quality data predicted by spatial modelling in urban areas.	To show how data fusion techniques can improve the quality of air quality estimated by spatial modelling techniques, in this thesis firstly three modelling approaches were used to produce maps of NO_2 concentrations in Sheffield. The techniques were geostatistical kriging interpolation, Airviro dispersion modelling, and LUR based on the generalised additive model. Measured NO_2 concentrations were fused with the estimated concentrations using universal kriging, which is an advanced kriging technique that is employed for data having more than one correlated variable. Six sets of measured and estimated NO_2 data were fused: (i) Fusion of Airviro estimations with LCS measurements; (ii) Fusion of Airviro estimations with diffusion tube measurements; (iii) Fusion of Airviro estimations with measurements of both LCS and diffusion tubes; (iv) Fusion of LUR estimations with LCS measurements; (v) Fusion of LUR estimations with diffusion tubes measurements; and (vi) Fusion of LUR measurements with both LCS and diffusion tubes measurements. Fused NO_2 were compared with measured concentrations. Maps produced by the fusion of LCS measurements and LUR estimations produced better results, with an r-value

	of 0.96, followed by the fusion of Airviro estimations with LCS measurements, with an r-value of 0.88, when compared to reference sensors. Fused maps developed by the fusion techniques produced better spatial coverage and better similarity with measured concentrations than the estimations of ordinary kriging interpolation, Airviro and LUR models. Therefore, the data fusion techniques were able to improve the air quality data quality and produced high-resolution maps.
How the concentrations of NO_2 measured by low-cost sensors vary on different temporal scales in comparison to reference sensors in urban areas; and what is the best time series model for modelling NO_2 concentrations.	The diurnal, weekly and annual cycles of low-cost sensors, especially Envirowatch E-MOTEs followed almost similar pattern as those shown by the reference sensors. However, there were some minor differences between various low-cost sensors installed at different locations within the city. In addition, differences in winter vs. summer and weekend vs. weekday concentrations were much prominent in reference sensors than in low-cost sensors. The annual cycles of NO_2 measured by AQMesh were different from each other and from the ones shown by the reference sensors. NO_2 concentrations were analysed employing univariate linear and nonlinear time series models and their performance was compared with a more advanced time series model using two exogenous variables (NO and O_3). For this purpose, time series data of NO, O_3 and NO_2 were obtained from a reference site in Sheffield, which are more accurate than the measurements from LCS and, therefore, more suitable for training and testing the model. Interesting, the linear auto-regressive integrated moving average (ARIMA) performed better than the nonlinear counterpart did; however, its performance in predicting NO_2 concentration was inferior to ARIMA with exogenous variables (ARIMAX). Using cross-validation ARIMAX demonstrated strong association with the measured concentrations, with a correlation coefficient of 0.84 and RMSE of 9.90. ARIMAX can be used as an early warning tool for predicting potential pollution episodes in order to be proactive in adopting precautionary measures.

9.3. Summary of the thesis

Air pollution levels in urban areas are driven by emission sources and influencing factors that control the dispersion of air pollutant emitted by different sources. For effective air quality management, the first and probably most important step is monitoring of the current levels of air pollutants and understanding the main drivers of air pollution in urban areas. Air pollutant levels demonstrate spatial variability at micro-levels in urban areas due to differences in emission sources and influencing factors that affect air pollutants dispersion such as meteorological parameters, building density and height, presence of open spaces, green areas, and street canyons. The current air quality monitoring network operated by DEFRA and Sheffield City Council is not dense enough to account for micro-level spatial variations in air pollutant concentrations. In Sheffield there are three air quality monitoring stations (AQMS) run by DEFRA and five run by Sheffield City Council. This suggests a need for further improvement in AQMN in Sheffield, supported by modelling studies. Until more recently, it was not possible to develop a dense air quality monitoring network in urban areas because of the high purchase prices of reference air quality sensors. However, since the emergence of low-cost sensors, it has become feasible to develop a dense multipurpose air quality monitoring network. The main objectives of the project are: (i) To structure a dense multipurpose AQMN made of several layers of air quality sensors including reference, low-cost sensors (LCS) and IoT sensors; (ii) To use LCS to render high quality measurements; (iii) To model the spatial variability of NO_2 employing LUR and dispersion models. ; (iv) To employ novel data fusion techniques for improving both model estimations and LCS measurements.; and (v) To analyse temporal trends of AQ employing TheilSen function, time series modelling and graphical presentations. Motivation and main objectives of the project are described in Chapter 1, which also describes the main drivers of air pollution and how different chapters (papers) address the objectives of the project.

Chapter 2 provides a detailed literature review of LCS, various AQ modelling, AQ monitoring and data fusion approaches. The aim is to find out as to what is the state of the art for AQ modelling, data fusion and AQ monitoring especially LCS in light of the current literature. AQ modelling is divided into two main types: dispersion modelling and statistical modelling. Each of the modelling type is further divided into several subcategories. Several research gaps have been identified in light of the literature review. The main research gaps identified are:

- It is a critical decision to make what criterion should be used for identifying a suitable site for a sensor installation. However, literature shows that in most of the network developed previously sensors are installed on ad hoc basis or based on a single criterion without following a formal approach. To address this, in this project we deployed a dense network of AQ sensors based on multiple criteria using sensors of different grades. This is related to objective 1 of the project.
- From the literature review, it was found that previously most the sensor calibration studies whether in the laboratories or outdoor fields were carried out for a short period of time mostly employing linear regression. Therefore, further research was required to compare the performance of low-cost sensors to each other and to reference sensor in outdoor fields for a longer period of time at least for a year, and develop a calibration approach employing a nonlinear regression model. This is related to objective 2**.**

- The literature review showed that previously several approaches have been employed for modelling the spatial variability of air pollutants in urban area. However, little was done on comparing different modelling approaches for analysing the spatial variability of pollutants in urban areas. This is related to objective 3 of the project.
- Literature review revealed that the application of data fusion techniques improve the quality of data, howver, little is done to demonstrate how these techniques help in improving the spatial modelling of air quality in urban areas using estimated and measured concentrations of LCS. This is related to objective 4.
- How the concentrations of NO_2 measured by LCS and reference sensors vary on different temporal scales in urban areas and what is the best time series model for modelling NO2 concentrations. This is related to objective 5 of the project.

In chapter 3, the performance of LCS, viz., Envirowatch EMOTEs was compared to each other and to a reference sensor installed nearby. The comparison was made for a full year for various gaseous pollutants, namely NO, NO_2 and CO. E-MOTEs were able to successfully capture the temporal variability such as diurnal, weekly and annual cycles in air pollutant concentrations and demonstrated significant similarity with reference instruments. NO_2 concentrations showed very strong positive correlation between various sensors. Mostly correlation coefficients were greater than 0.92 between different E-MOTEs. CO concentrations measured by different E-MOTES also had r - values mostly greater than 0.92, however, NO concentrations showed r - values less than 0.5. Demonstrating consistency of measurements measurement by different E-MOTEs. Several Multiple Linear Regression Models (MLRM) and Generalised Additive Models (GAM) were developed to calibrate the E-MOTE data and reproduce NO and NO_2 concentrations measured by the reference instruments using the measurements of E-MOTEs as predictors. The performance of MLRM and GAM was compared, where GAMs demonstrated significantly better performance than linear models by capturing the nonlinear association between the response and explanatory variables. The best GAM developed for reproducing NO_2 concentrations returned values of 0.95, 3.91, 0.81, 0.005, and 0.61 for Factor of two (FAC2), Root Mean Square Error (RMSE), coefficient of determination (R^2), Normalised Mean Biased (NMB) and Coefficient of Efficiency (COE), respectively, when predicted and measured concentrations of reference sensors were compared. The values of these metrics demonstrated that GAM calibration model could result in better results. This analysis showed that the LCS offered a more affordable alternative for providing real time high-resolution spatiotemporal AQ and meteorological parameter data with acceptable performance, which can be further improved by nonlinear calibration modelling techniques.

Reference sensors are very expensive to purchase, therefore, traditionally AQMN in urban areas are sparse and not dense enough for developing high-resolution air quality maps in urban areas. Since the emergence of low-cost sensors, it has become possible to create a purpose-designed dense network. Recently several researchers have designed AQMN, however, these networks are based on a single criterion or sensors are installed on ad hoc basis. Chapter 4 intended to address the issue of sparse AQMN by structuring a high-density network using various layers of sensors of different quality including reference instruments operated by DEFRA and Sheffield City Council, LCS deployed by the university of Sheffield and NO2 diffusion tubes. Here we propose a methodology supported by numerical, conceptual and GIS frameworks for structuring an AQMN using social, environmental and economic indicators as a case study in Sheffield, UK. The main factors used for AQMS site selection are population-weighted pollution concentration (PWPC) and weighted spatial variability (WSV) incorporating population density (social indicator), pollution levels and spatial variability of air pollutant concentrations (environmental indicator). Total number of sensors is decided on the basis of budget (economic indicator), whereas the number of sensors deployed in each output area is proportional to WSV. The purpose of AQ monitoring and its role in determining the location of AQMS is analysed. Furthermore, the existing AQMN is analysed and an alternative proposed following a formal procedure. In contrast to traditional networks, which are structured based on a single AQ monitoring approach, the proposed AQMN has several layers of sensors: Reference sensors recommended by EU and DEFRA, LCS (AQMesh and Envirowatch E-MOTEs) and IoT (Internet of Things) sensors. The core aim is to structure an integrated AQMN in urban areas, which will lead to the collection of AQ data with high spatiotemporal resolution. The use of LCS in the proposed network provides a cheaper option for setting up a purpose-designed network for greater spatial coverage. This is a new trend in AQ monitoring and can support the conventional AQMN. Data collected by the LCS can be used for detailed spatial mapping of air pollution, especially over small areas such as an urban area or a part of it, for atmospheric model validation, and for assessing population exposure. However, the robustness of LCS and the quality of the data collected is questionable and require outfield calibration, which was discussed in chapter 3.

In chapter 5, two air pollutants, viz., NOx and PM_{10} are monitored and modelled employing Airviro air quality dispersion modelling system in Sheffield. The aim was to determine the most significant emission sources and understand their spatial variability. NOx emissions (ton/year) from road traffic, point and area sources for the year 2017 were 5370, 6774 and 2425, whereas that of PM_{10} (ton/year) were 345, 1449 and 281, respectively. Airviro dispersion model was run using emission inventory and meteorological data for Sheffield. The model was run using different emission scenarios. The model was also run for different seasons of the year. Estimating annual concentrations, the model results showed three hotspots of NOx, namely the Sheffield City Centre, Darnall and Tinsley Roundabout (M1 J34S). High levels of NOx are also predicted on Sheffield Parkway (A630, A57) and between Meadowhall Shopping Centre and Sheffield Forgemasters International (a heavy engineering steel company). Sheffield City Centre probably experiences the highest levels of NOx, which is mainly due to high level of road traffic but various point and area sources also contribute. Pollution levels are highest in the busiest part of the city including St. Mary’s Gate, More Street, Eyre Street, Arundel Gate, Sheaf Street, Pond Street, Exchange Place and Castlegate. The train station, the bus station, the area of the Sheffield Hallam University and a busy shopping centre make this area very busy in terms of road traffics, exposing visitors, workers, students, commuters and residents to high levels of air pollution. Areas adjacent to the city centre also experience considerable amount of air pollution. Generally, pollution levels gradually decrease with distance from the city centre. However, due to prevailing southwesterly winds and the locations of some industrial sources, there is a north-eastern trend in air pollution levels i.e. north-eastern region towards M1 experiences considerably high amount of air pollution (39 – 52 $\mu g/m^3$). Model results showed that HGVs emitted a considerable amount of emissions in the city centre and their contribution was greater than that of the buses or cars. Generally atmospheric pollutant levels were higher in colder seasons probably due to (i) greater combustion of fossil fuels, mainly diesel, petrol, gas and coal to a lesser extent, and (ii) atmospheric stagnation and shallower atmospheric boundary-layer height, which discourage pollutant dispersion as compared to hotter seasons when the atmosphere is more turbulent and boundary layer height is wider. Annual average PM_{10} levels were also predicted. The areas with elevated PM_{10}

concentrations are shown outside Sheffield City Centre, in contrast to NOx. High PM_{10} concentrations were shown mainly between Sheffield Forgemasters International (a heavy engineering steel company) and Meadowhall shopping centre. This hotspot seems to be due to the point sources (heavy steel and other companies) in this area. However, this is also a busy area in terms of road traffic, which must be contributing a significant amount of emissions. The second hotspot of PM_{10} is shown in Attercliffe, which has six point sources emitting a significant amount of PM_{10} and other pollutants. The third hotspot of PM_{10} is shown in Sheffield Parkway having three point sources. The fourth hotspot is the north-eastern corner of the city centre in the Wicker and West Bar near Derek Dooley Way. When only point sources were considered as inputs to the model, the minimum, mean and maximum PM_{10} levels (µg/m3) were 0.24, 1.3 and 13.9, respectively, whereas when only road traffic was considered these levels were 0.04, 0.56 and 2.5, respectively. PM_{10} concentrations (µg/m^3) in various seasons of the year were also modelled and compared. PM_{10} concentrations slightly varied and showed slightly different spatial pattern in various seasons. The minimum, mean, and maximum PM10 concentrations (µg/m^3) in winter, spring, summer, and autumn were: 1.0, 4.12, 15.4; 0.66, 3.75, 15.4; 0.47, 3.47, 15.5; and 0.75, 4.49, 19.0, respectively. Autumn showed relatively higher average concentrations. The results of different emission scenarios showed that NOx concentrations were mainly from road traffic, whereas PM_{10} concentrations were from point sources. Spatiotemporal variability and public exposure to air pollution were analysed. NOx concentration was greater than 50 µg/m^3 in about 8 km^2 area, where more than 66 thousand people lived. Models validated by observations can be used to fill-in spatiotemporal gaps in measured data. The approach used presents spatiotemporal situation awareness maps that could be used for decision making and improving the urban infrastructure. Furthermore, dispersion models are important tools for urban air quality management, however, steps should be taken to minimise potential errors in emission data, meteorological data and complexity of the terrain. Particulates generated from vehicle wear and tear, resuspension of dust particles and emission from natural sources require special attention to improve model performance. In addition, further work is required to quantify people exposure to air pollution in Sheffield using dense network of static sensors and personal monitors. People can be exposed to air pollution in their houses (residents), work places and when commuting to-and-from work using various means of transports, e.g., buses, trains, trams, cars, cycles or walking. How exposure levels vary using various transport modes needs to be quantified in Sheffield.

It has been a common practice to use the linear regression for developing LUR models, however, the association between air pollutant levels and spatial features is not always linear, therefore, ideally nonlinear modelling approaches should be used for developing LUR models, which can help in understanding small scale spatial variability of different air pollutants in urban areas. In chapter 6, the GAM was fitted and predicted in R programming language and the predicted concentrations was then transferred to ArcGIS for producing maps of NO_2 in Sheffield. Alternatively, R has several special packages that can be used for mapping and spatial analysis of the predicted concentrations. In this paper spatial variability of NO_2 concentration is modelled in the city of Sheffield for year 2019. MLRM is a traditional and most commonly used approach for developing LUR models, here in addition to the linear approach, a nonlinear GAM model is employed and its benefits and the way it is applied are discussed. Three datasets of NO_2 measurements were used: (a) NO_2 data from 188 DT, (b) NO_2 from 40 LCS, and (c) NO_2 data from both DT and LCS. The first group performed better than the other two groups. Among predictor variables altitude (negative effect), major roads (positive effect), minor roads (positive effect), and commercial area (positive effect) had significant effect in all three groups. The model successfully captured the spatial variability of NO_2 in Sheffield, estimating high levels of NO_2 in the city centre and on major roads around the city. The eastern area between the city centre and motorway (M1) showed particularly high levels, whereas the western area (Peak District National Park) demonstrated lower levels of NO_2 concentrations. The main contributions of this work are summarised as follows: (a) An advanced nonlinear GAM is proposed for developing an LUR model, which outperforms the linear counterpart. (b) High resolution maps (100 m x 100 m) of NO_2 are developed in Sheffield using a nonlinear LUR model for quantifying public exposure to NO_2 and determining how the exposure varies at small scales in the city. (c) NO_2 data measured by a network of DT and LCS are integrated to developed high-resolution maps. (d) It is confirmed that Sheffield City Centre and its eastern sides experience relatively higher levels of NO_2 pollution. Future work could include developing a spatiotemporal LUR model using meteorology, traffic counts and fleet composition data. In this study, the effect of major and minor roads is analysed on NO_2 levels, however, road traffic and composition may vary from time to time on a given road. It is, therefore, important to capture temporal and spatial variability in meteorology and traffic data and feed it to the nonlinear LUR models. This will probably further improve the model performance, depending on the quality and temporal resolution of the data.

Airviro model was applied in Chapter 5 and LUR model in Chapter 6 for estimating the NO_2 concentrations in Sheffield. Here the performance of these models are compared for predicting NO_2 concentrations. In Chapter 6 several models were developed, here we used the performance of the GAM, which used data from NO_2 diffusion tubes for both modelling training and cross-validation. Correlation coefficient and RMSE values were 0.70 and 8.69 for GAM LUR model, and 0.48 and 44.01 for Airviro model, respectively. This shows that GAM LUR model performed better than the Airviro model. Comparison of the two models was not possible for PM_{10}, as no LUR model was developed for PM_{10} due to the lack of measured PM_{10} data to train the LUR model. Therefore, LUR is recommended for the future modelling of NO_2 in urban areas.

The aim of the chapter 7 is to analyse small-scale spatial variability in NO_2 concentrations with the help of pollution maps. Maps of NO_2 are produced using geostatistical interpolation, Airviro dispersion model and LUR model. Finally, NO_2 concentrations estimated by Airviro and LUR models are fused with measured NO_2 concentrations. Measured NO_2 data were used from 49 continuous AQMS including 3 Automatic Urban and Rural Network (AURN) sites run by DEFRA, 5 reference sites run by Sheffield City Council and 41 LCS installed by the Urban Flows Observatory, the University of Sheffield. In addition, NO_2 data were used from 188 diffusion tubes provided by Sheffield City Council. Air quality standards were exceeded at several locations in Sheffield, particularly in the city centre and on some busy roads, where annual mean NO_2 levels were higher than 40 µg/m^3. The highest levels of NO_2 were recorded by the Envirowatch E-MOTE installed next to the Sheffield train station taxi rank (136.81 µg/m^3), followed by the Sheaf Street/Sheaf Square pedestrian crossing (115.56 µg/m^3) and Arundel Gate (107.17 µg/m^3). Three modelling approaches were used to produce maps of NO_2 concentrations in Sheffield: (a) geostatistical kriging interpolation, (b) Airviro dispersion modelling, and (c) LUR based on the generalised additive model. Measured NO_2 concentrations were fused with the estimated concentrations using universal kriging, which is an advanced kriging technique that is employed for data having more than one correlated variable. Six sets of measured and estimated NO_2 data were fused: (i) Fusion of NO_2-Airviro with NO_2-LCS; (ii) Fusion of NO_2-Airviro with NO_2-DT; (iii) Fusion of NO_2-Airviro with NO_2-DTLCS; (iv) Fusion of NO_2-LUR with NO_2-LCS; (v) Fusion of NO_2-LUR with NO_2-DT; and (vi) Fusion of NO_2-LUR with NO_2-DTLCS. Fused NO_2 were compared with measured concentrations in terms of different statistical metrics including FAC2, r, RMSE, MAE and MBE. Maps produced by the fusion of NO_2-LCS and NO2-LUR produced better results, with an r-value of 0.96, followed by the fusion of NO_2-Airviro with NO_2-LCS, with an r-value of 0.88. Fused maps developed by universal kriging provided better spatial coverage and similarity with measured concentrations than the ordinary kriging interpolation, Airviro and LUR models. Fused maps combining measured and estimated concentrations produced more realistic concentrations and provided better spatial coverage. Data fusion added value to both measured and estimated concentrations: the measured data were improved by predicting spatiotemporal gaps, whereas the modelled data were improved by constraining it with observed data. Hot spots of NO_2 were shown in the city centre, eastern parts of the city towards the motorway (M1) and on some major roads. Road traffic is probably the dominant emission source of NO_2 in Sheffield. The main findings of this study are: (a) The universal kriging approach was capable of estimating realistic NO_2 concentration maps from the fusion of measured and modelled concentrations. The fused NO_2 concentrations inherited spatial patterns of the pollutant from the model estimations and adjusted the modelled values using the measured concentrations. (b) A huge number of sensors are required to provide reasonable spatial coverage at a city level, which is too expensive. Spatial modelling (e.g., dispersion model and LUR model) and data fusion approaches can provide city-level maps by integrating pollutant concentrations measured by the AQMS with modelled concentrations. (c) According to Schneider et al. [9], the accuracy of the data fusion depends on the number of sensors, their spatial distribution, their uncertainty and the accuracy of the estimated values to be fused with measured values. Here, we showed that in addition, the accuracy of the data fusion also depends on the uniformity of the sensors, meaning that using the same type of sensors can result in better accuracy. Increasing the number of sensors will not necessarily improve the model outputs, especially if sensors are not of the same type. For example, maps produced by the fusion of NO_2-LCS with NO_2-LUR (r-value 0.96) and NO_2-Airviro (r-value 0.88) produced better results than when both NO_2-DT and NO_2-LCS were used together and fused with NO_2-LUR or NO_2-Airviro, when the r-values decreased to 0.59 and 0.56, respectively. The uniqueness of this study is that it uses estimated NO_2 concentrations from two sources - a dispersion model and a land use regression model and measured NO_2 concentrations from three sources: diffusion tubes, reference sensors and low-cost sensors.

Chapter 8 analysed the temporal variability of NO_2 concentrations measured by 28 Envirowatch E-MOTEs, 13 AQMesh pods, and eight reference sensors (five run by Sheffield City Council and three run by the DEFRA). Density plots and time variation plots were used to compare the distributions and temporal variability of NO_2 concentrations. Density plots are a useful tool for comparing and analysing the distributions of NO2 concentrations measured by various sensors. Time variation plots were employed to characterise and compare the temporal variability of NO2 concentrations measured at different monitoring sites. Time variation plots visualise how NO2 concentrations vary during different time periods (e.g., diurnal, weekly and annual cycles) and help us understand their emission sources. Long-term trends, both adjusted and non-adjusted, showed significant reductions in NO_2 concentrations. At the Tinsley site, the non-adjusted trend was −0.94 (−1.12, −0.78) μgm^{-3}/year, whereas the adjusted trend was −0.95 (−1.04, −0.86) μgm^{-3}/year. At Devonshire Green, the non-adjusted trend was −1.21 (−1.91, −0.41) μgm^{-3}/year and the adjusted trend was −1.26 (−1.57, −0.83)

μgm^{-3}/year. Furthermore, NO_2 concentrations were analysed employing univariate linear and nonlinear time series models and their performance was compared with a more advanced time series model using two exogenous variables (NO and O_3). For this purpose, time series data of NO, O_3 and NO_2 were obtained from a reference site in Sheffield, which are more accurate than the measurements from LCS and, therefore, more suitable for training and testing the model. In this article, the three main steps used for model development are discussed: (i) model specification for choosing appropriate values for p, d and q, (ii) model fitting (parameters estimation), and (iii) model diagnostic (testing the goodness of fit). The linear auto-regressive integrated moving average (ARIMA) performed better than the nonlinear counterpart; however, its performance in predicting NO_2 concentration was inferior to ARIMA with exogenous variables (ARIMAX). Using cross-validation ARIMAX demonstrated strong association with the measured concentrations, with a correlation coefficient of 0.84 and RMSE of 9.90. ARIMAX can be used as an early warning tool for predicting potential pollution episodes in order to be proactive in adopting precautionary measures.

The next section describes as to how the main objectives of the project are addressed.

9.4. How are the main objectives of the project addressed?

The core aims of this PhD project were to analyse spatiotemporal variability of AQ and determine the main drivers of air pollution in Sheffield by deploying a dense AQMN based on multiple criteria, and employing various modelling and mapping techniques for AQ prediction and developing high resolution maps. The main objectives of the project and how they are addressed are given below:

Objective I: To define criteria for structuring a multipurpose AQMN and to structure a dense network made of several layers of AQ sensors including reference, LCS and IoT sensors in Sheffield.

Objective 1 is addressed in Chapter 4, where a methodology was proposed supported by numerical, conceptual and GIS frameworks for structuring AQMN using social, environmental and economic indicators as a case study in Sheffield. The main factors used for the selection of an AQMS were population-weighted pollution concentration (PWPC) and weighted spatial variability (WSV) incorporating population density (social indicator), pollution levels and spatial variability of air pollutant concentrations (environmental indicator). Total number of sensors was decided on the basis of budget (economic indicator), whereas the number of sensors deployed in each output area was proportional to WSV. The purpose of AQ monitoring and its role in determining the location of an AQMS was analysed. Furthermore, the existing AQMN in Sheffield was analysed and an alternative proposed following a formal procedure. In contrast to traditional networks, which are generally structured based on a single AQ monitoring approach, the proposed AQMN has several layers of sensors: Reference sensors recommended by DEFRA and LCS mainly AQMesh and Envirowatch E-MOTEs. The core aim was to structure an integrated AQMN in urban areas, which will lead to the collection of AQ data of high spatiotemporal resolution. The use of LCS in the proposed network provides a cheaper option for setting up a purpose-designed network for greater spatial coverage, especially in low- and middle-income countries.

Objective II: To deploy LCS for rendering high quality measurements of NO_2 concentrations. This objective assesses and improves the quality of LCS data using calibration models.

Objective II has been addressed in Chapter 3, which compared the measurements of Envirowatch E-MOTEs with the measurements of reference sensors. Ten E-MOTEs were deployed for a year (October 2016 to September 2017) in Sheffield, which measured several air pollutants (NO, NO_2, CO) and meteorological parameters. The measurements of E-MOTEs were compared to each other and to a reference instrument installed nearby. E-MOTEs were able to successfully capture the temporal variability such as diurnal, weekly and annual cycles in air pollutant concentrations and demonstrated significant similarity with reference instruments. NO_2 concentrations showed very strong positive correlation between various LCS sensors. Mostly correlation coefficients (r - values) were greater than 0.92. CO from different sensors also had r - values mostly greater than 0.92, however, NO showed r - value less than 0.5. Furthermore, several MLRM and GAM were developed to validate the E-MOTE data and reproduce NO and NO_2 concentrations measured by the reference instruments. GAMs demonstrated significantly better performance than the linear models by capturing the nonlinear association between the response and explanatory variables. The best GAM developed for reproducing NO_2 concentrations returned values of 0.95, 3.91, 0.81, 0.005, and 0.61 for FAC2, RMSE, R^2, NMB and COE, respectively.

Objective III: To model the spatial variability of AQ in urban areas using various modelling approaches including dispersion and land-use regression models.

Objective III of this project has been addressed in Chapter 5 and 6. In Chapter 5 NOx and PM_{10} concentrations were modelled employing Airviro air quality dispersion modelling system. The aim was to determine the most significant emission sources and analyse their spatial variability. NOx emissions (ton/year) from road traffic, point and area sources for the year 2017 were 5370, 6774 and 2425, whereas that of PM_{10} (ton/year) were 345, 1449 and 281, respectively, which are part of the emission database. The results showed three hotspots of NOx, namely the Sheffield City Centre, Darnall and Tinsley Roundabout (M1 J34S). High PM_{10} concentrations were shown mainly between Sheffield Forgemasters International (a heavy engineering steel company) and Meadowhall shopping centre. Several emission scenarios were tested which showed that NOx concentrations were mainly from road traffic, whereas PM_{10} concentrations were from point sources. Spatiotemporal variability and public exposure to air pollution were analysed. In Chapter 6 a nonlinear generalised additive model was proposed for LUR and its performance was compared to a linear model in Sheffield for the year 2019. Pollution models were estimated using NO_2 measurements obtained from 188 diffusion tubes and 40 LCS. Performance of the models was assessed by calculating several statistical metrics including correlation coefficient and RMSE. High-resolution (100m x 100m) maps demonstrated higher levels of NO_2 in the city centre, eastern side of the city and on major roads. Comparison of Airviro and LUR models showed that LUR model performed better for predicting NO_2 concentrations, and therefore is a better choice for model NO_2 concentrations in urban areas.

Objective IV: To explore the use of novel data fusion techniques to improve both model estimations and LCS measurements and further improve the quality of AQ maps.

Objective IV is addressed in Chapter 7, wherein NO_2 concentrations measured by diffusion tubes, reference sensors and LCS were fused with the estimations of Airviro and LUR model. Data fusion with universal kriging produced more realistic maps in terms of spatial coverage and comparability with measured data than the interpolated maps using ordinary kriging. The validity of the fused maps was analysed in terms of statistical metrics including correlation coefficient, RMSE, MAE, FAC2 and MBE. Maps produced by the fusion of NO_2-LCS with NO_2-LUR (r-value 0.96) and NO_2-Airviro (r-value 0.88) produced better results.

Objective V: To analyse temporal variability of AQ using both graphical and time series analysis.

In chapter 8 measurements from LCS are analysed employing density plots and time series modelling and visualisation. Density plots are a useful tool for comparing and analysing the distributions of NO_2 concentrations measured by various sensors. Time variation plots were employed to characterise and compare the temporal variability of NO_2 concentrations measured at different monitoring sites. Time variation plots visualise how NO_2 concentrations vary during different time periods (e.g., diurnal, weekly and annual cycles) and help us understand their emission sources. Long-term data for NO_2 concentrations showed negative trends at both the Sheffield Tinsley and Devonshire Green sites, indicating that pollutant emissions decreased because of stringent emission policies. However, the reductions in pollution varied both spatially and temporally. Moreover, further smart interventions are required to cut emissions and improve air quality to comply with air quality guidelines. Time series model with external variables demonstrated better performance compared to linear and nonlinear time series models without external variables.

.

9.5. Future work

In this PhD project only outdoor air pollution emissions were considered mainly road traffic and points sources, emission from indoor sources including residential houses were not considered. This needs further consideration for detailed modelling, especially for dispersion modelling. Quantification of the emissions from residential houses, especially from cooking and stove burners can further improve the performance of dispersion model and help quantify exposure to indoor pollution.

How indoor and outdoor AQ interacts is not fully characterised, therefore it is vital to characterise the interface between indoor and outdoor air pollution. How outdoor AQ affects indoor and vice versa, and to understand the flow of air between the indoor and outdoor environments is vital for characterising exposure to air pollution. Understanding how to retain heat whilst allowing adequate ventilation to maintain clean air is also important from health and energy saving point of view. Timing and rates of ventilation are important for improving comfortability of residential house and working environments.

People are exposed to air pollution in both indoor and outdoor environment, however, how the exposure vary from one house to another, from one office to another and what are the factors that affect air pollution levels and exposure need to be quantified. People are exposed to air pollution while commuting to-and-from work and the exposure varies in both time and space. Exposure also depends on the type of vehicle used for commuting and the route followed. All these factors need to be characterised in the future work.

In Sheffield a dense network of AQ sensors was deployed monitoring different air pollutants including both gaseous (e.g., NO_2, SO_2, O_3 and CO) and particle pollutants (e.g., PM_{10} and

$PM_{2.5}$). However, the chemical composition of particulate is not determined to quantify different types of chemical particles in particulate matter. This is, therefore, important to carry out chemical analysis of particle samples and characterise their chemical composition. How the chemical composition of particles vary both spatially and temporally need to be characterised.

9.6. Conclusion

In this chapter the PhD project is summarised highlighting its main objectives, methodology, main findings and recommendations for future work.

The main findings of the project are given below:

- The project proposes a nonlinear generalised additive model for low-cost sensors calibrations in outdoor environment. Low-cost sensors are a cheaper source of air quality data; however, they require robust outfield calibrations.
- A formal approach is proposed for structuring an air quality monitoring network in urban areas, which is be based on multiple criteria.
- Several modelling options are available for modelling the spatial variability of air pollutants in urban areas (e.g., Airviro dispersion model and LUR models). It is shown that LUR model based on nonlinear machine learning approach outperforms the dispersion modelling approach.
- Data fusion techniques (such as Universal krigging) are employed to integrate model estimations with measured concentrations. Such data fusion approaches are useful tools for improving data quality and producing high-resolution air quality maps.
- Time series modelling ARIMA with exogenous variables (ARIMAX) outperformed other linear and nonlinear time series models, and is proposed as an early warning tool for predicting potential pollution episodes in order to be proactive in adopting precautionary measures.

Potential future work may include: (a) characterisation of indoor air pollution and its interaction with outdoor air pollution, (b) quantification of human exposure to air pollution in different residential, working and commuting microenvironments, and (c) analysis of the chemical composition of particulate matter.

APPENDICES

Appendix 1: R Code

1. R code for Chapter 3 – Low-cost sensor calibrations

```
Require (openair)
temp1<-read.csv(file.choose(), header=T)##701 temp1<-import()#i used import
temp1<-import()#701
temp2<-import()#702

temp1$date <- as.POSIXct(strptime(temp1$date,format = "%d/%m/%Y %H:%M", tz =
"GMT"))
temp2$date <- as.POSIXct(strptime(temp2$date,format = "%d/%m/%Y %H:%M", tz =
"GMT"))
temp3$date <- as.POSIXct(strptime(temp3$date,format = "%d/%m/%Y %H:%M", tz =
"GMT"))
temp4$date <- as.POSIXct(strptime(temp4$date,format = "%d/%m/%Y %H:%M", tz =
"GMT"))
temp5$date <- as.POSIXct(strptime(temp5$date,format = "%d/%m/%Y %H:%M", tz =
"GMT"))
temp6$date <- as.POSIXct(strptime(temp6$date,format = "%d/%m/%Y %H:%M", tz =
"GMT"))
temp7$date <- as.POSIXct(strptime(temp7$date,format = "%d/%m/%Y %H:%M", tz =
"GMT"))
temp8$date <- as.POSIXct(strptime(temp8$date,format = "%d/%m/%Y %H:%M", tz =
"GMT"))
temp9$date <- as.POSIXct(strptime(temp9$date,format = "%d/%m/%Y %H:%M", tz =
"GMT"))
temp10$date <- as.POSIXct(strptime(temp10$date,format = "%d/%m/%Y %H:%M", tz =
"GMT"))

names(temp1) <- c("date", "CO_1", "NO_1", "NO2_1", "RH_1", "Temp_1")
names(temp2) <- c("date", "CO_2", "NO_2", "NO2_2", "RH_2", "Temp_2")
names(temp3) <- c("date", "CO_3", "NO_3", "NO2_3", "RH_3", "Temp_3")
names(temp4) <- c("date", "CO_4", "NO_4", "NO2_4", "RH_4", "Temp_4")
names(temp5) <- c("date", "CO_5", "NO_5", "NO2_5", "RH_5", "Temp_5")
names(temp6) <- c("date", "CO_6", "NO_6", "NO2_6", "RH_6", "Temp_6")
names(temp7) <- c("date", "CO_7", "NO_7", "NO2_7", "RH_7", "Temp_7")
names(temp8) <- c("date", "CO_8", "NO_8", "NO2_8", "RH_8", "Temp_8")
names(temp9) <- c("date", "CO_9", "NO_9", "NO2_9", "RH_9", "Temp_9")
names(temp10) <- c("date", "CO_10", "NO_10", "NO2_10", "RH_10", "Temp_10")

temp1_hr<-timeAverage(temp1, avg.time="hour")
temp2_hr<-timeAverage(temp2, avg.time="hour")
temp3_hr<-timeAverage(temp3, avg.time="hour")
```

```
temp4_hr<-timeAverage(temp4, avg.time="hour")
temp5_hr<-timeAverage(temp5, avg.time="hour")
temp6_hr<-timeAverage(temp6, avg.time="hour")
temp7_hr<-timeAverage(temp7, avg.time="hour")
temp8_hr<-timeAverage(temp8, avg.time="hour")
temp9_hr<-timeAverage(temp9, avg.time="hour")
temp10_hr<-timeAverage(temp10, avg.time="hour")

summary(temp1_hr)
temp1_hr<-subset(temp1_hr, NO2_1>0&NO2_1<100)
temp1_hr<-subset(temp1_hr, NO_1>0&NO_1<100)
summary(temp2_hr)
temp2_hr<-subset(temp2_hr, NO_2>0&NO_2<100)

summary(temp3_hr)
temp3_hr<-subset(temp3_hr, NO_3>0&NO_3<100)
summary(temp4_hr)
temp4_hr<-subset(temp4_hr, NO_4>0&NO_4<100)
summary(temp5_hr)
temp5_hr<-subset(temp5_hr, NO_5>0&NO_5<100)
summary(temp6_hr)
temp6_hr<-subset(temp6_hr, NO_6>0&NO_6<100)
summary(temp7_hr)
temp7_hr<-subset(temp7_hr, NO_7>0&NO_7<100)
summary(temp8_hr)
temp8_hr<-subset(temp8_hr, NO_8>0&NO_8<100)
summary(temp9_hr)
temp9_hr<-subset(temp9_hr, NO_9>0&NO_9<100)
summary(temp10_hr)
temp10_hr<-subset(temp10_hr, NO_10>0&NO_10<100)

install.packages ("plyr")
require(plyr)

merged<-Reduce(function(x, y) merge(x, y, all=TRUE), list(temp1_hr, temp2_hr, temp3_hr,
temp4_hr, temp5_hr,
temp6_hr, temp7_hr,temp8_hr, temp9_hr, temp10_hr))##
write.csv(merged, "merged_sep.csv", row.names=F)
timePlot(merged, pollutant=c("NO2_1","NO2_2","NO2_3", "NO2_4","NO2_5", "NO2_6",
"NO2_7",
"NO2_8", "NO2_9","NO2_10"), group = F, key.columns = 5, ylab="")
timePlot(merged, pollutant=c("NO_1","NO_2","NO_3", "NO_4","NO_5", "NO_6", "NO_7",
"NO_8", "NO_9","NO_10"), group = F, key.columns = 5, ylab="")
timePlot(merged, pollutant=c("CO_1","CO_2","CO_3", "CO_4","CO_5", "CO_6", "CO_7",
"CO_8", "CO_9","CO_10"), group = F, key.columns = 5, ylab="")
summaryPlot(temp1_hr)
scatterPlot(merged, x="date", y="NO2_1")
merged$month <- as.Date(merged$date,format = "%m", tz = "GMT")
head(merged$month)
names(merged)
```

```
plot(merged$date, merged$NO2_1, xlab="date", ylab="NO2 concentraiton (ppb)",
col="black", ylim=c(0,100))
points(merged$date, merged$NO2_2, col="red")
merged$month <- format(as.Date(merged$date),%m")
plot(merged$month, merged$no2_1, xlab="date", ylab="NO2 concentraiton (ppb)",
col="black")
timePlot(temp1_hr, c("NO_1", "NO2_1", "CO_1", "Temp_1", "RH_1"))
timePlot(temp2_hr, c("NO_2", "NO2_2", "CO_2", "Temp_2", "RH_2"))
timePlot(temp3_hr, c("NO_3", "NO2_3", "CO_3", "Temp_3", "RH_3"))
timePlot(temp4_hr, c("NO", "NO2_4", "CO_4", "Temp_4", "RH_4"))
timePlot(temp5_hr, c("NO_5", "NO2_5", "CO_5", "Temp_5", "RH_5"))
timePlot(temp6_hr, c("NO_6", "NO2_6", "CO_6", "Temp_6", "RH_6"))
timePlot(temp7_hr, c("NO_7", "NO2_7", "CO_7", "Temp_7", "RH_7"))
timePlot(temp8_hr, c("NO_8", "NO2_8", "CO_8", "Temp_8", "RH_8"))
timePlot(temp9_hr, c("NO_9", "NO2_9", "CO_9", "Temp_9", "RH_9"))
timePlot(temp10_hr, c("NO_10", "NO2_10", "CO_10", "Temp_10", "RH_10"))
head(merged)
##################################################
#correlation
#NO2
no2<-data.frame(merged$date,
merged$NO2_1,merged$NO2_2,merged$NO2_3,merged$NO2_4,
merged$NO2_5, merged$NO2_6, merged$NO2_7, merged$NO2_8, merged$NO2_9,
merged$NO2_10)
corPlot(no2)
names(no2) <- c("date", "NO2_1", "NO2_2","NO2_3", "NO2_4","NO2_5", "NO2_6",
"NO2_7",
"NO2_8", "NO2_9","NO2_10")
corPlot(no2, col="hue")
#type of color for correlation plots->“increment”, “heat”, “spectral”, “hue”, “greyscale”

#NO
no<-data.frame(merged$date, merged$NO_1,merged$NO_2,merged$NO_3,merged$NO_4,
merged$NO_5, merged$NO_6, merged$NO_7, merged$NO_8, merged$NO_9,
merged$NO_10)
names(no) <- c("date", "NO_1", "NO_2","NO_3", "NO_4","NO_5", "NO_6", "NO_7",
"NO_8", "NO_9","NO_10")
corPlot(no, col="hue")

#CO
co<-data.frame(merged$date, merged$CO_1,merged$CO_2,merged$CO_3,merged$CO_4,
merged$CO_5, merged$CO_6, merged$CO_7, merged$CO_8, merged$CO_9,
merged$CO_10)
names(co) <- c("date", "CO_1", "CO_2","CO_3", "CO_4","CO_5", "CO_6", "CO_7",
"CO_8", "CO_9","CO_10")
corPlot(co, col="hue")
###############################################################
timeVariation2(merged, c("NO2_2","NO2_3" ,"NO2_5", "NO2_6", "NO2_7",
"NO2_8", "NO2_9","NO2_10"), ylab="", ci=FALSE, key.columns = 5)
timeVariation(merged, c("NO_1", "NO_2","NO_3", "NO_4","NO_5", "NO_6", "NO_7",
"NO_8", "NO_9","NO_10"), ylab="", ci=F, key.columns = 5)
timeVariation(merged, c("CO_1","CO_3", "CO_4","CO_5", "CO_7",
"CO_8", "CO_9","CO_10"), ylab="", ci=F, key.columns = 5)
#############################################################################
## to find average of several columns
 no2$NO2_mean<-rowMeans(subset(no2,select=c(NO2_2,NO2_3,NO2_5, NO2_6, NO2_7,
NO2_8, NO2_9,NO2_10), na.rm = TRUE))
head(no2)
timeVariation(no2, "NO2_mean")
 no$NO_mean<-rowMeans(subset(no,select=c(NO_2,NO_3, NO_4,NO_5, NO_6, NO_7,
NO_8, NO_9,NO_10), na.rm = TRUE))
head(no)
timeVariation(no, "NO_mean", ci=T)
co$CO_mean<-rowMeans(subset(co,select=c(CO_1, CO_2,CO_3, CO_4,CO_5, CO_6,
CO_7,
CO_8, CO_9,CO_10), na.rm = TRUE))
head(co)
timeVariation(co, "CO_mean")
##############################################################################
mrg<-merge(temp2,dev, by="date", all=TRUE)
head(mrg)
## Devonshire Green Data
dev<-importAURN(site = "shdg", year =2016:2017, pollutant = "all")
dev2<-dev[,1:3]## to select only date, NO and NO2
dev3<-dev[,c(1:3, 12, 13)]## to select date, NO, NO2, ws, and wd
names(dev2)<-c("date", "NO", "NO2")
names(dev3)<-c("date", "NO", "NO2", "ws","wd")
dev2$NO2<-(dev2$NO2/1.88)
dev2$NO<-(dev2$NO/1.25)
dev3$NO2<-(dev3$NO2/1.88)
dev3$NO<-(dev3$NO/1.25)
dev4<-selectByDate(dev3, start = "2016/10/01", end = "2017/09/30")
timePlot(dev4, c("NO_DG", "NO2_DG"), group=T)
dev5<-subset(dev4, NO<200)
timePlot(dev4, c("NO_DG", "NO2_DG"), group=T, ylab="NO and NO2 concentrations
(ppb)", main="Devonshire Green")
timeVariation(dev4, c("NO", "NO2"), ylab="", main="Devonshire Green")
names(dev4)<-c("date", "NO_DG", "NO2_DG", "ws", "wd")##rename to be easily identified
in big dataset
dev_sensors<-merge(merged, dev4, by="date", all=TRUE)
head(dev_sensors)
nrow(dev_sensors)
dev_sensors$NO_mean<-rowMeans(subset(dev_sensors,select=c(NO_3, NO_4,NO_5,
NO_6, NO_7,
NO_8, NO_9,NO_10), na.rm = TRUE))
dev_sensors$CO_mean<-rowMeans(subset(dev_sensors,select=c(CO_1, CO_2,CO_3,
CO_4,CO_5, CO_6, CO_7,
CO_8, CO_9,CO_10), na.rm = TRUE))
dev_sensors$NO2_mean<-rowMeans(subset(dev_sensors,select=c(NO2_2,NO2_3,NO2_5,
NO2_6, NO2_7,
NO2_8, NO2_9,NO2_10), na.rm = TRUE))
names(dev_sensors)
timePlot(dev_sensors, pollutant=c("NO_mean","NO2_mean", "NO_DG", "NO2_DG"),
group=T, ylab="Concentrations (ppb)")
timePlot(dev_sensors, pollutant=c("NO_mean","NO2_mean", "NO_DG", "NO2_DG"),
group=F, ylab="Concentrations (ppb)")
timeVariation2(dev_sensors, pollutant=c("NO_mean","NO2_mean", "NO_DG",
"NO2_DG"), ylab="Concentrations (ppb)", ci=T)
timeVariation(dev_sensors, pollutant=c("NO_DG", "NO2_DG"), ylab="Concentrations
(ppb)", ci=T)
h<-dev_sensors
h<-subset(h, NO_DG>0&NO_DG<100)##removing some higher and lower values
timeVariation(h, pollutant=c("NO_mean","NO2_mean", "NO_DG", "NO2_DG"),
ylab="Concentrations (ppb)", ci=T)
no<-subset(h, select=c(NO_1, NO_2,NO_3, NO_4, NO_5, NO_6, NO_7, NO_8, NO_9,
NO_10))
no2<-subset(h, select=c(NO2_1, NO2_2,NO2_3, NO2_4, NO2_5, NO2_6, NO2_7, NO2_8,
NO2_9, NO2_10))
corPlot(no)
corPlot(no2)
################################################################
NO2<-dev_sensors[,c(4,9,14,19,24,29,34,39,44, 49, 53, 58)]
NO<-dev_sensors[, c(3,8,13,18,23,28,33,38,43,48,52,56)]
names(NO)<-c("NO_1", "NO_2", "NO_3", "NO_4", "NO_5","NO_6", "NO_7", "NO_8",
"NO_9", "NO_10",
"NO_DG", "NO_mean")
names(NO2)<-c("NO2_1", "NO2_2", "NO2_3", "NO2_4", "NO2_5","NO2_6", "NO2_7",
"NO2_8", "NO2_9", "NO2_10",
"NO2_DG", "NO2_mean")
names(dev_sensors)
corPlot(NO2, col="hue")
corPlot(NO, col="hue")
test<-dev_sensors[, c(52, 53, 56, 58)]
corPlot(test)
dev_sensors2<-dev_sensors[,c(-4,-8)]##to remove NO_2 and NO2_1 which have some bad
data
dev_sensors2$NO_mean<-rowMeans(subset(dev_sensors,select=c(no_1,no_3, no_4,no_5,
no_6, no_7,
no_8, no_9,no_10), na.rm = TRUE))
dev_sensors2$NO2_mean<-rowMeans(subset(dev_sensors,select=c(no2_2,no2_3,
no2_4,no2_5, no2_6, no2_7,
no2_8, no2_9,no2_10), na.rm = TRUE))
NO2<-dev_sensors[,c(9,14,19,24,29,34,39,44, 49, 53, 58)]
NO<-dev_sensors[, c(3,13,18,23,28,33,38,43,48,52,56)]
NO2$NO2_mean<-rowMeans(subset(NO2,select=c(no2_2,no2_3, no2_4,no2_5, no2_6,
no2_7,
no2_8, no2_9,no2_10), na.rm = TRUE))
 NO$NO_mean<-rowMeans(subset(NO,select=c(no_1, no_3, no_4,no_5, no_6, no_7,
no_8, no_9,no_10), na.rm = TRUE))
```

```
timePlot(dev_sensors, c("no_1", "no_3", "no_4","no_5", "no_6", "no_7", "no_8",
"no_9","no_10"))
corPlot(NO2)
windows()
corPlot(NO)
names(dev_sensors2)
test2<-dev_sensors2[, c(50, 51, 54, 56)]
corPlot(test2)
################################################################
#removing no_2 and no2_1 columns
timePlot(dev_sensors, pollutant=c("NO_mean","NO2_mean", "NO_DG", "NO2_DG"),
group=T, ylab="Concentrations (ppb)")
timeVariation(dev_sensors, pollutant=c("NO_mean","NO2_mean", "NO_DG", "NO2_DG"),
ylab="Concentrations (ppb)")
test<-subset(dev_sensors, select=c("NO_mean","NO2_mean", "NO_DG", "NO2_DG"))
corPlot(test)
y17<-selectByDate(dev_sensors, year=2017)##to select only 2017 as Oct and Nov in2016 are
eroneous
test2<-subset(y17, select=c("NO_mean","NO2_mean", "NO_DG", "NO2_DG"))
corPlot(test2)
train2<-na.omit(train)
cor<-cor(train2)
print(cor, digits=2)
cor2<-cor(train2, method="spearman")
print(cor2, digits=2)
###############################################################
Linear Regression Model
hello<-dev_sensors2[,50:58]
names(hello)<- c("NO_DG", "NO2_DG", "ws", "wd", "NO_mean",
"CO_mean","NO2_mean", "temp", "rh")
#select randomly 75% training dataset and 25% testing dataset
smp_size <- floor(0.75 * nrow(hello))
train_ind <- sample(seq_len(nrow(hello)), size = smp_size)
train <- hello[train_ind, ]
test <- hello[-train_ind, ]
lm_no<-lm(NO_mean~NO_DG, data=train)
p1<-predict(lm_no, newdata=test)
test$p1<-p1
modStats(test, mod="p1", obs="NO_mean")
lm_no2<-lm(NO2_mean~NO2_DG, data=train)
summary(lm_no2)
no2p<-predict(lm_no2, newdata=test)
test$no2p<-no2p
modStats(test, mod="no2p", obs="NO2_mean")
mlr<-lm(NO_mean~NO_DG+NO2_DG+NO2_mean+temp+ws+rh,data=train)
p3<-predict(mlr, newdata=test)
test$p3<-p3
modStats(test, mod="p3", obs="NO_mean")
mlr2<-lm(NO2_mean~NO_mean+NO_DG+NO2_DG+temp+ws+rh,data=train)
p4<-predict(mlr2, newdata=test)
```

```
test$p4<-p4
modStats(test, mod="p4", obs="NO2_mean")
mlr3<-lm(NO_DG~NO_mean+NO2_DG+NO2_mean+temp+ws+rh,data=train)
p5<-predict(mlr3, newdata=test)
test$p5<-p5
modStats(test, mod="p5", obs="NO_DG")
mlr4<-lm(NO2_DG~NO_mean+NO_DG+NO2_mean+temp+ws+rh,data=train)
p6<-predict(mlr4, newdata=test)
test$p6<-p6
modStats(test, mod="p6", obs="NO2_DG")
scatterPlot(test, x="NO2_DG", y="p6", mod.line=T, ylab="Predicted NO2 (ppb)",
xlab="Observed NO2_DG (ppb)", xlim=c(0, 70), main="MLRM")
##########################################################
r_gam<-c(0.41, 0.40, 0.55, 0.52, 0.84, 0.91)
fac2_gam<-c(0.98,0.80,0.98,0.84,0.53,0.95)
ioa_gam<-c(0.55,0.52,0.63,0.55,0.75,0.81)
nmb_gam<-c(0.014,0.012,0.010,0.011,0.008,0.005)
coe_gam<-c(0.101,0.048,0.633,0.097,0.50,0.614)
r_lm<-c(0.43, 0.39,0.52,0.42,0.72,0.80)
fac2_lm<-c(0.98,0.78,0.98,0.81,0.30,0.83)
ioa_lm<-c(0.55,0.52,0.62,0.52,0.56,0.705)
nmb_lm<-c(0.002,0.013,0.010,0.012,0.012,0.001)
coe_lm<-c(0.103,0.047,0.24,0.049,0.123,0.411)
gm<-data.frame(r_gam,fac2_gam, ioa_gam, nmb_gam, coe_gam)
lm<-data.frame(r_lm, fac2_lm, ioa_lm, nmb_lm, coe_lm)
plot(1:6, r_gam, col="blue", pch=2, ylim=c(0,1))
points(1:6, r_lm, col="red", pch=2)
points(1:6, fac2_gam,col="blue", pch=3)
points(1:6, fac2_lm,col="red", pch=3)
############################################################
require(mgcv)
set.seed(123)## set the seed to make your partition reproductible
gam_no<-gam(NO_mean~s(NO_DG), data=train)#....(7)
p7<-predict(gam_no, newdata=test)
test$p7<-p7
modStats(test, mod="p7", obs="NO_mean")
gam_no2<-gam(NO2_mean~s(NO2_DG), data=train)#....(8)
p8<-predict(gam_no2, newdata=test)
test$p8<-p8
modStats(test, mod="p8", obs="NO2_mean")
gam<-
gam(NO_mean~s(NO_DG)+s(NO2_DG)+s(NO2_mean)+s(temp)+s(ws)+s(rh),data=train)
p9<-predict(gam, newdata=test)
test$p9<-p9
modStats(test, mod="p9", obs="NO_mean")
gam2<-
gam(NO2_mean~s(NO_DG)+s(NO2_DG)+s(NO_mean)+s(temp)+s(ws)+s(rh),data=train)
p10<-predict(gam2, newdata=test)
test$p10<-p10
modStats(test, mod="p10", obs="NO2_mean")
```

```
gam3<-
gam(NO_DG~s(NO2_mean)+s(NO2_DG)+s(NO_mean)+s(temp)+s(ws)+s(rh),data=train)
p11<-predict(gam3, newdata=test)
test$p11<-p11
modStats(test, mod="p11", obs="NO_DG")
gam4<-
gam(NO2_DG~s(NO2_mean)+s(NO_DG)+s(NO_mean)+s(temp)+s(ws)+s(rh),data=train)
p12<-predict(gam4, newdata=test)
test$p12<-p12
modStats(test, mod="p12", obs="NO2_DG")
plot(gam4, page=1, ylab="NO2_DG conc. (ppb)")
scatterPlot(test, x="NO2_DG", y="p12", mod.line=T, ylab="Predicted NO2 (ppb)",
xlab="Observed NO2_DG (ppb)", main="GAM", xlim=c(0, 70), ylim=c(0, 60), col="blue")
summary(gam4)
####################################################################
temp<-subset(dev_sensors, select=c("temp_1", "temp_2", "temp_3", "temp_4", "temp_5",
"temp_6",
"temp_7", "temp_8", "temp_9", "temp_10"))
temp$temp_mean<-apply(temp, 1, mean)
dev_sensors$temp_mean<-temp$temp_mean
dev_sensors2$temp_mean<-temp$temp_mean
rh<-subset(dev_sensors, select=c("rh_1", "rh_2", "rh_3", "rh_4", "rh_5", "rh_6",
"rh_7", "rh_8", "rh_9", "rh_10"))
rh$rh_mean<-apply(rh, 1, mean)
dev_sensors$rh_mean<-rh$rh_mean
dev_sensors2$rh_mean<-rh$rh_mean
data_all<-merge(merged2, dev5, by="date", all=TRUE)
dev_sensors$ws<-data_all$ws
dev_sensors$pm2.5<-data_all$pm2.5
dev_sensors$o3<-data_all$o3
dev_sensors$wd<-data_all$wd
###################################
smooth = FALSE, spline = FALSE, linear = FALSE
method=“hexbin”
scatterPlot(dev_sensors, y="NO2_DG", x="NO2_mean", mod.line=T)
scatterPlot(dev_sensors, y="NO2_DG", x="NO2_mean", method="hexbin")
scatterPlot(dev_sensors, y="NO2_DG", x="NO2_mean", linear=T)
scatterPlot(dev_sensors, y="NO_DG", x="NO_mean", method="hexbin")
scatterPlot(dev_sensors, y="NO_DG", x="NO_mean", linear=T)
calendarPlot(dev_sensors, pollutant = "NO2_mean", year=2017)
calendarPlot(dev_sensors, pollutant = "NO2_mean", year=2016)
calendarPlot(dev_sensors, pollutant = "CO_mean", year=2017)
calendarPlot(dev_sensors, pollutant = "CO_mean", year=2016)
############################################################
polarPlot(dev_sensors2, "NO")
polarPlot(dev_sensors2, "NO2")
polarPlot(dev_sensors2, "NO2",type="rh_mean")
polarPlot(dev_sensors2, "NO",type="rh_mean")
polarPlot(dev_sensors2, "NO2", type="temp_mean")
polarPlot(dev_sensors2, "NO", type="temp_mean")
```

```
hello<-subset(dev_sensors2, select=c("NO", "NO2", "NO_mean", "NO2_mean"))
corPlot(hello)
############################################################################
tinsely<-importAURN(site = "she", year =2017, pollutant = c("no", "no2"))
tinsely2<-subset(tinsely,no>0&no<1000)
tinsely3<-subset(tinsely2,no2>0&no2<400)
tinsely3$no2<-(tinsely3$no2/1.88)
tinsely3$no<-(tinsely3$no/1.25)
timePlot(tinsely3, c( "no","no2"), group=T)
TheilSen(tinsely3, pollutant = "no2", ylab = "NO2 (ppb)")
TheilSen(tinsely3, pollutant = "no", ylab = "NO (ppb)")
statistic = "percentile"
percentile = c(5, 50, 95)
smoothTrend(tinsely3, pollutant = "no2", ylab = "NO2 concentration (ppb)",statistic =
"percentile",
percentile = c(25, 50, 75, 99))
smoothTrend(tinsely3, pollutant = "no", ylab = "NO concentration (ppb)",statistic =
"percentile",
percentile = c(25, 50, 75, 99))
names(tinsely3)<-c("date", "NO", "NO2", "site", "code", )
names(tinsely3)
timeVariation(tinsely3, c("NO", "NO2"), ylab="Concentrations (ppb)", normalise=T)
##################################################################################
require(leaps)
train2<-na.omit(train)
preds<-subset(train2, select=c("NO_mean", "NO_DG", "NO2_mean", "temp", "ws", "rh"))
a1<-regsubsets(preds, NO2_DG)
NO2_DG<-train2$NO2_DG
summary(a1)
head(train)
x2<-leaps(preds, NO2_DG, nbest=1, method="r2")
plot(x2$size-1, x2$r2, xlab="NO. of predictors", ylab="R2")
lines(spline(x2$size-1, x2$r2))
tt<-na.omit(dev_sensors)
NO2<-tt$NO2_DG
preds2<-subset(tt, select=c("no_1","no_2","no_3", "no_4", "no_5", "no_6", "no_7", "no_8",
"no_9", "no_10","no2_1","no2_2","no2_3", "no2_4", "no2_5", "no2_6", "no2_7", "no2_8",
"no2_9", "no2_10", "ws", "temp_mean","rh_mean"))
x3<-leaps(preds2, NO2, nbest=1, method="r2")
labels(preds2[1,])
preds3<-subset(tt, select=c("NO_mean", "NO2_mean","NO_DG", "ws",
"temp_mean","rh_mean"))
x4<-leaps(preds3, NO2, nbest=1, method="r2")
plot(x4$size-1, x4$r2, xlab="NO. of predictors", ylab="R2")
lines(spline(x4$size-1, x4$r2))
windows()
preds4<-subset(tt, select=c("NO_mean", "NO2_mean","NO_DG", "ws", "temp_mean"))
x5<-leaps(preds4, NO2, nbest=1, method="r2")
plot(x5$size-1, x5$r2, xlab="NO. of predictors", ylab="R2")
lines(spline(x5$size-1, x5$r2))
```

```
labels(preds4[1,])
###adjR2
preds2<-subset(tt, select=c("no_1","no_2","no_3", "no_4", "no_5", "no_6", "no_7", "no_8",
"no_9", "no_10","no2_1","no2_2","no2_3", "no2_4", "no2_5", "no2_6", "no2_7", "no2_8",
"no2_9", "no2_10", "ws", "temp_mean","rh_mean", "NO_DG", "NO_mean"))
x3<-leaps(preds2, NO2, nbest=1, method="adjr2")
plot(x3$size-1, x3$adjr2, xlab="NO. of predictors", ylab="adjR2")
lines(spline(x3$size-1, x3$adjr2))
labels(preds2[1,])
preds3<-subset(tt, select=c("NO_mean", "NO2_mean","NO_DG", "ws",
"temp_mean","rh_mean"))
x4<-leaps(preds3, NO2, nbest=1, method="adjr2")
plot(x4$size-1, x4$adjr2, xlab="NO. of predictors", ylab="adjR2")
lines(spline(x4$size-1, x4$adjr2))
windows()
preds4<-subset(tt, select=c("NO_mean", "NO2_mean","NO_DG", "ws", "temp_mean"))
x5<-leaps(preds4, NO2, nbest=1, method="adjr2")
plot(x5$size-1, x5$adjr2, xlab="NO. of predictors", ylab="adjR2")
lines(spline(x5$size-1, x5$adjr2))
labels(preds4[1,])
x5$which[which.max(x5$adjr2),]
timeVariation2<-fix(timeVariation)
timeVariation2(dev_sensors, pollutant=c("NO_mean","NO2_mean", "NO_DG",
"NO2_DG"), ylab="Concentrations (ppb)", ci=T)
timeVariation(dev_sensors, pollutant=c("NO_DG", "NO2_DG"), ylab="Concentrations
(ppb)", ci=T)
################################################################################
boxplot(no2[,c(-1, -12)], ylab="Concentrations (ppb)", xlab="Pollutants", ylim=c(0, 50))
boxplot(no[,c(-1, -12)], ylab="Concentrations (ppb)", xlab="Pollutants", ylim=c(0, 50))
boxplot(test, ylim=c(0, 150))
test2<-subset(test, test$NO_DG<60)
boxplot(test2, ylab="Concentrations (ppb)")
```

2. R Code for Chapter 6: Land Use Regression Model

```
#require(arcgisbinding)#for bridging r and arcGIS
#arc.check_product()
require(openair)#for corPlot and other visualisation
library(dplyr)#for renaming and selecting random sample
library(tidyverse)# for easy data manipulation and visualization
library(caret)# for easy machine learning workflow
library(leaps)# for computing stepwise regression
library(MASS)
#ozone.path <- system.file("extdata", "ca_ozone_pts.shp",
#                    package="arcgisbinding")
#ozone.arc.dataset <- arc.open(ozone.path)
```

```
        ######################################################################
        #Using Diffusion tubes data only
        dt<-read.csv("C:/Users/ci17sm/Desktop/AQdata2020/DTVar_selected157_CSV.csv")
        names(dt)
        dt<-dt[,-6]##to remove bus stops as it has only few observations
        #dt<-dt[,c(-5,-12)]
        (n<-nrow(dt))#157
        #75%
        set.seed(0000)
        (n75<-157*0.75)#118 rows
        (n25<-157*0.25)#39 rows
        train<-sample_n(dt,118)
        test<-sample_n(dt,39)
        lmtrain<-lm(NO2Conc~., data=train)
        summary(lmtrain)#R-squared 0.46 and adjusted R-squared 0.36
        pred_train<-predict(lmtrain)
        train$pred_train<-pred_train
        modStats(train, obs="NO2Conc", mod="pred_train")
        scatterPlot(train, x="NO2Conc", y="pred_train", mod.line=T,ylab="Predicted NO2",
        xlab="Measured NO2", xlim=c(0,60), ylim=c(0,60), main="LRM_Train")

        #Stepwise regression
        step.lmtrain <- stepAIC(lmtrain, direction = "both", trace = FALSE)
        summary(step.lmtrain)##R-squared 0.45, adjusted R-squared 0.41
        pred_swtrain<-predict(step.lmtrain)
        train$pred_swtrain<-pred_swtrain
        windows()
        scatterPlot(train, x="NO2Conc", y="pred_swtrain", mod.line=T,ylab="Predicted
NO2",
        xlab="Measured NO2", xlim=c(0,60), ylim=c(0,60), main="LM_Train_DT")
        modStats(train, obs="NO2Conc", mod="pred_swtrain")
        pred_swtest<-predict(step.lmtrain, newdata=test)
        test$pred_swtest<-pred_swtest
        windows()
        scatterPlot(test, x="NO2Conc", y="pred_swtest", mod.line=T,ylab="Predicted NO2",
        xlab="Measured NO2", xlim=c(0,60), ylim=c(0,60), main="LM_Test_DT")
        modStats(test, obs="NO2Conc", mod="pred_swtest")
         plot(train$NO2Conc, type="l")
        points(train$pred_swtrain, col="blue", type="l")
        plot(test$NO2Conc, type="l")
        points(test$pred_swtest, col="blue", type="l")
############################################################################
        ##Using Low_Cost Sensors data only
        lcs<-read.csv("C:/Users/ci17sm/Desktop/AQdata2020/LCS_LURVriables_CSV.csv")
        lcs<-lcs[,-6]##remove bus stops, as it has very few data points
        lcs$NO2Conc<-lcs$NO2Conc*1.91
        lmlcs<-lm(NO2Conc~.,data=lcs)
        summary(lmlcs)
        step.lmlcs <- stepAIC(lmlcs, direction = "both",
                      trace = FALSE)
        summary(step.lmlcs)
        set.seed(1234)
        (n75<-40*0.75)#30
        (n25<-40*0.25)#10
        trainlcs<-sample_n(lcs,30)
        testlcs<-sample_n(lcs,10)
lmtrainlcs<-lm(NO2Conc~Dist_MajRd+Dist_Comer+Dist_Resi+Altitude+Dist_Mway,
data=trainlcs)
        lmtrainlcs2<-lm(NO2Conc~., data=trainlcs)
        summary(lmtrainlcs2)
        step.lmtrainlcs <- stepAIC(lmtrainlcs2, direction = "both",
                    trace = FALSE)
        summary(step.lmtrainlcs)
        summary(lmtrainlcs2)#R-squared 0.30 and adjusted R-squared 0.15
        pred_trainlcs<-predict(lmtrainlcs)
        trainlcs$pred_trainlcs<-pred_trainlcs
        modStats(trainlcs, obs="NO2Conc", mod="pred_trainlcs")
        scatterPlot(trainlcs, x="NO2Conc", y="pred_trainlcs", mod.line=T,ylab="Predicted
NO2",
        xlab="Measured NO2", xlim=c(0,70), ylim=c(0,70), main="LRM_Train_LCS")
        pred_testlcs<-predict(lmtrainlcs, newdata=testlcs)
        testlcs$pred_testlcs<-pred_testlcs
        modStats(testlcs, obs="NO2Conc", mod="pred_testlcs")
        scatterPlot(testlcs, x="NO2Conc", y="pred_testlcs", mod.line=T,ylab="Predicted
NO2",
        xlab="Measured NO2", xlim=c(0,90), ylim=c(0,90), main="LRM_Test_LCS")
        #Stepwise regression
        #step.lmtrainlcs <- stepAIC(lmtrainlcs, direction = "both",
        #                   trace = FALSE)
        #summary(step.lmtrainlcs)##R-squared 0.89, adjusted R-squared 0.78
        #pred_swtrainlcs<-predict(step.lmtrainlcs)
        #trainlcs$pred_swtrainlcs<-pred_swtrainlcs
        #windows()
#scatterPlot(trainlcs, x="NO2Conc", y="pred_swtrainlcs", mod.line=T,ylab="Predicted
NO2",
        #xlab="Measured NO2", xlim=c(0,90), ylim=c(0,90), main="Stepwise_Train_LCS")
        #modStats(trainlcs, obs="NO2Conc", mod="pred_swtrainlcs")
        #pred_swtestlcs<-predict(step.lmtrainlcs, newdata=testlcs)
        #testlcs$pred_swtestlcs<-pred_swtestlcs
        windows()
        #scatterPlot(testlcs, x="NO2Conc", y="pred_swtestlcs", mod.line=T,ylab="Predicted
NO2",
        #xlab="Measured NO2", xlim=c(0,150), ylim=c(0,150),
main="Stepwise_Test_LCS")
        #write.csv(testlcs,"testlcs.csv")
        testlcs2<-read.csv("C:/Users/ci17sm/Desktop/AQdata2020/testlcs.csv")
        scatterPlot(testlcs2, x="NO2Conc", y="pred_swtestlcs", mod.line=T,ylab="Predicted
NO2",
        xlab="Measured NO2", xlim=c(0,100), ylim=c(0,100), main="Stepwise_Test_LCS")
##########################################################################
```

```
#using both diffusion tubes and sensors data
dtlcs<-
read.csv("C:/Users/ci17sm/Desktop/AQdata2020/NO2_Variables(100x100)_CSV.csv")
        names(dtlcs)
        (n<-nrow(dtlcs))#188
        #75%
        (n75<-188*0.75)#141 rows
        (n25<-188*0.25)#47 rows
        traindtlcs<-sample_n(dtlcs,141)
        testdtlcs<-sample_n(dtlcs,47)
        lmdtlcs<-lm(NO2Conc~.,data=traindtlcs)
        summary(lmdtlcs)#r-squared 0.31, adj r-squred0.18
        step.lmdtlcs <- stepAIC(lmdtlcs, direction = "both", trace = FALSE)
        summary(step.lmdtlcs)##R-squared 0.27, adjusted R-squared 0.23
        pred_traindtlcs<-predict(step.lmdtlcs)
        traindtlcs$pred_traindtlcs<-pred_traindtlcs
scatterPlot(traindtlcs, x="NO2Conc", y="pred_traindtlcs", mod.line=T,ylab="Predicted
NO2",
        xlab="Measured NO2", xlim=c(0,70), ylim=c(0,70),
main="Stepwise_Train_DTLCS")
        modStats(traindtlcs, obs="NO2Conc", mod="pred_traindtlcs")
        pred_testdtlcs<-predict(step.lmdtlcs, newdata=testdtlcs)
        testdtlcs$pred_testdtlcs<-pred_testdtlcs
        scatterPlot(testdtlcs, x="NO2Conc", y="pred_testdtlcs", mod.line=T,ylab="Predicted
NO2",
        xlab="Measured NO2", xlim=c(0,70), ylim=c(0,70), main="Stepwise_Test_DTLCS")
        modStats(testdtlcs, obs="NO2Conc", mod="pred_testdtlcs")
##############################################################################
        #Generalised additive modelling
#############################################################################
        require(mgcv)
        require(sp)
        dt<-read.csv("C:/Users/ci17sm/Desktop/AQdata2020/DTVar_selected157_CSV.csv")
        dt<-dt[,-6]##to remove bus stops as it has only few observations
        #dt<-dt[,c(-5,-12)]
        (n<-nrow(dt))#157
        gmdt<-gam(NO2Conc~s(Building)+s(Dist_Bstop)+s(Dist_MajRd)+s(Dist_MinRd)+
        s(Residential)+s(Commercial)+s(Altitude),data=dt)
        summary(gmtrain)
        plot(gmtrain)
        dt2<-subset(dt,
select=c("NO2Conc","Building","Dist_Bstop","Dist_MajRd","Dist_MinRd",
        "Residential","Commercial","Altitude"))
        corPlot(dt2)
        gm<-gam(NO2Conc~s(Dist_MajRd), data=dt)
        gm2<-gam(NO2Conc~s(Altitude), data=dt)
        gm3<-gam(NO2Conc~s(Residential), data=dt)
        dt3<-subset(dt2,Dist_Bstop<150)
        gm4<-gam(NO2Conc~s(Dist_Bstop), data=dt)
        plot(gm4)
        #75%
        (n75<-157*0.75)#118 rows
        (n25<-157*0.25)#39 rows
        set.seed(0000)
        train<-sample_n(dt,118)
        test<-sample_n(dt,39)
        ##in the model using output of stepwise reg model
        gmtrain<-
gam(NO2Conc~s(Building)+s(Dist_Bstop)+s(Dist_MajRd)+s(Dist_MinRd)+
        s(Residential)+s(Commercial)+s(Altitude),data=train)
        summary(gmtrain)
        plot(gmtrain)
        pred_gmtrain<-predict(gmtrain)
        train$pred_gmtrain<-pred_gmtrain
        modStats(train, obs="NO2Conc", mod="pred_gmtrain")
        scatterPlot(train, x="NO2Conc", y="pred_gmtrain", mod.line=T,ylab="Predicted
NO2",
        xlab="Measured NO2", xlim=c(0,60), ylim=c(0,60), main="GAM_Train_DT")
        pred_gmtest<-predict(gmtrain, newdata=test)
        test$pred_gmtest<-pred_gmtest
        windows()
        scatterPlot(test, x="NO2Conc", y="pred_gmtest", mod.line=T,ylab="Predicted NO2",
        xlab="Measured NO2", xlim=c(0,60), ylim=c(0,60), main="GAM_Test_DT")
        modStats(test, obs="NO2Conc", mod="pred_gmtest")
        griddata<-read.csv("C:/Users/ci17sm/Desktop/AQdata2020/grid_37605.csv")
        pred_gmgrid<-predict(gmtrain, newdata=griddata)
        names(griddata)
        griddata$pred_gmNO2dt<-pred_gmgrid
        write.csv(griddata, "griddata.csv")
        #this data is uploaded into arcGIS (and join into a polygon layer)to map it on
Sheffield area
#################################################################################
        ##Using Low_Cost Sensors data only
        lcs<-read.csv("C:/Users/ci17sm/Desktop/AQdata2020/LCS_LURVriables_CSV.csv")
        lcs<-lcs[,-6]##remove bus stops, as it has very few data points
        lcs$NO2Conc<-lcs$NO2Conc*1.91##convert into ug/m3
        #step.lmlcs <- stepAIC(lmlcs2, direction = "both",
        #                    trace = FALSE)
        #summary(step.lmlcs)
        #summary(lmlcs)
        #lcs<-lcs[,c(-5,-12)]
        #(n<-nrow(lcs))#40
        #75%
        set.seed(1234)
        (n75<-40*0.75)#30
        (n25<-40*0.25)#10
        trainlcs<-sample_n(lcs,30)
        testlcs<-sample_n(lcs,10)
        set.seed(1234)
        gmtrainlcs<-gam(NO2Conc~s(Dist_MajRd)+s(Dist_Comer)+s(Altitude),
data=trainlcs)
        summary(gmtrainlcs)#adjusted R-squared 0.66, deviance explained 79.6%
        pred_gmtrainlcs<-predict(gmtrainlcs)
        trainlcs$pred_gmtrainlcs<-pred_gmtrainlcs
        modStats(trainlcs, obs="NO2Conc", mod="pred_gmtrainlcs")
        scatterPlot(trainlcs, x="NO2Conc", y="pred_trainlcs", mod.line=T,ylab="Predicted
NO2",
        xlab="Measured NO2", xlim=c(0,70), ylim=c(0,70), main="GAM_Train_LCS")
        pred_gmtestlcs<-predict(gmtrainlcs, newdata=testlcs)
        testlcs$pred_gmtestlcs<-pred_gmtestlcs
        modStats(testlcs, obs="NO2Conc", mod="pred_gmtestlcs")
        scatterPlot(testlcs, x="NO2Conc", y="pred_gmtestlcs", mod.line=T,ylab="Predicted
NO2",
        xlab="Measured NO2", xlim=c(0,90), ylim=c(0,90), main="GAM_Test_LCS")
        griddata<-read.csv("C:/Users/ci17sm/Desktop/AQdata2020/grid_37605.csv")
        pred_gmlcsgrid<-predict(gmtrainlcs, newdata=griddata)
        names(griddata)
        griddata$pred_gmNO2lcs<-pred_gmlcsgrid
        write.csv(griddata, "griddata2.csv")
#############################################################################
        #using both diffusion tubes and sensors data
dtlcs<-
read.csv("C:/Users/ci17sm/Desktop/AQdata2020/NO2_Variables(100x100)_CSV.csv")
        names(dtlcs)
        (n<-nrow(dtlcs))#188
        #75%
        (n75<-188*0.75)#141 rows
        (n25<-188*0.25)#47 rows
        set.seed(2222)
        traindtlcs<-sample_n(dtlcs,141)
        testdtlcs<-sample_n(dtlcs,47)
        #gmtraindtlcs<-gam(NO2Conc~s(MajRd)+s(MinorRd)+s(Dist_Mway)+s(Industrial)+
        #s(Pop_num)+s(Dist_MajRd)+s(Commercial)+
s(Dist_Parks)+s(Dist_Comer)+s(Dist_Resi)+s(Altitude)+s(Dist_Bstop),
        #data=traindtlcs)
        #summary(gmtraindtlcs)
gmtraindtlcs<-gam(NO2Conc~s(Industrial)+s(Pop_num)+s(Dist_MajRd)+s(Commercial)+
        s(Dist_Parks)+s(Dist_Comer)+s(Dist_Resi)+s(Altitude), data=traindtlcs)
        summary(gmtraindtlcs)
        pred_gmtraindtlcs<-predict(gmtraindtlcs)
        traindtlcs$pred_gmtraindtlcs<-pred_gmtraindtlcs
scatterPlot(traindtlcs, x="NO2Conc", y="pred_gmtraindtlcs", mod.line=T,ylab="Predicted
NO2",
        xlab="Measured NO2", xlim=c(0,70), ylim=c(0,70), main="GAM_Train_DTLCS")
        modStats(traindtlcs, obs="NO2Conc", mod="pred_gmtraindtlcs")
        pred_gmtestdtlcs<-predict(gmtraindtlcs, newdata=testdtlcs)
        testdtlcs$pred_gmtestdtlcs<-pred_gmtestdtlcs
scatterPlot(testdtlcs, x="NO2Conc", y="pred_gmtestdtlcs", mod.line=T,ylab="Predicted
NO2",
        xlab="Measured NO2", xlim=c(0,70), ylim=c(0,70), main="GAM_Test_DTLCS")
        modStats(testdtlcs, obs="NO2Conc", mod="pred_gmtestdtlcs")
        pred_gmdtlcsgrid<-predict(gmtraindtlcs, newdata=griddata)
        names(griddata)
        griddata$pred_gmNO2dtlcs<-pred_gmdtlcsgrid
        write.csv(griddata, "griddata_dtlcs.csv")
path2<-"C:/Users/ci17sm/Documents/ArcGIS/Default.gdb/Join_Output_2"
        d2<-arc.open(path2)
        data100<-arc.select(object=d2,fields=c("POINT_X", "POINT_Y",
"RASTERVALU","MajRd", "MinorRd","Building","Avg_No2Con",
        "Industrial",
"Parks","Bus_Stops","St_Interse","Pop_num","Dist_MajRd","Dist_MinRd","Residentia","Co
mmercial"))
        names(dr2)
dr2<-dr %>% rename(NO2Conc=Avg_No2Con, Easting=POINT_X, Northing=POINT_Y,
Altitude=RASTERVALU,
Residential=Residentia,St_Intersect=St_Interse)#change names of NO2Conc etc
        names(dr2)
        dr3<-arc.select(object=d)
        names(dr3)
        dr4<-dr3 %>% rename(ID=OBJECTID,ID_1=OBJECTID_1,uid=UID,
easting=POINT_X,northing=POINT_Y,
        UnID=UniqID,altitude=RASTERVALU,Easting2=Avg_Eastin,
Northing2=Avg_Northi, NO2Conc=Avg_No2Con,
        Majln100=Avg_MajLn1,MajLn200=Avg_MajLn2, MajLn300=Avg_MajLn3,
MajLn500=Avg_MajLn5,
        MinLn100=Avg_MinLn1,MinLn200=Avg_MinLn2,MinLn300=Avg_MinLn3,
MinLn500=Avg_MinLn5, StInt100=Avg_StInt1,
        StInt200=Avg_StInt2, StInt300=Avg_StInt3, StInt500=Avg_StInt5,
BusStp100=Avg_BusStp,
        BusStp200=Avg_BusS_1,BusStp300=Avg_BusS_2,BusStp500=Avg_BusS_3,
AADT100=Avg_AADT10,
        AADT200=Avg_AADT20,AADT300=Avg_AADT30,AADT500=Avg_AADT50,Re
s100=Avg_ResAr1,Res200=Avg_ResAr2,
        Res300=Avg_ResAr3, Res500=Avg_ResAr5, Ind100=Avg_IndAr1,
Ind200=Avg_IndAr2,Ind300=Avg_IndAr3,
        Ind500=Avg_IndAr5, Com100=Avg_ComAr1, Com200=Avg_ComAr2,
Com300=Avg_ComAr3,Com500=Avg_ComAr5,
        Park100=Avg_ParkAr,Park200=Avg_Park_1,Park300=Avg_Park_2,Park500=Avg_P
ark_3,Wood100=Avg_WoodAr,
        Wood200=Avg_Wood_1,Wood300=Avg_Wood_2,
Wood500=Avg_Wood_3,Pop100=Avg_PopNum,Pop200=Avg_PopN_1,
        Pop300=Avg_PopN_2,Pop500=Avg_PopN_3,Build100=Avg_BuilAr,
Build200=Avg_Buil_1,Build300=Avg_Buil_2,
        Build500=Avg_Buil_3, Altitude=Avg_Altitu, DisTrain=Avg_DisTra,
DisBus=Avg_DisBus, TramID=Avg_Tram_I,
        TramDist=Avg_DisT_1)
        dr5<-dr4[,21:72]
        lm2<-lm(NO2Conc~.,data=dr5)
        step.model2 <- stepAIC(lm2, direction = "both",  trace = FALSE)
        summary(step.model2)
lm3<lm(NO2Conc~MajLn500+MinLn200+BusStp200+Ind200+Com300+Park100+Build20
0+Altitude, data=dr5)
        path<-"C:/Users/ci17sm/Documents/ArcGIS/Default.gdb/Join_Outputs_2_Selection"
        d<-arc.open(path)
        dr<-arc.select(object=d,fields=c("POINT_X", "POINT_Y",
"RASTERVALU","MajRd",
"MinorRd","Building","Avg_No2Con","Industrial","Parks","Bus_Stops","St_Interse","Pop_n
um","Dist_MajRd","Dist_MinRd","Residentia","Commercial"))
        names(dr)
        dr2<-dr %>% rename(NO2Conc=Avg_No2Con, Easting=POINT_X,
Northing=POINT_Y, Altitude=RASTERVALU,
        Residential=Residentia,St_Intersect=St_Interse)#change names of NO2Conc etc
        names(dr2)
        corPlot(dr2)
```

3. R Code for Chapter 7: Data fusion

```
#how to make meuse_grid from meuse data
#You can examine the fitted values for the variogram with the "summary()" function.
#The sill is the "Mean" value of the "psill" values while the nugget is the "Min"
#of the "psill". The range is in the "Mean" value for the "range".
require(gstat)
library(sp)
require(automap)
data(meuse)
coordinates(meuse) = ~x+y
meuse_grid = spsample(meuse, type = "regular", cellsize = c(40,40))
gridded(meuse_grid) = TRUE
plot(meuse_grid)
data(meuse.grid)
gridded(meuse.grid) = ~x+y
gridded(meuse_grid) = ~x+y
meuse_grid2 = spsample(meuse, type = "regular", cellsize = c(100,100))
gridded(meuse_grid) = TRUE
plot(meuse_grid2)
head(meuse)
nrow(meuse)
plot(meuse$x, meuse$y)
####################################
#sheffield example
#aq<-read.csv(file.choose(), header=T)
aq<-read.csv("C:/Users/ci17sm/Documents/AQ Data/2020/croppedAirviro.csv")
aq2<-subset(aq,select=c("x","y", "Altitude", "NO2_AV", "NO2_msrd"))
gridded(aq2) = ~x+y
names(aq2)<-c("x", "y","altitude","no2av","no2msrd")
names(aq2)
aq3<-subset(aq2, no2msrd>0)##remove rows having "zero" value
aq4<-subset(aq2, no2msrd<10)
aq4_grid<-read.csv("C:/Users/ci17sm/Documents/AQ Data/2020/croppedAirviro_grid.csv")
gridded(aq4_grid) = ~x+y
coordinates(aq3) = ~x+y
#aq_grid = spsample(aq3, type = "regular", cellsize = c(100,100))
#gridded(aq_grid) = TRUE
#plot(aq_grid)
plot(aq3$x, aq3$y, xlab="Easting", ylab="Northing")

kriging_result = autoKrige(no2msrd~1, aq3)#works
plot(kriging_result)#works
kriging_result2 = autoKrige(no2msrd~no2av+altitude, aq3, aq4_grid)##
plot(kriging_result2)
print(kriging_result2)
names(kriging_result2)
kr<-kriging_result2##
names(kr)#krige_output, exp_var, var_mod,sserr
names(kr$krige_output)
p<-kr$krige_output$var1.pred
aq4_grid$pred<-p
#####################################################################
kriging_result3 = autoKrige(no2msrd~no2av, aq3, aq2)##
plot(kriging_result3)
print(kriging_result3)
names(kriging_result3)
kr2<-kriging_result3
names(kr)#krige_output, exp_var, var_mod,sserr
names(kr2$krige_output)
p2<-kr2$krige_output$var1.pred
aq2$pred<-p2
write.csv(aq2, "aq2.csv")
#############################################################
#Variogram using automap
data(meuse)
coordinates(meuse) =~ x+y
variogram = autofitVariogram(no2msrd~1,aq3)
plot(variogram)
variogram2 = autofitVariogram(no2msrd ~ no2av + altitude, aq3)
plot(variogram2)
#################################################
aq<-read.csv("C:/Users/ci17sm/Documents/AQ Data/2020/DT_DTLCS_LCS_AV_LUR.csv")
names(aq)
aq2<-subset(aq, NO2DT>0)##remove rows having "zero" value to de used as a training dataset
##aq will be used as testing dataset, which contain the testing dataset as well
coordinates(aq2) = ~x+y
gridded(aq) = ~x+y
```

```
kr = autoKrige(NO2DT~NO2AV, aq2, aq)##
plot(kr)
print(kr)
names(kr)
names(kr)#krige_output, exp_var, var_mod,sserr
names(kr$krige_output)
p<-kr$krige_output$var1.pred
aq$pred_AVDT<-p
write.csv(aq, "aq_AVDT.csv")
########################################
aq<-read.csv("C:/Users/ci17sm/Documents/AQ
Data/2020/DT_DTLCS_LCS_AV_LUR.csv")
names(aq)
aq3<-subset(aq, NO2LCS>0)##remove rows having "zero" value to de used as a training
dataset
##aq will be used as testing dataset, which contain the testing dataset as well
coordinates(aq3) = ~x+y
gridded(aq) = ~x+y
kr2 = autoKrige(NO2LCS~NO2AV, aq3, aq)##
plot(kr2)
print(kr2)
names(kr2)#krige_output, exp_var, var_mod,sserr
names(kr2$krige_output)
p2<-kr2$krige_output$var1.pred
aq$pred_AVLCS<-p2
write.csv(aq, "aq_AVLCS.csv")
##################################################
aq<-read.csv("C:/Users/ci17sm/Documents/AQ
Data/2020/DT_DTLCS_LCS_AV_LUR.csv")
names(aq)
aq4<-subset(aq, NO2DT>0)##remove rows having "zero" value to de used as a training
dataset
##aq will be used as testing dataset, which contain the testing dataset as well
coordinates(aq3) = ~x+y
gridded(aq) = ~x+y
kr3 = autoKrige(NO2DT~NO2GAM, aq4, aq)##
plot(kr3)
print(kr3)
names(kr3)#krige_output, exp_var, var_mod,sserr
names(kr3$krige_output)
p3<-kr3$krige_output$var1.pred
aq$pred_lurdt<-p3
write.csv(aq, "aq_lurdt.csv")
#########################################################
kr4 = autoKrige(NO2LCS~NO2GAM, aq3, aq)##
plot(kr4)
print(kr4)
names(kr4)#krige_output, exp_var, var_mod,sserr
names(kr4$krige_output)
p4<-kr4$krige_output$var1.pred
```

```
aq$pred_lurlcs<-p4
write.csv(aq, "aq_lurlcs.csv")
####################################
aq5<-subset(aq, NO2LCSDT>0)
kr5 = autoKrige(NO2LCSDT~NO2GAM, aq5, aq)##
plot(kr5)
print(kr5)
names(kr5)#krige_output, exp_var, var_mod,sserr
names(kr4$krige_output)
p5<-kr4$krige_output$var1.pred
aq$pred_lurlcsdt<-p5
write.csv(aq, "aq_lurlcsdt.csv")
```

4. R code for Chapter 8 – time series analysis

```
#libraries
library(Hmisc)
#library('ggplot2')
library('forecast')
library('tseries')
library(tsDyn)
library(openair)
require(TSA)
modelt <- auto.arima(no2.train, xreg=NULL)
pred2<-predict_rolling(modelt, newdata=no2.test)
pred3 <- predict_rolling(modelt,newdata=no2.train)
str(pred2)
yx<-pred2$pred$no2.train
xx<-pred2$true$no2.train
test$yx<-yx
test$xx<-xx
scatterPlot(test, x="xx", y="yx",mod.line=T, xlim=c(0,90), ylim=c(0,70), ylab="predicted
NO2", xlab="Observed NO2")
par(mai=c(1,1,0.5,0.5))
plot(test$date, test$no2, xlab="Date", ylab=quickText("NO2 concentrations (ug/m3)"),
pch=1)
points(test$date, test$yx, col="red",pch=3)
legend("topleft",col=c("black", "red"), legend=c("observed","predicted"), pch=c(1,3))
modStats(test, mod="yx", obs="xx")
plot(rstandard(modelt),ylab ='Standardized Residuals',type='o'); abline(h=0)
qqnorm(residuals(modelt))
qqline(residuals(modelt))
hist(residuals(modelt), xlab="residuals", main="")
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
yx3<-pred3$pred$no2.train
xx3<-pred3$true$no2.train
train$yx3<-yx3
train$xx3<-xx3
```

```
scatterPlot(train, x="xx3", y="yx3",mod.line=T, xlim=c(0,90), ylim=c(0,70), ylab="predicted NO2", xlab="Observed NO2")
windows()
par(mai=c(1,1,0.5,0.5))
plot(train$date, train$no2, xlab="Date", ylab=quickText("NO2 concentrations (ug/m3)"), pch=1)
points(train$date, train$yx3, col="red",pch=3)
legend("topleft",col=c("black", "red"), legend=c("observed","predicted"), pch=c(1,3))
modStats(train, mod="yx3", obs="xx3")
##########################################################################
modelx <- auto.arima(no2.train, xreg=data.matrix(no.train))
newdata2<-no2.test
newdata<-no2.train
predx<-predict(modelx,newdata=no2.train, newxreg=data.matrix(no.train))
predx2<-predict(modelx,newdata=no2.test, newxreg=data.matrix(no.test))
train$fittedx.NO2<-predx$pred
test$fittedx2.NO2<-predx2$pred
plot(train$no2, train$fittedx.NO2)
plot(test$no2, test$fittedx2.NO2, ylim=c(0,60), xlim=c(0,60))
scatterPlot(train, x="no2", y="fittedx.NO2", mod.line=T, linear=T, xlim=c(0, 100))
plot(train$date, train$no2, xlab="Date", ylab=quickText("NO2 concentrations (ug/m3)"), pch=1)
points(train$date, train$fittedx.NO2, col="red",pch=3)
legend("topleft",col=c("black", "red"), legend=c("observed","fitted"), pch=c(1,3))
modStats(train, mod="fittedx.NO2", obs="no2")
modStats(test, mod="fittedx2.NO2", obs="no2")
plot(predx2$pred, main="")
plot(rstandard(model3),ylab ='Standardized Residuals',type='o', ylim=c(-4, 4)); abline(h=0)
qqnorm(residuals(model3), main=""); qqline(residuals(model3))
hist(residuals(model3), xlab="residuals", main="")
##########################################################################
df<-data.frame(nox.train, o3.train)
df2<-data.frame(nox.test, o3.test)
model4<-auto.arima(no2.train, xreg=data.matrix(df))#same
#model4<-auto.arima(train$no2, xreg=data.matrix(df))#same
pred4 <- predict(model4, newxreg=as.matrix(df))
pred5<-predict(model4, newxreg=as.matrix(df2), newdata=no2.test)
train$fitted_NO2<-pred4$pred
test$fitted_NO2<-pred5$pred
scatterPlot(train, x="no2", y="fitted_NO2", linear=T, mod.line=T)
scatterPlot(test, x="no2", y="fitted_NO2", linear=T, mod.line=T)
plot(train$date, train$no2, xlab="Date", ylab=quickText("NO2 concentrations (ug/m3)"), pch=1)
points(train$date, train$fitted_NO2, col="red",pch=3)
legend("topleft",col=c("black", "red"), legend=c("observed","fitted"), pch=c(1,3))
plot(test$date, test$no2, xlab="Date", ylab=quickText("NO2 concentrations (ug/m3)"), pch=1)
points(test$date, test$fitted_NO2, col="red",pch=3)
legend("topleft",col=c("black", "red"), legend=c("observed","fitted"), pch=c(1,3))
```

```
#timePlot(df, c("no2", "fitted_NO2"), group=T, ylab="log(NO2 concentrations)",
col=c("blue", "red"))
plot(pred4, main="")
modStats(train, mod="fitted_NO2", obs="no2")
modStats(test, mod="fitted_NO2", obs="no2")
plot(rstandard(model4),ylab ='Standardized Residuals',type='o'); abline(h=0)
qqnorm(residuals(model4)); qqline(residuals(model4))
hist(residuals(model4))

#The End
```