Development of Data Analytic Approaches for Air Quality Data

Stuart Kenneth Grange

Doctor of Philosophy

University of York

Chemistry

January 2019

Abstract

Continuous air quality monitoring networks were commissioned in the mid-twentieth century throughout the developed world to underpin the understanding of air pollution. These monitoring networks have produced a vast observational record which continues to grow. However, these data are generally used for simple tasks such as checking for compliance to legal standards or guidelines and the additional information contained in the data sets is not well leveraged to aid scientific understanding and inform policy makers. This thesis addresses this issue and has the goal of extracting additional information from "routine" air quality monitoring data using new, and novel data analyses with a focus on the impact of transportation activities across Europe. Specifically, this thesis outlines the development of bivariate polar plots with pair-wise statistics to aid source apportionment, the development of a European air quality database which much of this thesis's work is based on, a European-wide analysis of roadside nitrogen dioxide (NO_2), and the development of a framework and software to robustly detect and quantify changes in pollutant concentrations. The additional functionality of bivariate polar plots was useful for isolating the natural and anthropogenic sources of pollutants and is now included in the open source **openair** R package. The NO_2 analysis revealed that directly emitted NO_2 from road vehicles is decreasing across Europe and assumed emissions are too high resulting in pessimistic projections of future compliance. This conclusion is very important for policy makers to consider in their planning of disruptive interventions, most relevant of which are low emission zones because the observations suggest that the outlook is better than traditionally thought. For those analysing trends, a new technique has been developed that is highly effective at robustly characterising and quantifying the effects of interventions and the tools developed are available in the form of the open source **rmweather** R package.

Contents

List of figures

List of tables

Acknowledgements

The completion of my Doctor of Philosophy programme took support from a number of parties. I thank Anthony (Tony) Wild with the provision of the Wild Fund Scholarship (sponsor number: MKPTGOEMX) and I was also partially funded by the Natural Environment Research Council (NERC).

I thank my supervisors, David Carslaw and Ally Lewis for their research direction and input during my time in the Wolfson Atmospheric Chemistry Laboratories (WACL). My colleagues in WACL were always supportive and happy to help over the course of my programme. Mat Evans was my Independent Panel Member (IPM) and was supportive throughout my studies. One of the most important things in my life is my running and exploring the North York Moors, the Yorkshire Dales, and the Lake District was always one of my favourite activities while living in York. I thank York Knavesmire Harriers for the camaraderie during this time. I thank my friends and family at home in New Zealand and the support I received while I was living and studying (literally) a world away from home.

I was based at Empa in Dübendorf, Zürich, Switzerland for the second half of 2017. I thank Christoph Hüglin and Erini Boleti, and the wider Air Pollution/Environmental Technology group led by Lukas Emmenegger for their organisation for this placement and their support while living in Switzerland.

Author's declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References. This thesis is based on four peer reviewed publications with myself as the lead author. The details of the articles are below:

Grange, S. K., Lewis, A. C., and Carslaw, D. C. Source apportionment advances using polar plots of bivariate correlation and regression statistics. *Atmospheric Environment* 145 (2016), pp. 128–134. https://doi.org/10.1016/j.atmosenv.2016.09.016.

Grange S. K., Lewis, A. C., Moller, S. J., and Carslaw, D. C. Lower vehicular primary emissions of NO_2 in Europe than assumed in policy projections. *Nature Geoscience* 10.12 (2017), pp. 914–918. https://doi.org/10.1038/s41561-017-0009-0.

Grange, S. K., Carslaw, D. C., Lewis, A. C., Boleti, E., and Hueglin, C. Random forest meteorological normalisation models for Swiss PM_{10} trend analysis. *Atmospheric Chemistry and Physics* 18.9 (2018), pp. 6223–6239. https://www.atmos-chem-phys.net/18/6223/2018/.

Grange, S. K. and Carslaw, D. C. Using meteorological normalisation to detect interventions in air quality time series. *Science of the Total Environment* 653 (2018), pp. 578–588 https://doi.org/10.1016/j.scitotenv.2018.10.344.

The above publications compose Chapters 2, 4, 5 and 6 respectively and are supplemented with this published R package:

Grange, S. K. **rmweather**: Tools to Conduct Meteorological Normalisation on Air Quality Data. R package version 0.1.3. 2018. https://CRAN.R-project.org/package=rmweather.

Chapter 1

Introduction

1.1 Setting the scene

1.1.1 Urban air quality

Urban air quality has become a key environmental and public health issue over the last three decades. The consequences of poor air quality are exclusively negative and research continues to accurately quantify the health and financial cost of poor air quality on society, but the effects are considered vast.[1,2] Currently, the World Health Organisation (WHO) attribute 4.2 million deaths a year to exposure to poor outdoor air quality and estimate 91 % of the human population are exposed to polluted air.[3] Many global risk assessments of causes of death place outdoor air pollution well within the top ten risk factors which cause death and is considered the leading environmental cause of premature death (Figure 1.1).[4–6] However, exposure to poor air quality has a myriad of negative health effects other than premature death (Figure 1.2).[7] These factors place very large burdens on society and the global cost of air pollutant burdens has been estimated at $5.11 trillion (using 2013 as the analysis year).[4]

Urban air pollution is a global issue, but the negative consequences of poor air quality are disproportionally high in the under-developed and developing world.[9] However, the developed world also faces significant chal-

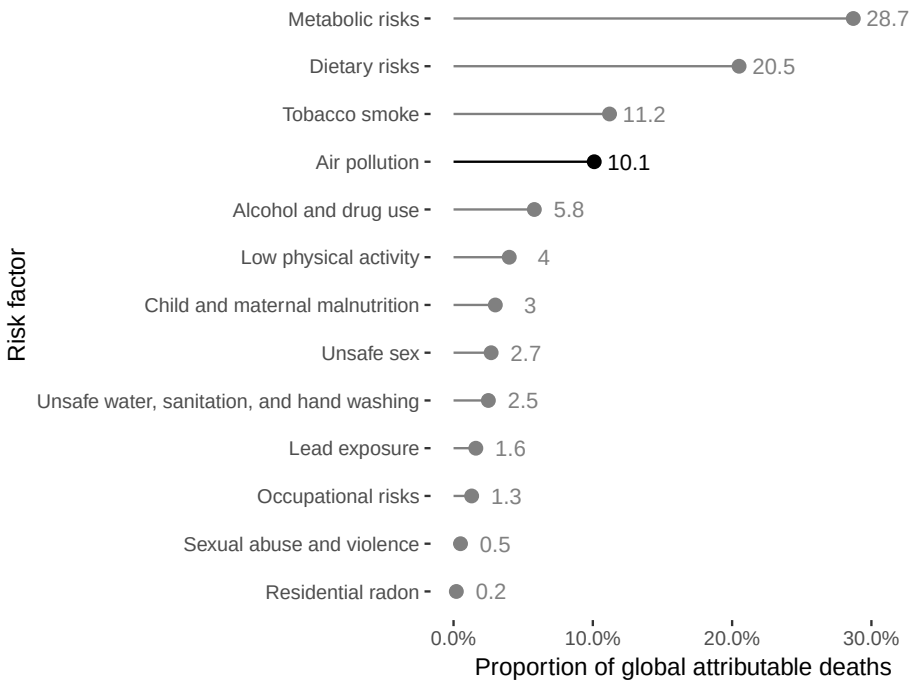

Figure 1.1: Global attributable deaths by risk factor.[4] Air pollution is the leading environmental cause of premature death.

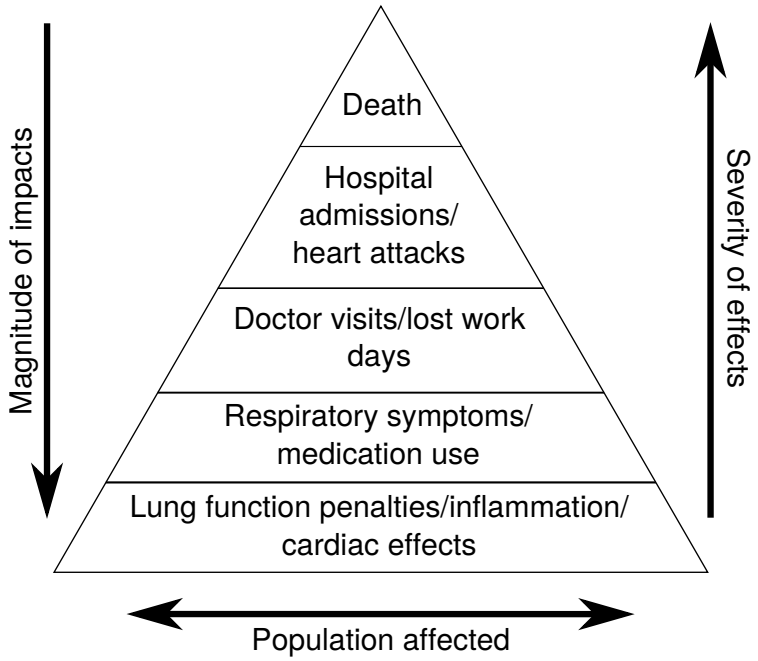

Figure 1.2: The pyramid of effects to poor air quality within a population, adapted from United States Environmental Protection Agency [8].

lenges and invest the most and lead the way in poor air quality mitigation efforts.[10] In 2008, the landmark of the majority of the world's population resided in urban areas was met[11] which makes outdoor urban air quality a relevant issue for the majority of the world's population.[12]

Air pollution has both natural and anthropogenic sources.[3] The majority of issues associated with poor air quality are due to anthropogenic processes, but the combination of the presence of natural and human sourced pollutants can be important at times.[13] Management of air pollution focuses on anthropogenic processes because such activities are more easily controlled compared to natural processes. It can be seen that poor air quality is an issue which can be solved, but there are significant economic implications of doing so.[14] The costs involved with managing air quality leads to exposure inequalities and this can be observed when comparing underdeveloped, developing, and developed countries, but also when exploring the differences among relative rich and poor within countries.[15]

1.1.2 Air pollutants

There are thousands of pollutants which contaminate urban atmospheres. However, there are a handful of pollutants which are considered "classical" pollutants, are regularly monitored, where legislation exists to control them, and are explored in this thesis. These pollutants are: particulate matter (PM) of various size fractions the most commonly encountered of which are PM_{10} and $PM_{2.5}$ (particulate matter 10 and 2.5 micrometers or less in diameter respectively), ozone (O_3), oxides of nitrogen (NO_x), sulfur dioxide (SO_2), and carbon monoxide (CO). Both SO_2 and CO have been effectively controlled in the developed world because of improvements in the quality of fuels and combustion processes and are rarely an operational issue in the outdoor environment. The World Health Organisation (WHO) outline ambient air quality guidelines for these pollutants with the exception of CO because there is strong evidence about the negative consequences of these pollutants (Table 1.1).[9] Table 1.1 demonstrates that air pollutants

have both acute and chronic (short and long term) effects hence different guideline values for different aggregation periods.

Table 1.1: World Health Organisation (WHO) Air Quality Guideline values.[9]

Pollutant	Value	Summary type
$PM_{2.5}$	$10\,\mu g\,m^{-3}$	Annual mean
$PM_{2.5}$	$25\,\mu g\,m^{-3}$	24-hour mean
PM_{10}	$20\,\mu g\,m^{-3}$	Annual mean
PM_{10}	$50\,\mu g\,m^{-3}$	24-hour mean
O_3	$100\,\mu g\,m^{-3}$	8-hour mean
NO_2	$40\,\mu g\,m^{-3}$	Annual mean
NO_2	$200\,\mu g\,m^{-3}$	24-hour mean
SO_2	$20\,\mu g\,m^{-3}$	Annual mean
SO_2	$500\,\mu g\,m^{-3}$	24-hour mean

For the negative consequences of air pollution to be realized, exposure to the pollutant must occur. Therefore, there must be an interaction of emission or generation of pollutants, dispersion, and a subsequent exposure for the negative effects to be felt by individuals and populations. Urban areas are more susceptible to this interaction because of the combination of high population density and intensive consumption of resources to meet economic demands which results in the release of pollutants. At an individual level, little can be done to avoid or reduce exposure to air pollutants and therefore effective management requires the action of local, regional, and national policy makers.[9,14]

The mechanisms of air pollutants' negative effects on human health are diverse and are dependent on the pollutant, but the common diseases in order of prevalence are generally considered: heart disease, strokes, pulmonary disease, lower respiratory infections, and lung cancer.[9] There are however far more less dramatic negative effects (Figure 1.2) which are not reported and are poorly quantified due to complicated relationships with

other risk factors. For excellent reviews and summaries of the human health effects of air quality see Cohen et al. [16], Nel [17], Curtis et al. [18], Pope and Dockery [19], and Kampa and Castanas [20].

1.1.3 European context

In Europe, poor air quality within and near roadside environments has become the dominating focus of air quality management. Air quality issues surrounding road transport are far more common than industrial or residential (generally wood burning) activities, but these former issues are important in some European locations nevertheless.[21–25] There are examples of very large and disruptive intervention efforts being applied to European urban areas such as low emission zones (also called clean air zones), the banning of private vehicles, and the promotion of cycling and public transportation in an attempt to improve air quality.[26–31]

The majority of these roadside air quality issues have arisen due to Europe's rapid, and unique dieselisation of its passenger vehicle fleet.[32] New sales of diesel-powered passenger vehicles in Europe peaked at 56 % in 2011 (Figure 1.3) and in most countries compose ≈45 % of the in-service fleet.[33] Diesel powered vehicles emit more NO_x, the sum of nitrogen oxide (NO) and nitrogen dioxide (NO_2), and very fine particulates compared to gasoline (petrol) fuelled vehicles. The Volkswagen (VW) diesel emission scandal, also known as "dieselgate" or "emissiongate"[34–36] broke in late September 2015 and further promoted issues related to vehicle emissions and poor air quality in the public, and therefore political domains. The European Environment Agency (EEA) estimates that 399 000, 75 000, and 13 600 premature European deaths are caused by exposure to $PM_{2.5}$, NO_2, and O_3 respectively.[37,38]

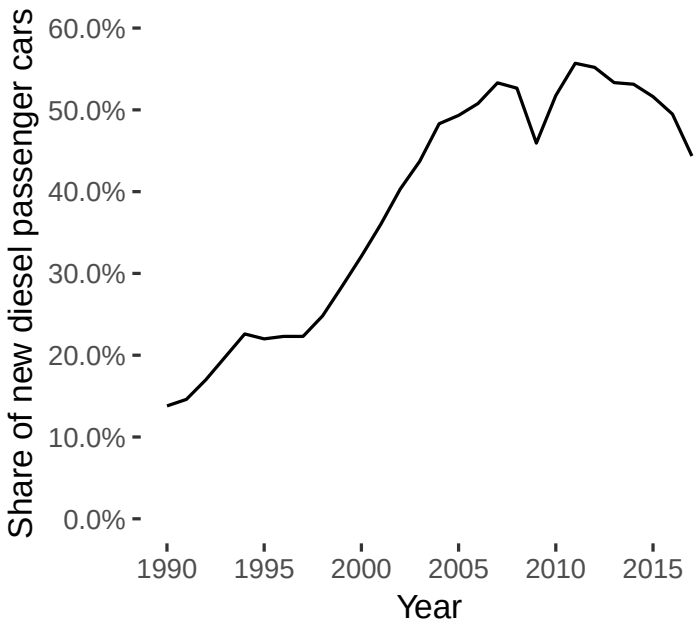

Figure 1.3: Market share of diesel powered passenger vehicles sold in Europe between 1990 and 2017.[33]

1.1.3.1 Volkswagen diesel emission scandal

The Volkswagen diesel emission scandal, also known as "dieselgate" outlined a major issue with the generality of vehicle emissions as measured in test procedures to real-world environments. A defeat device was first found within a popular Volkswagen AG (VW) 2.0 litre diesel engine (engine code: EA189) manufactured between 2009 and 2015 in the form of a piece of software in September 2015 by a research group at West Virginia University.[39,40] This defeat device was designed to subvert laboratory testing by changing engine dynamics when a test cycle was detected and then deliver better performance and fuel economy in all other situations, but at the consequence of a factor of 10 to 40 greater NO_x emissions than the legal limits impose.[39,41–43] The software within the vehicles' electronic control module (ECU) would evaluate the electronic stability control status, barometric pressure, speed of vehicle, steering position, and engine run time to determine if the vehicle was within a test cycle. If these tests were positive, the ECU would change engine operation to reduce emissions to enable the

vehicle to pass the test.[44]

Eleven million vehicles worldwide were embroiled in the scandal and some critics have labelled the deception as a crime against humanity due to the air quality consequences of the increased NO_x emissions.[41,45] In the United States, VW was legally required to purchase vehicles from owners in some situations, and the locations where these vehicles were stored gives some indication on how many vehicles were involved in the scandal (Figure 1.4).

To date, similar and sophisticated VW-like defeat devices have not been been found in other manufacturers' vehicles with the possible exception of Daimler AG (Mercedes-Benz),[47] but it seems that most manufacturers have engaged in emission control manipulation to enable their vehicles to pass type approval tests, rather than focus on true solutions to reduce emission in realistic on-road operation cycles.[48,49] Porsche and BMW have recalled 60 000 and 12 000 vehicles respectively for diesel emission control reasons, however neither have admitted illegal activities and Porsche no longer offers any diesel powered vehicles.[50–52] This has given rise to an issue where diesel passenger vehicles are several times more NO_x polluting when driven on normal roads compared to what the test procedure data and the emission standards suggest.[53] Work is ongoing to address and quantify this discrepancy because it is especially important for European urban areas with their high rates of dieselisation.

In the urban air quality domain, the disconnect between the progressively stringent vehicular emission standards and the lack of decreasing pollutant concentrations were highlighted well before the diesel emission scandal.[34,54] Authors had noted that during the time when passenger vehicle emissions were tightened from Euro 1 to Euro 6 between 1992–2014, the emission limits for NO_x as an example decreased by 92 % (Figure 1.5). However, during this period, ambient pollutant concentrations have decreased, but not even close to the magnitude of change enforced by the emission limits. Diesel passenger vehicle numbers increased during this period and

Figure 1.4: A holding pen for 21 000 Volkswagen AG vehicles embroiled in the diesel emission scandal at the Southern California Logistics Airport (from Worstall [46]).

there is a lag period before the old vehicles are replaced with new vehicles compliant to modern emission standards, but these factors did not explain the lack of concentration decrease observed. In hindsight, it is clear that there was a mismatch between vehicular emissions being reported during laboratory test cycles and those being generated by vehicles operating in their true environment on roads. The diesel emission scandal demonstrated that this disconnect is very real and that some manufacturers have deployed sophisticated and illegal subversion tactics which help explain why ambient pollutant concentrations have not decreased at the rates which could be expected and has led to significant air quality consequences.

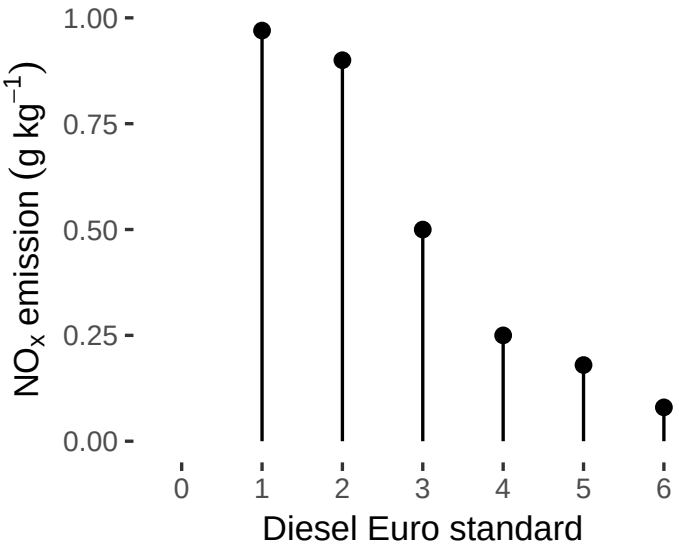

Figure 1.5: Euro NO_x emission standards for diesel powered passenger vehicles.

1.2 Key chemistry of vehicular NO_x emissions

NO_x is formed from the combustion of fuels due to the presence of high temperatures, nitrogen (N), and oxygen (O). For most combustion processes, the majority of NO_x is in the form of NO with a smaller component of NO_2.[55,56] NO formation by combustion processes is described by the ex-

tended Zeldovich mechanism where dioxygen (O_2) and dinitrogen (N_2) are broken and react to form NO (Equation 1.1):

$$\begin{aligned} N_2 + O &\rightleftharpoons NO + N \\ N + O_2 &\rightleftharpoons NO + O \\ N + OH &\rightleftharpoons NO + H \end{aligned} \tag{1.1}$$

For the combustion of hydrocarbon fuels used in road vehicles, there is no significant NO_x sourced from nitrogen containing compounds in the fuel itself.[57]

The mechanisms determining engine-out NO_x emissions for internal combustion engines are excess oxygen and temperature. Diesel engines emit more NO_x because of their lean combustion cycle which creates combustion environments with excess oxygen and higher peak flame temperatures when compared to (usually) stoichiometric gasoline (petrol) engines.[58,59] Gasoline engines emit very little if any NO_2 but diesel engines can emit a significant amount of NO_2 in absolute terms, and their NO_2/NO_x ratio is higher.[58] NO_2 formation within a combustion chamber is described by Equation 1.2:

$$NO + HO_2 \rightarrow NO_2 + OH \tag{1.2}$$

The NO_2 conversion back to NO is shown in Equation 1.3:

$$NO_2 + O \rightarrow NO + O_2 \tag{1.3}$$

In the presence of excess oxygen, the NO_2/NO_x ratio in a diesel engine is driven primarily by temperature with lower temperatures generally resulting in higher NO_2/NO_x ratios while higher temperatures result in lower NO_2/NO_x ratios.[56] Therefore, the highest NO_2/NO_x ratios tend to favour low load situations such as idling and slow driving cycles such as those encountered in congested traffic conditions.[58] Modern automotive diesel engines with their very lean operation cycles and (usually) forced induction

systems such as turbochargers also generally result in higher NO_2/NO_x ratios when compared to the older, previous generation diesel engines.[60]

1.2.1 Diesel after-treatment technology

Modern diesel after-treatment technology further complicates the NO_2/ NO_x ratio emitted out the tailpipe. Diesel oxidation catalysts (DOC) can significantly alter the ratio by oxidising some NO to NO_2.[61] However, the reduction of NO_2 to NO does occur which is rather unintuitive but DOC are specifically designed to treat other species such as CO and hydrocarbons.

Diesel particulate filters (DPF) use NO_2 as an oxidant to control particulate/soot emissions because NO_2 is a more effective particulate oxidant at lower temperatures compared to oxygen. To burn particulate in diesel exhaust, temperatures greater than $\approx$250–300 °C are required and such temperatures are not commonly found in diesel exhaust. DPFs are installed downstream of DOCs to allow the systems to work together and use NO_2 as a particulate oxidant.[62] If DPFs are poorly optimised, not all NO_2 is consumed in this process and increases the NO_2/NO_x ratio emitted from the tailpipe.

The two principal after treatment technologies to reduce NO_x emissions to comply with the latest Euro 6 emission standards are lean NO_x traps (LNT; also known as NO_x adsorbers) and selective catalytic reduction (SCR) devices. LNTs operate by storing NO_x in an adsorbent, commonly barium oxide (BaO) and periodically switching to a "regeneration" cycle every few minutes where NO_x is reduced to elemental nitrogen under a rich burning phase which lasts a few seconds.[64] SCR devices also reduce NO_x, but with the use of reducing agents, usually urea sourced from a diesel exhaust fluid, better known by its commercial name AdBlue.[65]

The NO_x reduction phases for both of these after-treatment strategies operate more rapidly with higher NO_2/NO_x ratios than lean burn diesel engines provide directly from the combustion chamber. Therefore, DOCs usually based on platinum (Pt) are installed upstream to oxidise a fraction

of NO to NO_2 before being exposed to the reduction step.[65,66] In the same way as particulate control can cause higher tailpipe NO_2/NO_x ratios due to poor optimisation, the same issue can result from modern NO_x after-treatment technology resulting in a greater fraction of NO_2 being emitted into the roadside atmosphere which can have air quality consequences.

1.3 Measurement of ambient NO_x

The vast majority of instruments used for the measurement of ambient NO_x exploit the chemiluminescence principle.[67,68] Chemiluminescence is defined as the release of electromagnetic radiation (light) from a chemical reaction. A chemiluminescence NO_x gas analyser detects NO concentration after mixing a gas sample with excess O_3 from an on-board generator within a reaction chamber (Equation 1.4 and Figure 1.6).

$$\begin{aligned} NO + O_3 &\rightarrow NO_2^* + O_2 \\ NO_2^* &\rightarrow NO_2 + h\nu \end{aligned} \tag{1.4}$$

The NO + O_3 reaction generates excited NO_2 molecules which emit photons which are in turn detected by a photomultiplier tube and the concentration of NO can be quantified.

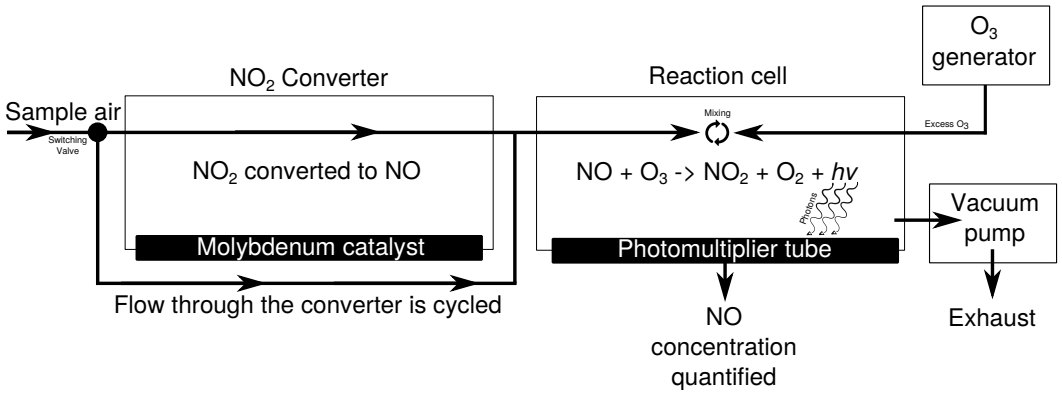

Figure 1.6: Conceptual diagram of a chemiluminescent NO_x analyser.

This measurement principle does not allow for the direct detection of NO_2 and leads to an unusual situation where despite NO_2 being a pollutant

with legal limits, it is indirectly quantified with the standard instrumentation. To indirectly detect NO_2, a converter is used upstream of the reaction cell within the analyser (Figure 1.6). Such converters use heated molybdenum catalysts which reduce NO_2 to NO before being introduced into the reaction cell and results in a measurement of total NO_x. A measurement cycle is used where sample air is directed through the the NO_2 converter and a second cycle (or mode) where the converter is bypassed (Figure 1.6). The separate and cyclical measurement of NO and NO_x allows for calculation of NO_2 (NO_x minus NO).

Chemiluminescence NO_x analysers can be subjected to significant measurement uncertainty, primarily due to less than 100 % NO_2 conversion efficiency and errors resulting from negative and positive interfering species other than NO and NO_2.[69–71] Compounds within the NO_y family, including nitric acid (HNO_3), nitrous acid (HNO_2/HONO), and peroxyacetyl nitrate (PAN) are positive interferents in chemiluminescence NO_x analysers and therefore are detected as NO_2.[68,72] With the exception of HONO, these interfering species are much more important in rural locations where air masses have been exposed to chemical processes and ageing when compared to roadside environments.[68] Combustion is a source of HONO however, and this compound is detected as NO_2 by chemiluminescence analysers.[73] The amount of HONO in the roadside atmosphere is poorly quantified, but it is thought to be small and therefore will this interference is very unlikely to influence results presented in this thesis (especially those described in Chapter 4). Despite the known issues of chemiluminescence NO_x analysers, they are still widely used and installed, but newer technologies which are not as sensitive to interferences such as blue light (photolytic) converters are becoming increasingly popular.[74]

1.4 Air quality monitoring data

To underpin air quality knowledge and research, continuous monitoring networks began to commissioned in the mid-twentieth century. At first, such networks were crude and labour intensive. As technology improved, the instrumentation and monitoring activities became progressively more sophisticated and automated. By the 1980s, instrumentation had reached an appropriate level of maturity for common gaseous pollutants and the monitoring of outdoor air became a legal requirement in many jurisdictions for CO, NO_x, SO_2, and O_3. PM monitoring required further instrument development than these common gases and continues to evolve today due to the heterogeneous nature of PM and the vast range of sources.[75]

Air quality has grown into a data intensive domain. With most counties operating air quality monitoring networks to accurately determine concentrations of pollutants within their boundaries, the observational record has continually grown since instruments became available and the amount of data collected is now vast and continues to increase.[76] In the last ten years, standardised reporting has been mandatory for a number of federal and super-governmental bodies. This allows for a common infrastructure to release and share air quality data. For European Union (EU) member states and other cooperating countries or areas, the older AirBase system and the current Air Quality e-Reporting (AQER) systems are publicly accessible.[77,78]

Despite the wealth of data collected and the efforts required to gain air quality observations, they are generally only used to generate simple summaries such as annual means to determine compliance to legal limits.[79] Such analyses do not use these data to their potential and does little to understand the physical and chemical processes giving rise to elevated concentrations nor understand the characteristics of air pollution sources.

Within this "routine" air quality observational record there is potential for new insights. Conducting careful and novel data analysis on these data

sets could give insight into what gives rise to poor air quality and how pollutant sources operate. This is the main focus of this thesis, to analyse routine air quality data to gain additional and novel insight into physical and chemical atmospheric processes. The advantages of leveraging such data is that there is large amount of it and the measurements available are standardised allowing for much easier comparisons among multiple locations and times when compared with data collected during specialised and/or during short term campaigns.

The atmosphere is an extremely complicated system.[80] The complication of the atmosphere makes probing the system very challenging in most situations. In respect to urban air quality, a combination of source activity, chemistry, and meteorological controls all play a role in determining the concentrations of air pollutants. These families of processes also interact with one another leading to a system being a function of competing processes, all with differing dominance at different times and states.

Owing to the atmosphere's complexity, changes in pollutant concentrations are often nuanced and take time to move to a different state. It is rare to see pollutant concentrations abruptly move from one regime to another.[81] This gives rise to frustration when source activities change because the effect of the change are often obscured by other atmospheric processes which results in questions around the efficacy of intervention and management efforts. Therefore, questions such as "has a low emission zone improved air quality" can be very difficult to answer.

Questions concerning the behaviour of such source activity can be asked of routine monitoring data. The challenge of such analyses is how to extract new and interesting things related to source characteristics from the observations. Currently, the world is experiencing major growth in all things data related with the utilisation of data-focused programming languages, "big data", and machine learning all becoming important components in a diverse range of fields and domains.

To further our understanding about why and how poor air quality is

experienced, more in depth data analysis is required, preferably with common and open methods and tools. Projects such **openair** have developed and have resulted in the release of a set of free, open source, and cross platform tools accessible to anyone with a modern computer.[79,82] The **openair** project focuses on giving data users an integrated tool set to help with air quality data analysis which has significant uptake around the world. **openair** has many functions for particular types of data visualisation common in the atmospheric sciences such as wind roses, bivariate polar plots, time series plots, and trend plots as well as utilities such as a flexible aggregator allowing data to be aggregated to different time resolutions.

An important component in any data analysis activity is importing or loading data.[83] **openair** contains importing functions which fetch up to date air quality observations from web servers for many British monitoring locations. An **openair** companion package named **worldmet** accesses NOAA's Integrated Surface Database (ISD) which imports meteorological observations from the worldwide ISD network.[84,85] Meteorological observations are often necessary to have to analyse air quality data in a meaningful way. Other projects such as OpenAQ are becoming prominent because they act as aggregators for data and offer a framework to enable consistent interaction with a heterogeneous collection of data sources.[86]

The provision and release of data to the public is not the only component to allow for usability with file formats and file structures also requiring substantial thought. The modern system used for European air quality data transmission (AQER) uses Extensible Markup Language (XML)[87] as the file format. XML is a common format for data transmission across the web, but in the case of AQER the schema complication seems unnecessary with deeply nested data structures being encountered. Data analysis activities generally require two dimensional tabular objects which the nested XML elements must be formatted into before analysis. Programmatically, this can pose a substantial challenge and would stop many data users being able to leverage these data. To address this issue in part, the **smonitor** Europe

database was developed (Chapter 3) to allow the research questions to be answered.

1.5 What this research will contribute

This thesis will present a number of linked analyses with the overarching theme of extracting new and novel insights from routine air quality monitoring data. New techniques were researched, applied, and the tools, generally software, were distributed to others.

1.5.1 Objectives

The primary objective of this thesis is to extract new and novel information from routine ambient air quality monitoring data. Additional information to aid understanding of emission source behaviour and atmospheric physical and chemical processes on transportation activities is the main area of focus with the ultimate aim to help with effective air quality management. The tools which result from these developments are developed in a manner which are distributable to other data users. To demonstrate the developments, four separate, but related case studies will be presented. These developments are (*i*) enhancements to a particular type of data visualisation called bivariate polar plots with weighted pair-wise statistics to aid source apportionment, (*ii*) the development of an air quality monitoring database with a formal data model called **smonitor** Europe to enable a scalable and convenient way to manage the European air quality observations and their metadata, (*iii*) a trend analysis of European primary (directly emitted) vehicular NO_2 and the implications for future compliance to the European ambient NO_2 limits, and (*iv*) the presentation of a meteorological normalisation framework to remove the variability of pollutant concentrations due to changes in weather in an air quality time series. The meteorological normalisation framework is presented in two components because the two case studies have different objectives. The first component is a formal trend anal-

ysis for PM_{10} concentrations across Switzerland and the second component is the exploration of interventions expected to impact pollutant concentrations in two locations in the United Kingdom. As a side effect of the meteorological normalisation case studies, the **rmweather** R package which applies the technique is introduced. All case studies have the common theme of investigating source characteristics relating to transportation activities and extracting additional information from routine air quality monitoring data.

1.5.2 Bivariate polar plots with pair-wise statistics

Bivariate polar plots are a popular technique in air quality data analysis used for source apportionment.[88,89] In their most common and simplest form, pollutant concentrations are aggregated into wind speed and wind direction bins and are displayed on polar coordinates. Because of the patchiness of wind speed and directions experienced by most monitoring sites, the surface is usually modelled to give some amount of interpolation which results in a more aesthetically pleasing plot, for example, Figure 1.7. Although it is intuitive to plot wind speed on the radial scale (moving from the centre point outwards), any variable can be used and the use of other surface meteorological variables such as ambient temperature can be very effective at illuminating source processes.[79]

1.5.2.1 Pair-wise statistics

An extension to bivariate polar plots involves the use of pair-wise statistics. Pair-wise statistics is a broad term for a statistical value which represents two quantities in a singular fashion. Pearson correlation coefficients (r) and slopes from simple least squared regression models (usually denoted as m or β) are examples of pair-wise statistics. High levels of correlation between two air pollutants can often give insight to the pollutants' sources. For example, metals such as vanadium (V) and nickel (Ni) are emitted together

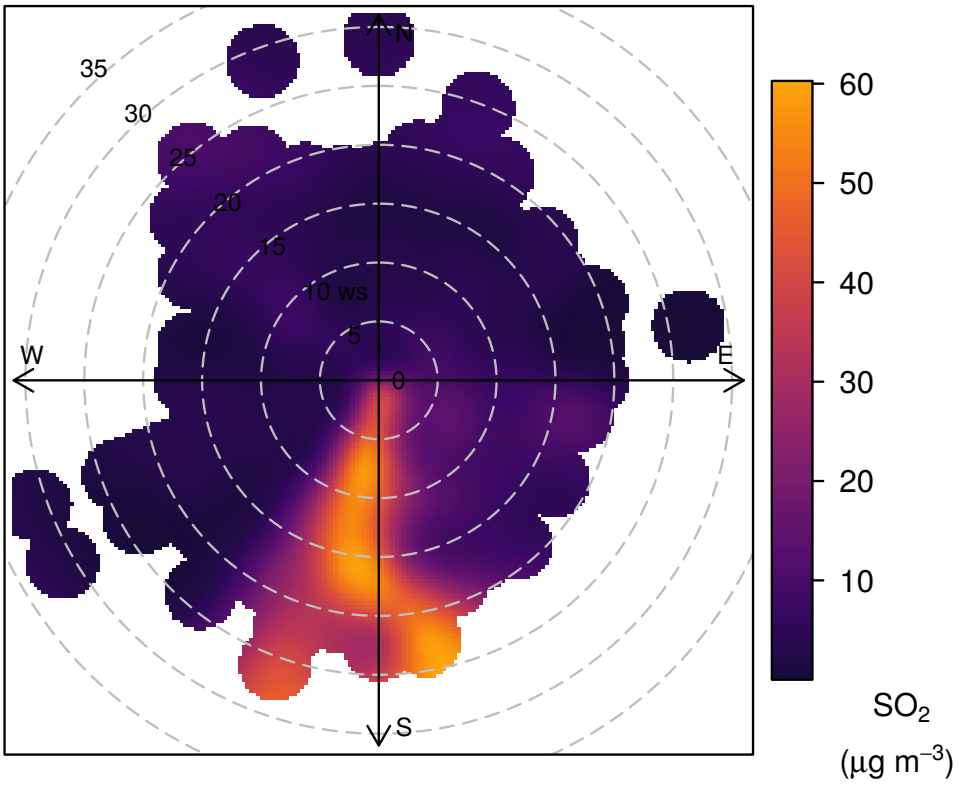

Figure 1.7: A bivariate polar plot displaying mean SO_2 concentrations between 2001 and 2010 at Dover Landon Cliff. South of the monitoring site is the Port of Dover complex, an area with significant SO_2 emissions resulting from the combustion of marine fuels.

in certain ratios when certain types of heavy fuel oils are burnt and show high levels of correlation[90–92] and the correlation between different particulate matter (PM) size fractions can indicate emission sources, especially when investigating the division between natural and anthropogenic processes.[93,94]

Perhaps more useful than correlation for source apportionment is the slope between two pollutants. Using the slope between two pollutants or compounds is a very common analysis procedure used in source appointment,[95] but when used with a bivariate polar plot, the sources can be further disaggregated by wind direction which is a powerful technique when conducting exploratory data analysis (EDA).

Chapter 2 presents an analysis using these tools. The enhancements to the `polarPlot` function in **openair** has been implemented and has been available in the package since November 2016[†]. Since publication, the approach has been used and reported by others.[96]

1.5.3 Data

For productive data analysis activities, data users require performant access to standardised data. If this process is done well, the standardised data underpins high quality analysis activities and complex questions can be asked of the data without having to repeatedly handle data formatting issues due to a common framework imposed on the data. To this end, the **smonitor** Europe database was designed and commissioned to aid much of the research presented in this thesis.[97,98]

smonitor Europe primarily contains European air quality time series data and a collection metadata units to support the observations in a useful way. The **smonitor** relational data model was developed from the ground up with an emphasis on usability.[97] The primary development focus was to ensure data can be queried from the database as quickly and as easily as possible. Other options do exist, most notably the 52°North implementa-

[†]**openair** git commit 64db36d

tion of a Sensor Observation Service (SOS), but this was not used due to its unneeded complication for this application.[99–101] **smonitor** also contains other components to allow for easy updating, invalidating, and the calculation of aggregations, but in the case of **smonitor** Europe, these are not used because it only serves a data storage function. The database technology used is PostgreSQL (version 9.5).[102] The **smonitor** data model has proven to be highly scalable with the current disk size being 413 Gb containing 12 800 sites† and 4.1×10^9 observations while still maintaining good performance.

The primary data sources for **smonitor** Europe are the data repositories maintained by the European Environment Agency (EEA). For data between 1969 and 2012 (inclusive), the AirBase repository was used.[77,103] For the 2013 and beyond reporting years, a new data reporting system called Air Quality e-Reporting (AQER) for the compliance to the 2004/107/EC and 2008/50/EC air quality directives was implemented and was used.[78] The AirBase data are supplied as formatted text files which were easy to accommodate into the **smonitor** framework. However, the AQER was a far greater challenge because the self describing XML file format is used which required far more programmatic development to parse and format these data files. The XML documents submitted consist of separate "observational units" which requires the joining of many documents to one another to make the files usable. Additionally, because AQER is a relatively new system, many member states do this process poorly and errors are commonly encountered.

Since the initial development of **smonitor** Europe, the EEA have addressed some of these issues and now have an interactive data portal to access member states' observations.[104] There is however a significant time lag from when documents are submitted to when they become available and the full data sets are not reported which results in using the submitted XML still being the best option to get these data currently.

The final primary data source used is NOAA's Integrated Surface Database

† **smonitor** Europe interactive site map

(ISD).[84] Unlike the EEA's data repositories, the ISD is not concerned with air quality observations, rather surface meteorological/weather data. For many air quality analyses, high quality metrological data is necessary too. The ISD's scope is worldwide so for sites within a rough European boundary, data were retrieved and inserted into **smonitor** Europe. An additional 100 monitoring sites have been inserted into **smonitor** Europe from other data sources including the Centre for Environmental Data Analysis (CEDA),[105] EBAS,[106] and the World Data Centre for Greenhouse Gases (WDCGG).[107] This development allows easy access to European time series with the same functions and framework enabling high quality data analysis to occur without repeatably dealing with tedious data formatting issues. **smonitor** Europe has also seen use by others in their research, for example Hu et al. [108]. Chapter 3 presents a technical note on **smonitor** Europe outlining the technical development choices and how the database is structured and used.

1.5.4 European vehicular primary NO_2 trends

Roadside environments in Europe remain polluted with NO_x and many EU member states are non-compliant to the legal ambient air quality limits.[37,109] Many European urban areas have experienced a somewhat counter-intuitive situation where NO_x concentrations have generally decreased while NO_2, a component of NO_x, has not decreased at the same rate, and in some situations has increased (Figure 1.8).[60,110–112] The vehicular emission standards in Europe exclusively prescribe limits for NO_x with no regard to NO_x's components.[113] However, air quality standards only exist for NO_2 and highlights a disconnect between vehicle emission control and local air quality management.

This disconnect has given rise to a situation where vehicular NO_x emissions have decreased, but NO_2 emissions have not been reduced at the same rate, *i.e.* the amount of directly emitted (usually called "primary") NO_2 has changed.[114–116] This has been attributed to changes in the composition of

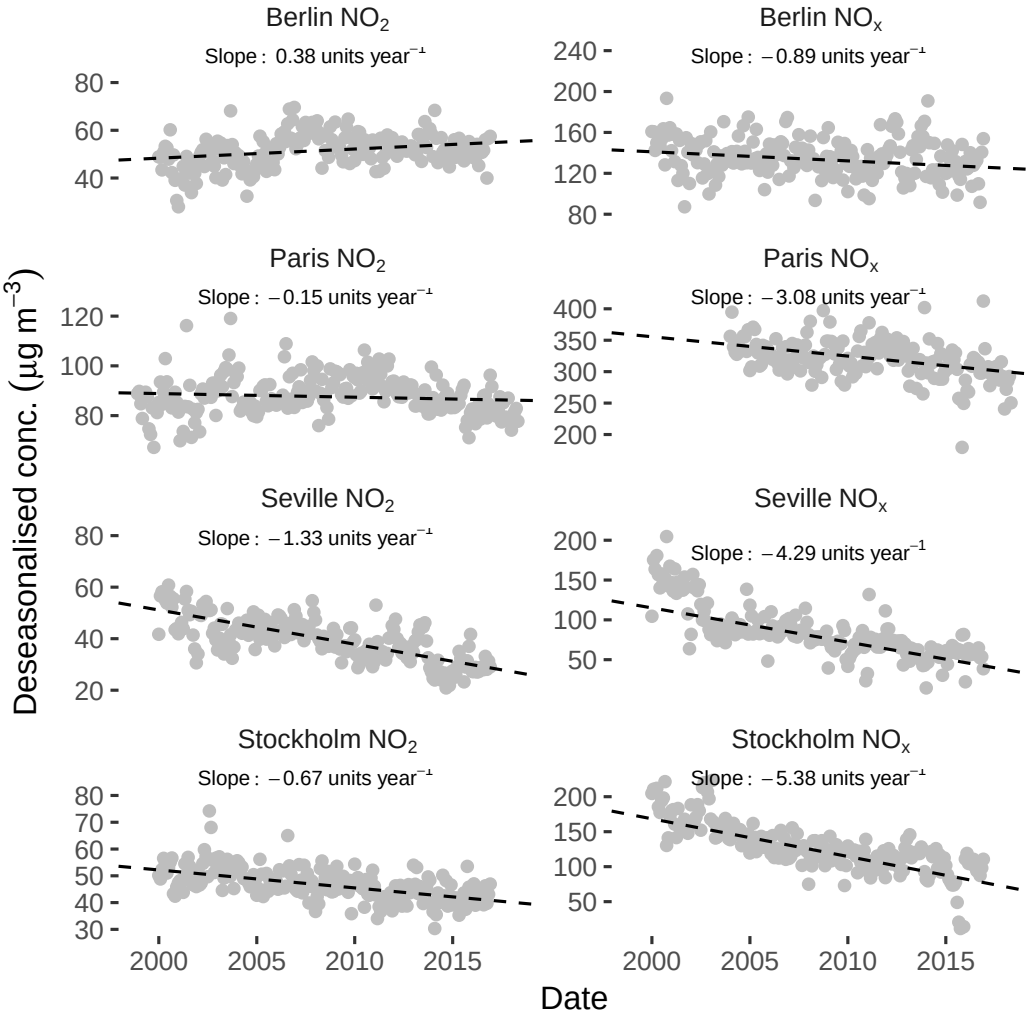

Figure 1.8: Deseasonalised NO_x and NO_2 trends between 2000 and 2016 for four European cities experiencing different climates. NO_2 has constantly reduced at a slower rate than NO_x (and in the case of Berlin, has increased), indicating composition changes of total NO_x throughout Europe.

the European passenger vehicle fleet. The European passenger vehicle fleet has undergone significant dieselisation since ≈1990 and critically, the control of NO_x emissions for diesel engines is much more difficult compared to petrol powered vehicles.[62] The main reasons for this are that diesel engines commonly use lean air-fuel ratio combustion conditions and the inability for three-way catalysts to operate in oxygen rich exhaust environments. Diesel vehicles equipped with diesel particulate filters (DPF) for PM emission control often use NO_2 as an oxidant for soot removal and some untransformed NO_2 can also be directly emitted.[62,117,118]

The issue around changes in primary NO_2 emissions is a subtle one and has only been reported sparingly.[119] Despite this, there is evidence that the importance of primary NO_2 emissions on roadside NO_2 concentrations have decreased since ≈2010.[62] This has been explained by the lower emission limits imposed by the newer Euro standards, degradation of diesel oxidation catalysts (DOC) and perhaps DPF, and a process so-called "catalyst thrifting" where manufacturers and original equipment manufacturers (OEMs) have reduced the amount of platinum group metals within the catalyst washcoat due to concerns regarding costs.

Chapter 4 outlines a pan-European trend analysis of primary NO_2 between 1990 and 2016 using **smonitor** Europe (Section 1.5.3 and Chapter 3) as a data source. Since publication, other studies have confirmed that the trends presented in our European scale analysis are constant in their local environments.[120,121]

1.5.5 Robust trend and intervention exploration

Trend analysis and the exploration of interventions using air quality data are very common procedures.[122] Unfortunately, a myriad of processes resulting from the complexity of the atmosphere can exacerbate or obscure changes in emission sources when only concentration values are used for these types of analyses.[123] For trend analysis and intervention exploration to be conducted robustly, changes in meteorology/weather should be con-

trolled or accounted for over the period of analysis.[124–128]

To address this, a framework to control for meteorology in an air quality time series called *meteorological normalisation* was developed and the functionality demonstrated. The chief philosophy of meteorological normalisation is to reduce variation of pollutant concentrations by statistical modelling. To achieve this, random forest,[129–131] an ensemble decision tree machine learning technique is used to build a predictive model and it is used repeatably to predict a sampled and simulated observational record which result in a time series representing concentrations in "average weather". This prepared time series can then be exposed to formal trend analysis and/or other EDA techniques.

Random forest was the algorithm chosen for this procedure because of its useful attributes: it is fast to train and predict, is conceptually simple, it is resistant to overfitting, is not extremely sensitive to its hyperparameters, is non-parametric, can handle interaction and collinearity among explanatory variables, and it is not a black box technique which allows the learning process to be evaluated.[132–138] These advantages coalesce to form a technique which can be used for a wind range of inputs without specialist machine learning or statistical modelling knowledge, hence, functions can be developed to be user friendly, flexible, and stable. There were three components to this research (*i*), trend analysis of PM_{10} data across Switzerland, (*ii*) an analysis of known air quality interventions related to transportation activities and their impact on air quality, and (*iii*) the development and release of a tool to conduct analyses such as these in the form of an R package called **rmweather**.

1.5.5.1 Swiss PM_{10}

The Swiss federal air quality monitoring network (NABEL) has a very high quality observational record of daily PM_{10} collected by gravimetric sampling techniques.[139,140] Switzerland is not within the EU and therefore has different air quality management plans and regulations when compared to

other European countries, however, the differences are generally subtle.[141] To test the effectiveness of local PM emission control and wider European reductions of PM precursors, 31 PM_{10} sites were analysed with the meteorological normalisation technique across Switzerland with urban traffic, urban background, suburban, rural motorway, rural, and rural mountain site type classifications. After the meteorological normalisation technique had been applied, a formal trend analysis was conducted.

In all but two Swiss monitoring sites, PM_{10} concentrations were found to be significantly decreasing between 1997 and 2016 at rates between -0.09 and -1.16 $\mu g\,m^{-3}$ $year^{-1}$. Suburban monitoring sites experienced less of a decrease than expected which was speculated to be a result of rapid urbanisation in many Swiss locations with many suburban monitoring sites becoming "more urban" in their characteristics during the period of analysis. The legibility of the random forest models allowed physical and chemical atmospheric processes to be illuminated. Most interesting of which was that high temperatures, high boundary layer heights can still result in elevated PM_{10} concentrations due to high rates of secondary generation of PM. Chapter 5 describes and presents this research. Like other research projects presented in this thesis, **smonitor** Europe was used as the data source (Section 1.5.3 and Chapter 3).

1.5.5.2 Intervention exploration

When interventions or management activities are conducted in an attempt to improve air quality, it is very common to use ambient air quality data to attempt to find and quantify the effect of the interventions. However, effects can be difficult to observe in time series resulting from the atmosphere's complexity and where changes usually occur progressively and in a nuanced fashion. This often leads to ambiguous conclusions on the efficacy of air quality management.[28,29,142] The meteorological normalisation technique offers a tool to reduce variability in an air quality time series so changes in concentrations and therefore emissions can be more easily ex-

posed.

In Chapter 6, two well known interventions which were expected to result in improvements to air quality were explored in ambient air quality data. The first intervention was the progressive reductions in the allowed sulfur content in marine fuels and its effect on Dover's (a port city in the South East of England) SO_2's monitoring data.[143] In August 2006, regulations were introduced for EU ports which applied a 1.5% sulfur content limit for marine fuel oil and in 2010, another limit was imposed of 1% for berthed vessels.[144] The estimated change in sulfur fuel content for 2006 was 44%[145] and the SO_2 time series showed a 45% change in pre- and post-intervention concentrations after the meteorological normalisation technique had been applied. This near exact representation of the intervention was reflected because SO_2 was sourced almost exclusively from the port's activities for the two monitoring sites investigated.

London is one of Europe's largest cities and has a significant roadside NO_x and NO_2 issue.[26] Some of London's roadside monitoring sites report the highest NO_2 concentrations in Europe and in some cases being non-compliant to the EU's hourly NO_2 standards within the first week of the year.[146–149] To combat traffic congestion, a number of transport interventions have been applied in London such as the Inner London Congestion Charge Zone (CCZ) in 2003, the Greater London Low Emission Zone (in 2008 and made more stringent in 2012), and the T-Charge (2017) with other management processes planned.[31] These events were also expected to improve London's roadside NO_2 concentrations due to reducing the volume of traffic and the incentivisation of modern vehicles.

London Marylebone Road is a monitoring site located on the A501 road which marks the northern boundary if the CCZ. London Marylebone Road is a prominent site due to the long monitoring record, a number of specialised pollutants are monitored, generally high pollutant concentrations, and it has a complicated irregular street canyon siting.[150–152]

When London Marylebone Road's NO_x and NO_2 time series was exposed

to the meteorological normalisation technique, it was very clear that NO_x and NO_2 were not behaving in the same way. Normalised NO_x concentrations, and therefore emissions have remained static since the implementation of the CCZ in 2003 to the end of the analysis period (2016). These constant NO_x emissions were present even with the progressively stringent vehicular emission control being applied across Europe during the period of analysis. However, in the case of NO_2 emissions, they increased when the CCZ was introduced.

Coinciding with the introduction of the CCZ, was the retrofitting of many of the buses which service the local routes with continuously regenerating diesel particulate filters (CRDPF or also known by their commercial name CRT filters).[153] These CRDPF devices oxidised NO to NO_2 within the exhaust stream for particulate matter emission control. However, these devices did not reduce all NO_2 after being oxidised resulting in NO_2 being directly emitted into the roadside atmosphere, but without altering the amount of total NO_x emissions. When these retrofitted Euro III buses were progressively replaced with fleets of Euro IV and V vehicles, the NO_2 concentrations reduced and at the end of 2016, NO_2 emissions were near pre-CCZ levels.

These observations link the European analysis of primary NO_2 reported in Section 1.5.4 and Chapter 4. Although the particular processes responsible for the NO_2 emissions observed in London Marylebone Road are unlikely to be the same for other European urban areas, there is consistency between the different analyses. "Zooming" into the problem documented at a European scale to an individual monitoring site demonstrates important implications for air quality management with only the use of routine monitoring data and applications of novel data analysis tools.

1.5.5.3 rmweather R package

The meteorological normalisation technique requires a large amount of programming logic and careful treatment of input data sets to be reliable and

stable. To ensure the meteorological normalisation tools could be distributed, accessed by, and contributed to by other data users, the **rmweather** R package[154] was developed alongside the Swiss PM_{10} trend (Chapter 5) and British intervention analyses (Chapter 6). **rmweather** was first accepted onto CRAN (The Comprehensive R Archive Network) in May 2018,[155] R's official code repository which demonstrates the development has met rigorous quality control and can be installed by any R user easily with `install.packages("rmweather")`. **rmweather** depends on other user contributed packages, most notably **ranger** (short for RANdom forest GEneRator) to gain access to a performant C++, multithreaded, and cross platform random forest algorithm.[156] The development of **rmweather** fits within into R's modern "tidy data" analysis framework.[83,157]

1.6 Structure of thesis

This thesis uses four articles as stand alone chapters and the details are presented in the thesis's declaration. The contents of the chapters are faithful to the accepted and published articles, however minor formatting changes have been made to section headings, figures and captions, citations and reference styles, and the inclusion of supplementary material for the sake of completeness and consistency. Every chapter contains its own bibliography to ensure the chapters form coherent units without the need for additional document sections. There are minor inconsistencies among the different chapters regarding counts of observations and sites used for the different analyses. This has arisen due to the nature of working with time series data where the data available continuously grows and the analyses being conducted at different points during this growth.

Chapter 2 presents an R package called **polarplotr**.[158] Soon after publication, this package was made redundant because the functionality was incorporated into the established **openair** package. Similarly, Chapter 5 references an R package called **normalweatherr** which was developed to

perform the analysis presented.[159] The **normalweatherr** package can be considered the Mk I version of the published **rmweather** package and although is still accessible, it has been deprecated in favour of **rmweather**.

1.7 References

[1] Fenger, J. Urban air quality. *Atmospheric Environment* 33.29 (1999), pp. 4877–4900. DOI: 10.1016/S1352-2310(99)00290-3. URL: http://www.sciencedirect.com/science/article/pii/S1352231099002903.

[2] Fenger, J. Air pollution in the last 50 years - From local to global. *Atmospheric Environment* 43.1 (2009), pp. 13–22. DOI: 10.1016/j.atmosenv.2008.09.061. URL: http://www.sciencedirect.com/science/article/B6VH3-4TNWH49-8/2/abef8a2a473d977523bdab2a9c548512.

[3] World Health Organization. Ambient air pollution: A global assessment of exposure and burden of disease. 2016. URL: http://www.who.int/iris/bitstream/10665/250141/1/9789241511353-eng.pdf?ua=1.

[4] The World Bank. The Cost of Air Pollution: strengthening the Economic Case for Action. The World Bank and Institute for Health Metrics and Evaluation. University of Washington, Seattle. 2016. URL: http://documents.worldbank.org/curated/en/781521473177013155/The-cost-of-air-pollution-strengthening-the-economic-case-for-action.

[5] Ritchie, H. and Roser, M. Causes of Death. Published online at OurWorldInData.org. 2018. URL: https://ourworldindata.org/causes-of-death.

[6] Pimpin, L., Retat, L., Fecht, D., Preux, L. de, Sassi, F., Gulliver, J., Belloni, A., Ferguson, B., Corbould, E., and Jaccard, A. Estimating the costs of air pollution to the National Health Service and social care: An assessment and forecast up to 2035. *PLoS medicine* 15.7 (2018), e1002602. URL: https://doi.org/10.1371/journal.pmed.1002602.

[7] Organisation for Economic Co-operation and Development (OECD). The economic consequences of outdoor air pollution: policy highlights. 2016. URL: http://www.oecd.org/env/the-economic-consequences-of-outdoor-air-pollution-9789264257474-en.htm.

[8] United States Environmental Protection Agency. Benefits Mapping and Analysis Program (BenMAP): How BenMAP-CE Estimates the Health and Economic Effects of Air Pollution. 2018. URL: https://www.epa.gov/benmap/

how-benmap-ce-estimates-health-and-economic-effects-air-pollution.

[9] World Health Organisation. Fact-sheet: Ambient (outdoor) air quality and health. 2 May 2018. 2018. URL: http://www.who.int/en/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health.

[10] Gulia, S., Shiva Nagendra, S. M., Khare, M., and Khanna, I. Urban air quality management-A review. *Atmospheric Pollution Research* 6.2 (2015), pp. 286–304. URL: http://www.sciencedirect.com/science/article/pii/S1309104215302373.

[11] The World Bank. Urban population (% of total). United Nations Population Division. World Urbanization Prospects: 2014 Revision. 2018. URL: https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS.

[12] Gurjar, B., Butler, T., Lawrence, M., and Lelieveld, J. Evaluation of emissions and air quality in megacities. *Atmospheric Environment* 42.7 (2008), pp. 1593–1606. DOI: 10.1016/j.atmosenv.2007.10.048. URL: http://www.sciencedirect.com/science/article/B6VH3-4R1KVWY-3/2/0320e6d16c1a0fe19e5103bf2e3b7740.

[13] Mayer, H. Air pollution in cities. *Atmospheric Environment* 33.24-25 (1999), pp. 4029–4037. DOI: 10.1016/S1352-2310(99)00144-2. URL: http://www.sciencedirect.com/science/article/B6VH3-3X114HN-K/2/854a71aeec808868db672f793014db69.

[14] Whitty, C. Air Pollution: Its Impact on Health and Possible Solutions. Gresham College. 2018. URL: https://www.gresham.ac.uk/lectures-and-events/air-pollution-its-impact-on-health-and-possible-solutions.

[15] Mage, D., Ozolins, G., Peterson, P., Webster, A., Orthofer, R., Vandeweerd, V., and Gwynne, M. Urban air pollution in megacities of the world. *Atmospheric Environment* 30.5 (1996), pp. 681–686. DOI: 10.1016/1352-2310(95)00219-7. URL: http://www.sciencedirect.com/science/article/B6VH3-3WBXSB7-20/2/a1baba4a3cf62af08ffab2aee17569a0.

[16] Cohen, A. J., Ross Anderson, H., Ostro, B., Pandey, K. D., Krzyzanowski, M., Künzli, N., Gutschmidt, K., Pope, A., Romieu, I., Samet, J. M., and Smith, K. The Global Burden of Disease Due to Outdoor Air Pollution. *Journal of Toxicology and Environmental Health, Part A* 68.13-14 (2005), pp. 1301–1307. DOI: 10.1080/15287390590936166. URL: https://doi.org/10.1080/15287390590936166.

[17] Nel, A. Air Pollution-Related Illness: Effects of Particles. *Science* 308.5723 (2005), p. 804. URL: http://science.sciencemag.org/content/308/5723/804.abstract.

[18] Curtis, L., Rea, W., Smith-Willis, P., Fenyves, E., and Pan, Y. Adverse health effects of outdoor air pollutants. *Environment International* 32.6 (2006), pp. 815–830. URL: http://www.sciencedirect.com/science/article/pii/S0160412006000444.

[19] Pope, C. A. and Dockery, D. W. Health Effects of Fine Particulate Air Pollution: Lines that Connect. *Journal of the Air & Waste Management Association* 56.6 (2006), pp. 709–742. DOI: 10.1080/10473289.2006.10464485. URL: https://doi.org/10.1080/10473289.2006.10464485.

[20] Kampa, M. and Castanas, E. Human health effects of air pollution. *Environmental Pollution* 151.2 (2008), pp. 362–367. DOI: 10.1016/j.envpol.2007.06.012. URL: http://www.sciencedirect.com/science/article/B6VB5-4P83KJH-1/2/6428f5cc65435308092fdad408e6ae92.

[21] Szidat, S., Prévôt, A. S. H., Sandradewi, J., Alfarra, M. R., Synal, H.-A., Wacker, L., and Baltensperger, U. Dominant impact of residential wood burning on particulate matter in Alpine valleys during winter. *Geophysical Research Letters* 34.5 (2007), p. L05820. URL: http://dx.doi.org/10.1029/2006GL028325.

[22] Herich, H. and Hueglin, C. Residential Wood Burning: A Major Source of Fine Particulate Matter in Alpine Valleys in Central Europe. *The Handbook of Environmental Chemistry*. Springer Berlin Heidelberg, 2013, pp. 1–18. URL: http://dx.doi.org/10.1007/698_2013_229.

[23] Fountoukis, C., Butler, T., Lawrence, M., Denier van der Gon, H., Visschedijk, A., Charalampidis, P., Pilinis, C., and Pandis, S. Impacts of controlling biomass burning emissions on wintertime carbonaceous aerosol in Europe. *Atmospheric Environment* 87.0 (2014), pp. 175–182. URL: http://www.sciencedirect.com/science/article/pii/S1352231014000259.

[24] Fuller, G. W., Tremper, A. H., Baker, T. D., Yttri, K. E., and Butterfield, D. Contribution of wood burning to PM_{10} in London. *Atmospheric Environment* 87.0 (2014), pp. 87–94. URL: http://www.sciencedirect.com/science/article/pii/S1352231013009825.

[25] Beddows, D. C. S. and Harrison, R. M. Identification of specific sources of airborne particles emitted from within a complex industrial (steelworks) site. *Atmospheric Environment* 183 (2018), pp. 122–134. URL: http://www.sciencedirect.com/science/article/pii/S1352231018302176.

[26] Carslaw, D. C. and Beevers, S. D. The efficacy of low emission zones in central London as a means of reducing nitrogen dioxide concentrations. *Transportation Research Part D: Transport and Environment* 7.1 (2002), pp. 49–64. DOI: 10.1016/S1361-9209(01)00008-6. URL: http://www.sciencedirect.com/science/article/pii/S1361920901000086.

[27] Lutz, M. and Rauterberg-Wulff, A. Ein Jahr Umweltzone Berlin: Wirkungsuntersuchungen. Ein Jahr Umweltzone Berlin: Wirkungsuntersuchungen. 2009. URL: http://www.berlin.de/sen/umwelt/luftqualitaet/en/luftreinhalteplan/umweltzone_allgemeines.shtml.

[28] Boogaard, H., Janssen, N. A., Fischer, P. H., Kos, G. P., Weijers, E. P., Cassee, F. R., Zee, S. C. van der, Hartog, J. J. de, Meliefste, K., Wang, M., Brunekreef, B., and Hoek, G. Impact of low emission zones and local traffic policies on ambient air pollution concentrations. *Science of The Total Environment* 435–436.0 (2012), pp. 132–140. DOI: 10.1016/j.scitotenv.2012.06.089. URL: http://www.sciencedirect.com/science/article/pii/S0048969712009229.

[29] Holman, C., Harrison, R., and Querol, X. Review of the efficacy of low emission zones to improve urban air quality in European cities. *Atmospheric En-*

vironment 111.0 (2015), pp. 161–169. URL: http://www.sciencedirect.com/science/article/pii/S1352231015300145.

[30] Nieuwenhuijsen, M. J. and Khreis, H. Car free cities: Pathway to healthy urban living. *Environment International* 94 (2016), pp. 251–262. URL: http://www.sciencedirect.com/science/article/pii/S0160412016302161.

[31] Transport for London. Driving. 2018. URL: https://tfl.gov.uk/modes/driving/.

[32] Cames, M. and Helmers, E. Critical evaluation of the European diesel car boom - global comparison, environmental effects and various national strategies. *Environmental Sciences Europe* 25.1 (2013), pp. 1–22. DOI: 10.1186/2190-4715-25-15. URL: http://dx.doi.org/10.1186/2190-4715-25-15.

[33] International Council on Clean Transportation Europe. European vehicle market statistics, 2017/2018. 2017. URL: https://www.theicct.org/publications/european-vehicle-market-statistics-20172018.

[34] Lewis, A. C., Carslaw, D. C., and Kelly, F. J. Diesel pollution long underreported. *Nature* 526 (2015), p. 195. URL: http://dx.doi.org/10.1038/526195c.

[35] Brand, C. Beyond 'Dieselgate': Implications of unaccounted and future air pollutant emissions and energy use for cars in the United Kingdom. *Energy Policy* 97 (2016), pp. 1–12. URL: http://www.sciencedirect.com/science/article/pii/S030142151630341X.

[36] Zachariadis, T. After 'dieselgate': Regulations or economic incentives for a successful environmental policy? *Atmospheric Environment* 138 (2016), pp. 1–3. URL: http://www.sciencedirect.com/science/article/pii/S1352231016303430.

[37] European Environment Agency. Air quality in Europe — 2017 report. EEA Report No 13/2017. 2017. URL: https://www.eea.europa.eu/publications/air-quality-in-europe-2017.

[38] European Environment Agency. Premature deaths attributable to $PM_{2.5}$, NO_2 and O_3 exposure in 41 European countries and the EU-28, 2014. 2017. URL: https://www.eea.europa.eu/highlights/improving-air-quality-in-european/premature-deaths-2014.

[39] Barrett, S. R. H., Speth, R. L., Eastham, S. D., Dedoussi, I. C., Ashok, A., Malina, R., and Keith, D. W. Impact of the Volkswagen emissions control defeat device on US public health. *Environmental Research Letters* 10.11 (2015), p. 114005. URL: http://stacks.iop.org/1748-9326/10/i=11/a=114005.

[40] Gewin, V. Turning point: Daniel Carder. *Nature* 527.7578 (2015), pp. 401–401. DOI: doi:10.1038/nj7578-401a. URL: https://www.nature.com/nature/journal/v527/n7578/full/nj7578-401a.html.

[41] Schiermeier, Q. The science behind the Volkswagen emissions scandal. *Nature News* (2015). URL: https://www.nature.com/news/the-science-behind-the-volkswagen-emissions-scandal-1.18426.

[42] Anenberg, S. C., Miller, J., Minjares, R., Du, L., Henze, D. K., Lacey, F., Malley, C. S., Emberson, L., Franco, V., Klimont, Z., and Heyes, C. Impacts and mitigation of excess diesel-related NO_x emissions in 11 major vehicle markets. *Nature* 545 (2017), p. 467. URL: http://dx.doi.org/10.1038/nature22086.

[43] Fenske, J. Engineering Explained: Volkswagen Gassed Monkeys To Prove Diesels Are Clean. How Volkswagen's Diesel Defeat Device Works. Feb 25, 2018. 2018. URL: https://www.youtube.com/watch?v=glHNu91zUdA.

[44] Reitze, A. W. The Volkswagen Air Pollution Emissions Litigation. *Environmental Law Reporter* 46.174 (2016). University of Utah College of Law Research Paper.

[45] Cadogan, J. How Volkswagen Betrayed the World. October 4, 2015. 2015. URL: https://www.youtube.com/watch?v=-VZFP3lW4gU.

[46] Worstall, T. A Parking Lot With 21,000 VW Diesels Parked For Scrap. The Continental Telegraph. March 25, 2018. 2018. URL: https://www.contentaltelegraph.com/uncategorized/a-parking-lot-with-21000-vw-diesels-parked-for-scrap/.

[47] Kable, G. Daimler to recall 774,000 Mercedes models due to emission 'defeat devices': Recall is understood to include newest, Euro 6 diesel engines; number of UK cars affected is unknown. Autocar.co.uk. 12 June 2018. 2018. URL: https://www.autocar.co.uk/car-news/industry/daimler-recall-774000-mercedes-models-due-emission-defeat-devices.

[48] German, J. The emissions test defeat device problem in Europe is not about VW. 2016. URL: http://www.theicct.org/blogs/staff/emissions-test-defeat-device-problem-europe-not-about-vw.

[49] Transport Environment. *#Dieselgate continues: new cheating techniques*. 2016. URL: https://www.transportenvironment.org/sites/te/files/publications/2016_05_Dieselgate_continues_briefing.pdf.

[50] Rauwald, C. and Jennen, B. Porsche Slapped With Diesel Recall by German Regulator. Bloomberg. 18 May 2018, 12:55 BST Updated on 18 May 2018, 14:04 BST. 2018. URL: https://www.bloomberg.com/news/articles/2018-05-18/porsche-slapped-with-diesel-recall-by-german-regulator.

[51] Deutsche Welle. BMW to recall 12,000 cars over faulty emissions software. 24.02.2018. 2018. URL: https://p.dw.com/p/2tFv4.

[52] Attwood, J. Porsche confirms it will no longer offer diesel engines. German manufacturer hasn't offered a diesel-powered machine since February, and will now focus on hybrids and EVs. Autocar: 23 September 2018. 2018. URL: https://www.autocar.co.uk/car-news/new-cars/porsche-confirms-it-will-no-longer-offer-diesel-engines.

[53] McGee, P. All new diesel cars fail EU emissions standards, says study. Even worst-rated petrol vehicles fare better in real-world conditions, campaigners find. The Financial Times, June 6, 2018. 2018. URL: https://www.ft.com/content/9d052960-6903-11e8-b6eb-4acfcfb08c11.

[54] Degraeuwe, B. and Weiss, M. Does the New European Driving Cycle (NEDC) really fail to capture the NO_x emissions of diesel cars in Europe? *Environmental Pollution* 222 (2017), pp. 234–241. URL: http://www.sciencedirect.com/science/article/pii/S0269749116327476.

[55] Lavoie, G. A., Heywood, J. B., C., J., and Keck. Experimental and Theoretical Study of Nitric Oxide Formation in Internal Combustion Engines. *Combustion Science and Technology* 1.4 (1970), pp. 313–326. DOI: 10.1080/00102206908952211. URL: https://doi.org/10.1080/00102206908952211.

[56] Olsen, D. B., Kohls, M., and Arney, G. Impact of Oxidation Catalysts on Exhaust NO_2/NO_x Ratio from Lean-Burn Natural Gas Engines. *Journal of the Air & Waste Management Association* 60.7 (2010), pp. 867–874. DOI: 10.3155/1047-3289.60.7.867. URL: https://doi.org/10.3155/1047-3289.60.7.867.

[57] Liviu-Constantin, S. Simplified mechanism used to estimate the NO_x emission of Diesel engine. Proceedings of the 2nd International Conference on Manufacturing Engineering, Quality and Production Systems. 2010. URL: https://pdfs.semanticscholar.org/0164/41ccd82a030db4bdfb25eb8ee98d404c76f2.pdf.

[58] Heywood, J. B. Internal combustion engine fundamentals (1988). URL: http://www.eng.auburn.edu/~pjones/MECH.5830.6830.6836/Internal_Combustion_Engines_Fundamentals_by_J.B.Heywood.pdf.

[59] Myung, C.-L., Jang, W., Kwon, S., Ko, J., Jin, D., and Park, S. Evaluation of the real-time de-NO_x performance characteristics of a LNT-equipped Euro-6 diesel passenger car with various vehicle emissions certification cycles. *Energy* 132 (2017), pp. 356–369. URL: http://www.sciencedirect.com/science/article/pii/S0360544217308472.

[60] Carslaw, D. C., Beevers, S. D., Tate, J. E., Westmoreland, E. J., and Williams, M. L. Recent evidence concerning higher NO_x emissions from passenger cars and light duty vehicles. *Atmospheric Environment* 45.39 (2011), pp. 7053–7063. DOI: 10.1016/j.atmosenv.2011.09.063. URL: http://www.sciencedirect.com/science/article/pii/S1352231011010260.

[61] He, C., Li, J., Ma, Z., Tan, J., and Zhao, L. High NO_2/NO_x emissions downstream of the catalytic diesel particulate filter: An influencing factor study. *Journal of Environmental Sciences* 35 (2015), pp. 55–61. URL: http://www.sciencedirect.com/science/article/pii/S1001074215002090.

[62] Carslaw, D. C., Murrells, T. P., Andersson, J., and Keenan, M. Have vehicle emissions of primary NO_2 peaked? *Faraday Discussions* 189.0 (2016), pp. 439–454. DOI: 10.1039/C5FD00162E. URL: http://dx.doi.org/10.1039/C5FD00162E.

[63] Grange, S. K. and Carslaw, D. C. Using meteorological normalisation to detect interventions in air quality time series. *Science of The Total Environment* 653 (2019), pp. 578–588. URL: http://www.sciencedirect.com/science/article/pii/S004896971834244X.

[64] Fang, H. L., Huang, S. C., Yu, R. C., Wan, C. Z., and Howden, K. A Fundamental Consideration on NO_x Adsorber Technology for DI Diesel Application. *SAE 2002 Transactions Journal of Engines* 111.3 (2002). URL: https://doi.org/10.4271/2002-01-2889.

[65] Koebel, M., Madia, G., and Elsener, M. Selective catalytic reduction of NO and NO_2 at low temperatures. *Catalysis Today* 73.3 (2002), pp. 239–247. URL: http://www.sciencedirect.com/science/article/pii/S0920586102000068.

[66] Vrabie, V., Scarpete, D., and Zbarcea, O. The New Exhaust Aftertreatment System For Reducing NO_x Emissions Of Diesel Engines: Lean NO_x Trap (LNT). A Study. *Trans Motauto World* 1.4 (2016), pp. 35–38. URL: https://stumejournals.com/journals/tm/2016/4/35.

[67] Grosjean, D. and Harrison, J. Response of chemiluminescence NO_x analyzers and ultraviolet ozone analyzers to organic air pollutants. *Environmental Science and Technology* 19.9 (1985), pp. 862–865. DOI: 10.1021/es00139a016. URL: https://doi.org/10.1021/es00139a016.

[68] Steinbacher, M., Zellweger, C., Schwarzenbach, B., Bugmann, S., Buchmann, B., Ordóñez, C., Prevot, A. S. H., and Hueglin, C. Nitrogen oxide measurements at rural sites in Switzerland: Bias of conventional measurement techniques. *Journal of Geophysical Research* 112.D11 (2007), 13pp. DOI: 10.1029/2006JD007971. URL: http://dx.doi.org/10.1029/2006JD007971.

[69] Fuchs, H., Dubé, W. P., Lerner, B. M., Wagner, N. L., Williams, E. J., and Brown, S. S. A Sensitive and Versatile Detector for Atmospheric NO_2 and NO_x Based on Blue Diode Laser Cavity Ring-Down Spectroscopy. *Environ-*

mental Science and Technology 43.20 (2009), pp. 7831–7836. DOI: 10.1021/es902067h. URL: https://doi.org/10.1021/es902067h.

[70] Villena, G., Bejan, I., Kurtenbach, R., Wiesen, P., and Kleffmann, J. Interferences of commercial NO_2 instruments in the urban atmosphere and in a smog chamber. *Atmospheric Measurement Techniques* 5.1 (2012), pp. 149–159. URL: https://www.atmos-meas-tech.net/5/149/2012/.

[71] Reed, C., Evans, M. J., Di Carlo, P., Lee, J. D., and Carpenter, L. J. Interferences in photolytic NO_2 measurements: explanation for an apparent missing oxidant? *Atmospheric Chemistry and Physics* 16.7 (2016), pp. 4707–4724. URL: https://www.atmos-chem-phys.net/16/4707/2016/.

[72] Kirchstetter, T. W., Harley, R. A., and Littlejohn, D. Measurement of Nitrous Acid in Motor Vehicle Exhaust. *Environmental Science & Technology* 30.9 (1996), pp. 2843–2849. DOI: 10.1021/es960135y. URL: http://dx.doi.org/10.1021/es960135y.

[73] Gutzwiller, L., Arens, F., Baltensperger, U., Gäggeler, H. W., and Ammann, M. Significance of Semivolatile Diesel Exhaust Organics for Secondary HONO Formation. *Environmental Science and Technology* 36.4 (2002), pp. 677–682. DOI: 10.1021/es015673b. URL: https://doi.org/10.1021/es015673b.

[74] Sadanaga, Y., Fukumori, Y., Kobashi, T., Nagata, M., Takenaka, N., and Bandow, H. Development of a Selective Light-Emitting Diode Photolytic NO_2 Converter for Continuously Measuring NO_2 in the Atmosphere. *Analytical Chemistry* 82.22 (2010), pp. 9234–9239. DOI: 10.1021/ac101703z. URL: https://doi.org/10.1021/ac101703z.

[75] Mazzei, F., D'Alessandro, A., Lucarelli, F., Nava, S., Prati, P., Valli, G., and Vecchi, R. Characterization of particulate matter sources in an urban environment. *Science of The Total Environment* 401.1–3 (2008), pp. 81–89. URL: http://www.sciencedirect.com/science/article/pii/S0048969708002751.

[76] Vitolo, C., Elkhatib, Y., Reusser, D., Macleod, C. J., and Buytaert, W. Web technologies for environmental Big Data. *Environmental Modelling & Software* 63 (2015), pp. 185–198. URL: http://www.sciencedirect.com/science/article/pii/S1364815214002965.

[77] European Environment Agency. *AirBase – The European air quality database (Version 8)*. 2014. URL: https://www.eea.europa.eu/data-and-maps/data/airbase-the-european-air-quality-database-8.

[78] European Environment Agency. *Eionet Central Data Repository*. 2017. URL: http://cdr.eionet.europa.eu/.

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

[80] Hamilton, J. F. Using Comprehensive Two-Dimensional Gas Chromatography to Study the Atmosphere. *Journal of Chromatographic Science* 48.4 (2010), pp. 274–282. DOI: 10.1093/chromsci/48.4.274. URL: http://dx.doi.org/10.1093/chromsci/48.4.274.

[81] Carslaw, D. C. and Carslaw, N. Detecting and characterising small changes in urban nitrogen dioxide concentrations. *Atmospheric Environment* 41.22 (2007), pp. 4723–4733. DOI: 10.1016/j.atmosenv.2007.03.034. URL: http://www.sciencedirect.com/science/article/pii/S1352231007002919.

[82] Carslaw, D. and Ropkins, K. *openair: Open-source tools for the analysis of air pollution data*. 2015.

[83] Wickham, H. and Grolemund, G. *R for Data Science. Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media, 2016, 522pp. URL: http://r4ds.had.co.nz/.

[84] NOAA. Integrated Surface Database (ISD). 2016. URL: https://www.ncdc.noaa.gov/isd.

[85] Carslaw, D. ***worldmet**: Import Surface Meteorological Data from NOAA Integrated Surface Database (ISD)*. R package version 0.7.5. 2017. URL: http://github.com/davidcarslaw/worldmet.

[86] OpenAQ. OpenAQ Platform. An open platform for air quality data. 2018. URL: https://openaq.org/.

[87] World Wide Web Consortium. Extensible Markup Language (XML) 1.1 (Second Edition). 2006. URL: https://www.w3.org/TR/2006/REC-xml11-20060816/.

[88] Carslaw, D. C. and Beevers, S. D. Characterising and understanding emission sources using bivariate polar plots and k-means clustering. *Environmental Modelling & Software* 40 (2013), pp. 325–329. URL: http://www.sciencedirect.com/science/article/pii/S136481521200237X.

[89] Uria-Tellaetxe, I. and Carslaw, D. C. Conditional bivariate probability function for source identification. *Environmental Modelling & Software* 59.0 (2014), pp. 1–9. URL: http://www.sciencedirect.com/science/article/pii/S1364815214001339.

[90] Font, A., Hoogh, K. de, Leal-Sanchez, M., Ashworth, D. C., Brown, R. J. C., Hansell, A. L., and Fuller, G. W. Using metal ratios to detect emissions from municipal waste incinerators in ambient air pollution data. *Atmospheric Environment* 113 (2015), pp. 177–186. URL: http://www.sciencedirect.com/science/article/pii/S1352231015300753.

[91] Masiol, M., Hopke, P. K., Felton, H. D., Frank, B. P., Rattigan, O. V., Wurth, M. J., and LaDuke, G. H. Analysis of major air pollutants and submicron particles in New York City and Long Island. *Atmospheric Environment* 148 (2017), pp. 203–214. URL: http://www.sciencedirect.com/science/article/pii/S1352231016308500.

[92] Masiol, M., Hopke, P. K., Felton, H. D., Frank, B. P., Rattigan, O. V., Wurth, M. J., and LaDuke, G. H. Source apportionment of $PM_{2.5}$ chemically speciated mass and particle number concentrations in New York City. *Atmospheric Environment* 148 (2017), pp. 215–229. URL: http://www.sciencedirect.com/science/article/pii/S1352231016308512.

[93] Allen, A. G., Nemitz, E., Shi, J. P., Harrison, R. M., and Greenwood, J. C. Size distributions of trace metals in atmospheric aerosols in the United Kingdom. *Atmospheric Environment* 35.27 (2001), pp. 4581–4591. URL: http://www.sciencedirect.com/science/article/pii/S135223100100190X.

[94] Taiwo, A. M., Beddows, D. C. S., Shi, Z., and Harrison, R. M. Mass and number size distributions of particulate matter components: Comparison of an industrial site and an urban background site. *Science of The Total Environment* 475 (2014), pp. 29–38. URL: http://www.sciencedirect.com/science/article/pii/S004896971301557X.

[95] Davy, P., Trompetter, W. J., Markwitz, A., and Weatherburn, D. C. Elemental Analysis And Source Apportionment Of Ambient Particulate Matter At Masterton, New Zealand. *International Journal of PIXE* 15.3/4 (2005), pp. 225–231. URL: http://search.ebscohost.com/login.aspx?direct=true&db=a9h%5C&AN=18995642%5C&site=ehost-live%5C&scope=site.

[96] de Foy, B. City-level variations in NOx emissions derived from hourly monitoring data in Chicago. *Atmospheric Environment* 176 (2018), pp. 128–139. URL: http://www.sciencedirect.com/science/article/pii/S1352231017308695.

[97] Grange, S. K. ***smonitor**: A framework and a collection of functions to allow for maintenance of air quality monitoring data*. 2018. URL: https://github.com/skgrange/smonitor.

[98] Grange, S. K. smonitor Europe air quality monitoring sites. 2018. URL: http://skgrange.github.io/www/maps/smonitor_europe_sites/smonitor_europe_sites.html.

[99] 52°North. SOS 4.x Documentation. 2018. URL: https://wiki.52north.org/SensorWeb/SensorObservationServiceIVDocumentation.

[100] 52°North. Standardized, web-based upload and download of sensor data and sensor metadata 52°North Sensor Observation Service (SOS). URL: https://github.com/52North/SOS.

[101] 52°North. 52°North Sensor Observation Service 4.x Database model. 2018. URL: https://wiki.52north.org/SensorWeb/SensorObservationServiceDatabaseModel.

[102] PostgreSQL Global Development Group. PostgreSQL. Version 9.5. URL: https://www.postgresql.org/.

[103] Simoens, D. AirBase version 8 data products on EEA data service. European Environment Agency. 2014. URL: http://www.eea.europa.eu/data-and-maps/data/airbase-the-european-air-quality-database-8.

[104] The European Environment Agency. Air Quality e-Reporting (AQ e-Reporting). 2018. URL: https://www.eea.europa.eu/data-and-maps/data/aqereporting-8.

[105] Centre for Environmental Data Analysis. Centre for Environmental Data Analysis (CEDA). Data and information services for environmental science. 2017. URL: http://www.ceda.ac.uk/.

[106] Norwegian Institute for Air Research. EBAS. 2017. URL: http://ebas.nilu.no/.

[107] Japan Meteorological Agency. World Data Centre for Greenhouse Gases (WDCGG). 2017. URL: http://ds.data.jma.go.jp/gmd/wdcgg/.

[108] Hu, L., Keller, C. A., Long, M. S., Sherwen, T., Auer, B., Da Silva, A., Nielsen, J. E., Pawson, S., Thompson, M. A., Trayanov, A. L., Travis, K. R., Grange, S. K., Evans, M. J., and Jacob, D. J. Global simulation of tropospheric chemistry at 12.5 km resolution: performance and evaluation of the GEOS-Chem chemical module (v10-1) within the NASA GEOS Earth system model (GEOS-5 ESM). *Geoscientific Model Development* 11.11 (2018), pp. 4603–4620. URL: https://www.geosci-model-dev.net/11/4603/2018/.

[109] Kiesewetter, G., Borken-Kleefeld, J., Schöpp, W., Heyes, C., Thunis, P., Bessagnet, B., Terrenoire, E., Gsella, A., and Amann, M. Modelling NO_2 concentrations at the street level in the GAINS integrated assessment model: projections under current legislation. *Atmospheric Chemistry and Physics* 14.2 (2014), pp. 813–829. DOI: https://doi.org/10.5194/acp-14-813-2014. URL: http://www.atmos-chem-phys.net/14/813/2014/.

[110] Alvarez, R., Weilenmann, M., and Favez, J.-Y. Evidence of increased mass fraction of NO_2 within real-world NO_x emissions of modern light vehicles — derived from a reliable online measuring method. *Atmospheric Environment* 42.19 (2008), pp. 4699–4707. URL: http://www.sciencedirect.com/science/article/pii/S1352231008000964.

[111] Keuken, M., Roemer, M., and van den Elshout, S. Trend analysis of urban NO_2 concentrations and the importance of direct NO_2 emissions versus ozone/NO_x equilibrium. *Atmospheric Environment* 43.31 (2009), pp. 4780–4783. URL: http://www.sciencedirect.com/science/article/pii/S1352231008006298.

[112] Carslaw, D. C. and Rhys-Tyler, G. New insights from comprehensive on-road measurements of NO_x, NO_2 and NH_3 from vehicle emission remote sensing in London, UK. *Atmospheric Environment* 81.0 (2013), pp. 339–347. URL: http://www.sciencedirect.com/science/article/pii/S1352231013007140.

[113] European Commission. Transport Emissions — Air pollutants from road transport. 2016. URL: http://ec.europa.eu/environment/air/transport/road.htm.

[114] Carslaw, D. C. Evidence of an increasing NO_2/NO_x emissions ratio from road traffic emissions. *Atmospheric Environment* 39.26 (2005), pp. 4793–4802. DOI: https://doi.org/10.1016/j.atmosenv.2005.06.023. URL: http://www.sciencedirect.com/science/article/pii/S1352231005005443.

[115] Grice, S., Stedman, J., Kent, A., Hobson, M., Norris, J., Abbott, J., and Cooke, S. Recent trends and projections of primary NO_2 emissions in Europe. *Atmospheric Environment* 43.13 (2009), pp. 2154–2167. DOI: https://doi.org/10.1016/j.atmosenv.2009.01.019. URL: http://www.sciencedirect.com/science/article/pii/S1352231009000508.

[116] Ligterink, N. E., Kadijk, G., and van Mensch, P. *Determination of Dutch NO_x emission factors for Euro-5 diesel passenger cars*. TNO 2012 R11099. 2012.

[117] Johnson, T. V. Review of Diesel Emissions and Control. *SAE International Journal of Fuels and Lubricants* 3.1 (2010), pp. 16–29. URL: http://dx.doi.org/10.4271/2010-01-0301.

[118] Wild, R. J., Dubé, W. P., Aikin, K. C., Eilerman, S. J., Neuman, J. A., Peischl, J., Ryerson, T. B., and Brown, S. S. On-road measurements of vehicle NO_2/NO_x emission ratios in Denver, Colorado, USA. *Atmospheric Environment* 148 (2017), pp. 182–189. URL: http://www.sciencedirect.com/science/article/pii/S1352231016308469.

[119] Williams, M. L. and Carslaw, D. C. New Directions: Science and policy — Out of step on NO_x and NO_2? *Atmospheric Environment* 45.23 (2011), pp. 3911–3912. DOI: https://doi.org/10.1016/j.atmosenv.2011.04.067. URL: http://www.sciencedirect.com/science/article/pii/S1352231011004560.

[120] Matthaios, V. N., Kramer, L. J., Sommariva, R., Pope, F. D., and Bloss, W. J. Investigation of vehicle cold start primary NO_2 emissions from ambient monitoring data in the UK and their implications for urban air quality. *Atmospheric Environment* (2018). URL: http://www.sciencedirect.com/science/article/pii/S1352231018308033.

[121] Casquero-Vera, J. A., Lyamani, H., Titos, G., Borrás, E., Olmo, F. J., and Alados-Arboledas, L. Impact of primary NO_2 emissions at different urban sites exceeding the European NO_2 standard limit. *Science of The Total Environment* 646 (2019), pp. 1117–1125. URL: http://www.sciencedirect.com/science/article/pii/S0048969718328560.

[122] Porter, P. S., Rao, S. T., Zurbenko, I. G., Dunker, A. M., and Wolff, G. T. Ozone Air Quality over North America: Part II: An Analysis of Trend Detection and Attribution Techniques. *Journal of the Air & Waste Management Association* 51.2 (2001). PMID: 28060596, pp. 283–306. DOI: 10.1080/10473289.2001.10464261. URL: http://dx.doi.org/10.1080/10473289.2001.10464261.

[123] Anh, V., Duc, H., and Azzi, M. Modeling anthropogenic trends in air quality data. *Journal of the Air & Waste Management Association* 47.1 (1997), pp. 66–71. DOI: doi:10.1080/10473289.1997.10464406. URL: https://doi.org/10.1080/10473289.1997.10464406.

[124] Rao, S. T. and Zurbenko, I. G. Detecting and Tracking Changes in Ozone Air Quality. *Journal of the Air & Waste Management Association* 44.9 (1994), pp. 1089–1092. DOI: 10.1080/10473289.1994.10467303. URL: http://dx.doi.org/10.1080/10473289.1994.10467303.

[125] Pryor, S., McKendry, I., and Steyn, D. Synoptic-scale meteorological variability and surface ozone concentrations in Vancouver, British Columbia. *Journal of Applied Meteorology* 34.8 (1995), pp. 1824–1833. DOI: 10.1175/1520-0450(1995)034<1824:SSMVAS>2.0.CO;2. URL: https://doi.org/10.1175/1520-0450(1995)034%3C1824:SSMVAS%3E2.0.CO;2.

[126] Libiseller, C. and Grimvall, A. Model selection for local and regional meteorological normalisation of background concentrations of tropospheric

ozone. *Atmospheric Environment* 37.28 (2003), pp. 3923–3931. URL: http://www.sciencedirect.com/science/article/pii/S1352231003005028.

[127] Libiseller, C., Grimvall, A., Waldén, J., and Saari, H. Meteorological normalisation and non-parametric smoothing for quality assessment and trend analysis of tropospheric ozone data. *Environmental Monitoring and Assessment* 100.1 (2005), pp. 33–52. DOI: 10.1007/s10661-005-7059-2. URL: http://dx.doi.org/10.1007/s10661-005-7059-2.

[128] Wise, E. K. and Comrie, A. C. Extending the Kolmogorov–Zurbenko Filter: Application to Ozone, Particulate Matter, and Meteorological Trends. *Journal of the Air & Waste Management Association* 55.8 (2005), pp. 1208–1216. DOI: 10.1080/10473289.2005.10464718. URL: http://dx.doi.org/10.1080/10473289.2005.10464718.

[129] Breiman, L. Bagging predictors. *Machine Learning* 24.2 (1996), pp. 123–140. DOI: 10.1007/BF00058655. URL: http://dx.doi.org/10.1007/BF00058655.

[130] Breiman, L. Random Forests. *Machine Learning* 45.1 (2001), pp. 5–32. DOI: 10.1023/A:1010933404324. URL: http://dx.doi.org/10.1023/A:1010933404324.

[131] Tong, W., Hong, H., Fang, H., Xie, Q., and Perkins, R. Decision Forest: Combining the Predictions of Multiple Independent Decision Tree Models. *Journal of Chemical Information and Computer Sciences* 43.2 (2003), pp. 525–531. DOI: 10.1021/ci020058s. URL: http://dx.doi.org/10.1021/ci020058s.

[132] Caruana, R. and Niculescu-Mizil, A. An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 161–168.

[133] Friedman, J. H. Recent Advances in Predictive (Machine) Learning. *Journal of Classification* 23.2 (2006), pp. 175–197. URL: http://dx.doi.org/10.1007/s00357-006-0012-4.

[134] Biau, G., Devroye, L., and Lugosi, G. Consistency of Random Forests and Other Averaging Classifiers. *Journal of Machine Learning Research* 9 (2008), pp. 2015–2033.

[135] Kotsiantis, S. B. Decision trees: a recent overview. *Artificial Intelligence Review* 39.4 (2013), pp. 261–283. URL: http://dx.doi.org/10.1007/s10462-011-9272-4.

[136] Ziegler, A. and König, I. R. Mining data with random forests: current options for real–world applications. *WIREs Data Mining and Knowledge Discovery* 4.1 (2013), pp. 55–63. DOI: 10.1002/widm.1114. URL: https://doi.org/10.1002/widm.1114.

[137] Jones, Z. and Linder, F. Exploratory Data Analysis using Random Forests. 73rd annual MPSA conference, April 16-19, 2015, Chicago, United States of America. 2015. URL: https://pdfs.semanticscholar.org/e7b7/3565b07a7f1369a20b1055f222423f0feb34.pdf.

[138] Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., and Gräler, B. Random Forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ Preprints* 6 (2018), e26693v3. URL: https://doi.org/10.7287/peerj.preprints.26693v3.

[139] Federal Office for the Environment. Messstationen des NABEL — Stations de mesure NABEL. Technischer Bericht NABEL 2013. 2014. URL: https://www.bafu.admin.ch/dam/bafu/en/dokumente/luft/fachinfo-daten/nabel-messstationen.pdf.download.pdf/nabel-messstationen.pdf.

[140] Federal Office for the Environment. National Air Pollution Monitoring Network (NABEL). 2017. URL: https://www.bafu.admin.ch/bafu/en/home/topics/air/state/data/national-air-pollution-monitoring-network--nabel-.html.

[141] Heldstab, J., Leippert, F., Wüthrich, P., Künzle, T., and Stampfli, M. PM_{10} and $PM_{2.5}$ ambient concentrations in Switzerland. Modelling results for 2005, 2010, 2020. Federal Office for the Environment. 2013. URL: https://www.bafu.admin.ch/dam/bafu/en/dokumente/luft/uw-umwelt-wissen/pm10_und_pm2_5_immissioneninderschweizzusammenfassung.pdf.download.pdf/pm10_and_pm2_5_ambientconcentrationsinswitzerland.pdf.

[142] Panteliadis, P., Strak, M., Hoek, G., Weijers, E., Zee, S. van der, and Dijkema, M. Implementation of a low emission zone and evaluation of effects on air quality by long-term monitoring. *Atmospheric Environment* 86.0 (2014), pp. 113–119. URL: http://www.sciencedirect.com/science/article/pii/S1352231013009801.

[143] Ricardo Energy & Environment. Kent Air Quality database. 2018. URL: http://www.kentair.org.uk.

[144] International Maritime Organization. Revised MARPOL Annex VI. Annex VI of MARPOL addresses air pollution from ocean-going ships. 2005. URL: http://www.imo.org/en/OurWork/Environment/PollutionPrevention/AirPollution/Pages/Air-Pollution.aspx.

[145] Entec. Defra UK Ship Emissions Inventory—Final Report. Doc Reg No. 21897-01. 2010. URL: https://uk-air.defra.gov.uk/assets/documents/reports/cat15/1012131459_21897_Final_Report_291110.pdf.

[146] Cecil, N. Oxford Street pollution levels breached EU annual limit just four days into 2015. Evening Standard. 6 January 2015. 2015. URL: https://www.standard.co.uk/news/london/oxford-street-pollution-levels-breached-annual-limit-just-four-days-into-2015-9959849.html.

[147] Carrington, D. London breaches annual air pollution limit for 2017 in just five days. The Guardian. 6 Jan 2017. 2017. URL: https://www.theguardian.com/environment/2017/jan/06/london-breaches-toxic-air-pollution-limit-for-2017-in-just-five-days.

[148] Dunham, M. Brixton Road breaches annual air pollution limit in five days. BBC. 6 January 2017. 2017. URL: https://www.bbc.co.uk/news/uk-england-london-38529928.

[149] Cecil, N. Pollution breached EU limits at nearly 50 sites in London last year - Worst area for pollution in London revealed to be Brixton Road in Lambeth. Evening Standard. 5 January 2018. 2018. URL: https://www.standard.co.uk/news/london/pollution-breached-eu-limits-at-nearly-50-sites-in-london-last-year-a3732701.html.

[150] Carslaw, D. C. and Beevers, S. D. Investigating the potential importance of primary NO_2 emissions in a street canyon. *Atmospheric Environment* 38.22 (2004), pp. 3585–3594. URL: http://www.sciencedirect.com/science/article/pii/S1352231004003267.

[151] Westmoreland, E. J., Carslaw, N., Carslaw, D. C., Gillah, A., and Bates, E. Analysis of air quality within a street canyon using statistical and dispersion modelling techniques. *Atmospheric Environment* 41.39 (2007), pp. 9195–9205. URL: http://www.sciencedirect.com/science/article/pii/S1352231007006863.

[152] Jeanjean, A. P. R., Buccolieri, R., Eddy, J., Monks, P. S., and Leigh, R. J. Air quality affected by trees in real street canyons: The case of Marylebone neighbourhood in central London. *Urban Forestry & Urban Greening* 22 (2017), pp. 41–53. DOI: https://doi.org/10.1016/j.ufug.2017.01.009. URL: http://www.sciencedirect.com/science/article/pii/S1618866716303740.

[153] DieselNet.com. CRT Filter. Revision 2005.01c. 2005. URL: https://www.dieselnet.com/tech/dpf_crt.php.

[154] Grange, S. K. ***rmweather**: Tools to Conduct Meteorological Normalisation on Air Quality Data*. R package version 0.1.2. 2018. URL: https://CRAN.R-project.org/package=rmweather.

[155] R Core Team. The Comprehensive R Archive Network. 2018. URL: https://cran.r-project.org/.

[156] Wright, M. N. and Ziegler, A. **ranger**: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* 77.1 (2017), pp. 1–17. DOI: 10.18637/jss.v077.i01. URL: http://dx.doi.org/10.18637/jss.v077.i01.

[157] Wickham, H. Tidy Data. *Journal of Statistical Software* 59.10 (2014). URL: http://www.jstatsoft.org/v59/i10.

[158] Carslaw, D. and Grange, S. *polarplotr: Functions to plot polar-plots*. 2016. URL: https://github.com/davidcarslaw/polarplotr.

[159] Grange, S. K. ***normalweatherr**: Package to conduct meteorological/weather normalisation on air quality data (deprecated)*. 2017. URL: https://github.com/skgrange/normalweatherr.

Chapter 2

Enhancements to bivariate polar plots

This work was originally published in *Atmospheric Environment* on 14 September, 2016.[†]

2.1 Abstract

This paper outlines the development of enhanced bivariate polar plots that allow the concentrations of two pollutants to be compared using pair-wise statistics for exploring the sources of atmospheric pollutants. The new method combines bivariate polar plots, which provide source characteristic information, with pair-wise statistics that provide information on how two pollutants are related to one another. The pair-wise statistics implemented include weighted Pearson correlation and slope from two linear regression methods. The development uses a Gaussian kernel to locally weight the statistical calculations on a wind speed-direction surface together with variable-scaling. Example applications of the enhanced polar plots are presented by using routine air quality data for two monitoring sites in London, United Kingdom for a single year (2013). The London examples demonstrate that the combination of bivariate polar plots, correla-

[†]https://doi.org/10.1016/j.atmosenv.2016.09.016

tion, and regression techniques can offer considerable insight into air pollution source characteristics, which would be missed if only scatter plots and mean polar plots were used for analysis. Specifically, using correlation and slopes as pair-wise statistics, long-range transport processes were isolated and black carbon (BC) contributions to $PM_{2.5}$ for a kerbside monitoring location were quantified. Wider applications and future advancements are also discussed.

2.2 Introduction

Determining how variables are related to one-another is a key component of data analysis and statistics. Within the atmospheric sciences, exploring the relationships between chemical constituents and meteorological parameters is extremely common and the techniques for comparing, correlating, and determining relationships are very diverse. Analysis involving the correlation of two pollutants can often be insightful because it can lead to the identification of emission source characteristics, as can investigation into ratios or slopes from regression analysis between two pollutants.[1] Within atmospheric disciplines, data analysis can also benefit from being able to integrate wind behaviour.[2] The use of wind speed and direction can be informative because it often leads to the suggestion of source locations and source characteristics, such as height of emission above the surface.[3,4]

Exploration of relationships among variables can be achieved with many different methods that can range from the simple to the numerically complex. However, a technique that is used very widely is the simple x-y scatter plot.[5] Scatter plots are useful because they allow for the visualisation of variables and model fitting can be evaluated quickly and simply with visual feedback. Regression techniques, most commonly ordinary least-squared regression, are often employed to formally quantify how x and y are related. The use of least-squared regression is however technically questionable in many cases, and despite a large collection of alternative techniques

available, its use remains a persistent feature of air quality data analysis. The use of simple scatter plots is usually carried out with entire datasets or with simple or superficial filtering and therefore have potential to hide some discrete relationships which are present in the global data if they do not conform to the mean rate of change.[6]

Slopes from regression models relating two pollutants to one another are often used in applications that use monitoring data such as emission inventories and pollutant models. When measurements are not available, slopes for the unknown pollutants are often substituted from the literature, short-term monitoring, or data collected at a near-by location. However, the use of simple and static ratios is likely to be deficient in many situations because they can be expected to be highly dependent on source and time.[7] To differentiate sources in air quality data, techniques other than simple scatter plots often need to be used.

A common method for source characterisation is the use of bivariate polar plots.[4,8–10] Polar plots are typically used to visualise and explore mean pollutant concentrations for single species based on wind speed and wind direction. In the atmospheric sciences, it is intuitive to plot wind direction (from 0 to 360 ° clockwise from north) on the angular "axis" and wind speed to be used for the radial scale. Aggregation functions other than the arithmetic mean can be used and different variables apart from wind speed can be used for the radial scale. For example, atmospheric temperature or stability are often useful variables to use. The main attribute for the choice of radial-axis variable is that it helps to differentiate between different source characteristics in some way due to different source types responding differently to values of the angular scale. Despite the range of potential options, wind speed is widely used to help discriminate different source types and is particularly useful when used together with wind direction and the concentration of a species.[11,12]

This type of polar plot analysis has, in part, become wide-spread due to the open source `polarPlot` function available in the **openair** R pack-

age.[13,14] Other similar techniques such as non-parametric wind regression have also shown their ability to determine source locations for various pollutants by using polar plots.[3,15,16]

2.2.1 Objectives

Combining correlation and regression techniques with those that provide information on source apportionment potentially offers considerably more insight into air pollution sources. The use of wind behaviour has the potential to evaluate correlation and slopes based on source locations and therefore different processes. It is common for emission inventories to use ratios for pollutants when they are not measured or when high quality data is lacking. It is hypothesised that the combination of correlation, regression, and polar plots could lead to significant additions to data analysis by understanding how different pollutants are related to one another depending on source.

In this paper, the combination of bivariate polar plots approaches with correlation and regression techniques is considered for comparing two pollutants. This combination of methods is then used to aid the interpretation of air quality data. The primary objectives of this paper are as follows. First, to develop methods to combine bivariate polar plot techniques with correlation and a range of linear regression approaches. Second, apply the methods to commonly available measurements of air pollutants to demonstrate the new insights made possible by these techniques. Third, to consider the wider potential uses of the approaches in air quality science. The software developed has been released with an open source licence and can be found in the **polarplotr** R package.[17]

2.3 Methods

2.3.1 Function development

2.3.1.1 Kernel weighting and scaling

The plotting mechanism for polar plots when using wind direction as the polar axis generally involves first aggregating a time-series into wind speed and direction intervals (or bins). The specific intervals and numbers of the bins can be altered for a particular application, but all combinations of the two types of bins are summarised by an aggregation function such as the mean or maximum. In the **openair** `polarPlot` function, a smoothed surface is fitted to these binned summaries using a generalised additive model (GAM) to create a continuous surface which can be plotted with polar coordinates. Further details of the approach can be found in Carslaw and Beevers [9] and Uria-Tellaetxe and Carslaw [10].

When applying a simple aggregation function, the number of observations in a time-series which compose a discrete wind speed and direction bin is not critical for the calculation or the visual presentation of the surface, except at the edges of the plot where there are (usually) few observations. However, when calculating correlations or relationships between two variables, it becomes important to consider the minimal number of observations which would create a valid summary. If there are too few observations for a particular bin and a statistic such as the correlation or slope is calculated between a pair of variables, it is likely that unreliable summaries will be generated due to large variations between neighbouring bins. To overcome this limitation, for each wind speed and direction bin, the entire time-series was evaluated but observations were *weighted* by their proximity to a wind speed and direction bin *i.e.*, wind speed or direction values further from the bin centre are weighted less than those closer to the centre of the bin. Like previous works such as Henry et al. [3] and Henry et al. [15], a weighting kernel was used to create weighting variables.

The weighting kernel used was the Gaussian kernel (Equation 2.1). The Gaussian kernel has infinite tails and therefore all input bins are given a non-zero weighting, but observations furthest from the bin being analysed have very small weights associated with them. The Gaussian kernel was used for weighting both wind speed and direction because it is considered more utilitarian than many other kernels such as the Epanechnikov kernel which have finite bounds and therefore at times, will give observations weights of zero which can cause ambiguity issues.

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \tag{2.1}$$

To ensure the weighing variable was appropriate for the particular wind speed and direction application, the input wind speed and direction variables required scaling. The scaling process used was simple; the wind variables were multiplied by an integer to increase their bounds and therefore influence within the weighting kernel. The variables were also normalised to ensure that all observations had values between zero and one. This normalisation step is not strictly necessary when the Gaussian kernel is used, but is needed for some other kernels and ensures the output of process always had a known range.

If the weighting operated too locally, the inherently variable nature of wind behaviour was represented in the plotted surface as noise. Conversely, if weighting was extended too far, isolated areas of 'real' peaks were obscured due to over-smoothing. It is difficult to determine an optimal set of scaling values for wind speed and direction for every application, therefore a series of heuristic simulations were performed to determine the ideal integer scaling values.

It was found that within a central range the final output was rather insensitive to the scaling values. One reason for this relative insensitivity will be due to the inherent random variability of concentrations as a function of either wind speed or wind direction due to atmospheric turbulence. This indicates that within a central band of values, the scaling process is not par-

ticularly influential. It is possible for other applications that these scaling magnitudes will have to be tuned and therefore the defaults can be altered by the user. An example of the scaling defaults used in the `polarPlot` function are shown in Figure 2.1. Figure 2.1 allows visualisation of the Gaussian weighting kernel for both the wind speed and direction variables as well as the extent of the default scaling procedure for a single bin for $4.8\,\mathrm{m\,s^{-1}}$ and 230 degrees.

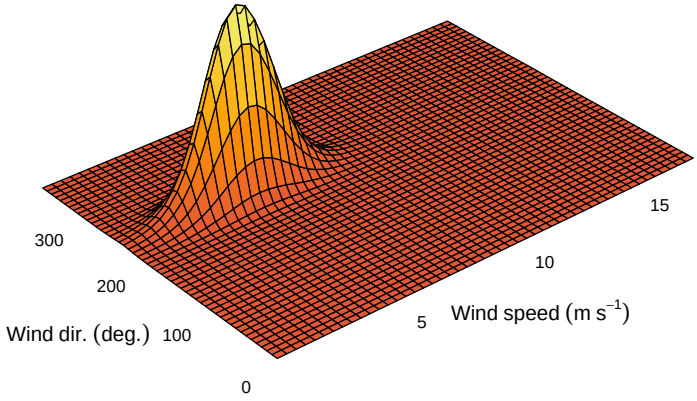

Figure 2.1: Three-dimensional surface of weights for a single wind speed and direction bin ($4.8\,\mathrm{m\,s^{-1}}$ and 230 degrees respectively). The surface is normalised and therefore intensity units are not informative.

After the appropriate weights have been calculated, the calculation of any pair-wise statistic that allows for weighting could be calculated between two pollutants. The first methods implemented were the Pearson correlation coefficient and two linear regression methods. Using these two groups of techniques allowed for the investigation of the correlation between two pollutants and the investigation of the slope between pollutants, but with the inclusion of wind speed and direction.

2.3.1.2 Correlation

Correlation is a measure of how well two (or more) variables are associated to one-another. Correlation is a useful measure for air pollutants because pollutants which demonstrate high levels of correlation are often emitted

from the same source, or undergo similar chemical and physical transformations in the atmosphere. For use in polar plots, the correlation statistic implemented was the weighted Pearson correlation coefficient (r).[18,19]

2.3.1.3 Regression

Regression is a very common statistical technique and is often used to describe and investigate relationships among variables.[20] Regression is a large topic and only the linear regression techniques considered for the polar plot function will be discussed. Of particular interest is the estimate of the slope from a linear regression between two species. The slope will often reveal useful information concerning source characteristics, for example, the amount of PM_{10} that is in the fine fraction ($PM_{2.5}$), or the ratios of combustion products such as CO and NO_x which can be compared with emission inventory estimates.

The first regression technique implemented was weighted least-squares linear regression. This is very similar to ordinary least-squares linear regression, but the weighted sum of squares are minimised which has the effect of creating a model which preferentially represents a local area of the input data rather than the entire set. Because of the common presence of outliers in air pollution time-series measurements, other regression methods such as robust regression can offer advantages over the least-squares regression for use in the enhanced polar plots.

Robust regression extends least-squares regression techniques in attempting to better handle situations where the parametric assumptions of the least-squares regression method are violated. These violations are usually involved with the presence of outliers and heteroscedasticity (non-equal variances). Primarily, the power of robust regression lies in the resistance to the influence of outliers. Robust regression achieves this by substituting the least-squares estimator for a more robust estimator.[21] There are many types of robust estimators, but they all operate by first classing observations as outliers or not-outliers and then reducing the influence of the outliers on

the regression model.[22] The procedures for calculating robust estimators are iterative and more computationally demanding when compared to the calculation of the least-squares estimator. This is noticeable to a user of the `polarPlot` function because additional run-time is needed when the robust regression techniques are used. The robust regression functions were supplied by the **MASS** package and the estimator used was the M-estimator because this estimator allows the use of weights.[23]

2.3.2 Data

Data analysis was conducted on hourly air quality monitoring data for two sites included in the United Kingdom's Automatic Urban and Rural Network (AURN). The two sites were London Marylebone Road and London North Kensington (Table 2.1 and Figure 2.2). Monitoring data for 2013 were downloaded using the **openair** `importAURN` function. Both monitoring sites measure a large complement of chemical and particulate species and achieve high data capture rates. The particulate matter measurements were focused on for polar plot analysis and PM_{10} and $PM_{2.5}$ at London Marylebone Road and London North Kensington are monitored by TEOM-FDMS (Tapered Element Oscillating Microbalance-Filter Dynamics Measurement System) instruments. This enhanced method is not as susceptible to removing volatile and semi-volatile components in the monitored air-stream as standard heated TEOMs.[24,25] Hourly black carbon (BC) data were also used and these data were sourced directly from the AURN monitoring database after personal communication with Ricardo Energy & Environment. More detailed site and instrument details can be found see at `https://uk-air.defra.gov.uk/`.

Meteorological data for 2013 from London Heathrow (a major airport) in western London were used to represent regional conditions for the two air quality monitoring sites. Hourly data from the London Heathrow site were obtained from the NOAA Integrated Surface Database (ISD) and access was gained with the **worldmet** R package.[26,27] The data from Heathrow Airport

Table 2.1: Details of locations of air quality and meteorological monitoring sites in London providing data for this study.

Site name	Latitude	Longitude	Elevation	Site type
London North Kensington	51.5211	-0.2134	5	Urban background
London Marylebone Road	51.5225	-0.1546	35	Urban traffic
London Heathrow	51.4780	-0.4610	25.3	Meteorological only

Figure 2.2: Locations of air quality and meteorological monitoring sites in London providing data for this study. The map's internal polygons show London's Boroughs, the City of London, and the River Thames.

were used in preference to other local surface measurements, which tend to be strongly influenced by local buildings.

2.4 Results & discussion

2.4.1 London North Kensington PM_{10} and $PM_{2.5}$

London North Kensington is an urban background site (Table 2.1 and Figure 2.2) and it is expected that a wide range of sources will contribute particle concentrations, including both local (London) and long-range (conti-

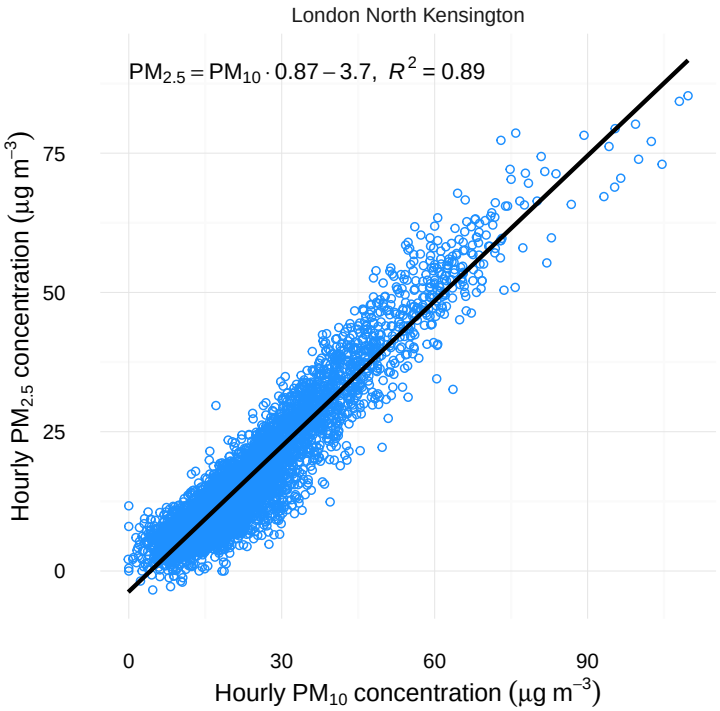

Figure 2.3: Simple x-y scatter plot of $PM_{2.5}$ and PM_{10} for 2013 at London North Kensington. Fitted line and equation represents the ordinary least-squared regression model.

nental Europe) sources. A scatter plot of $PM_{2.5}$ and PM_{10} shows that the two particle size fractions showed a good degree of correlation during 2013 (Figure 2.3). From Figure 2.3 alone there is no obvious indication that different source types contribute to the overall scatter of points. The mean ratio between $PM_{2.5}$ and PM_{10} was 0.87, as determined by the ordinary least-squares linear regression model and it explained 89 % of the variation (Figure 2.3).

The usual use of polar plots, by calculating the mean concentration for wind speed and directions bins, show that the there were multiple sources of PM_{10} and $PM_{2.5}$ at London North Kensington in 2013 (Figure 2.4a and Figure 2.4b). Figure 2.4 suggests that locally-sourced particulate matter were present, as potentially indicated by the elevated concentrations at low wind speeds, but the highest concentrations were experienced with easterly winds when wind speeds were high ($\approx 10\,\mathrm{m\,s^{-1}}$). By contrast, NO_x, a pol-

lutant which is dominated by local (London) emissions, showed that only when wind speeds were low, were elevated concentrations experienced due to a lack of pollutant dispersion (Figure 2.4c). However, when the $PM_{2.5}$ and PM_{10} data are plotted with a correlation statistic binned by wind speed and direction, the situation is more revealing than the scatter plot and mean polar plots would suggest alone (Figure 2.5).

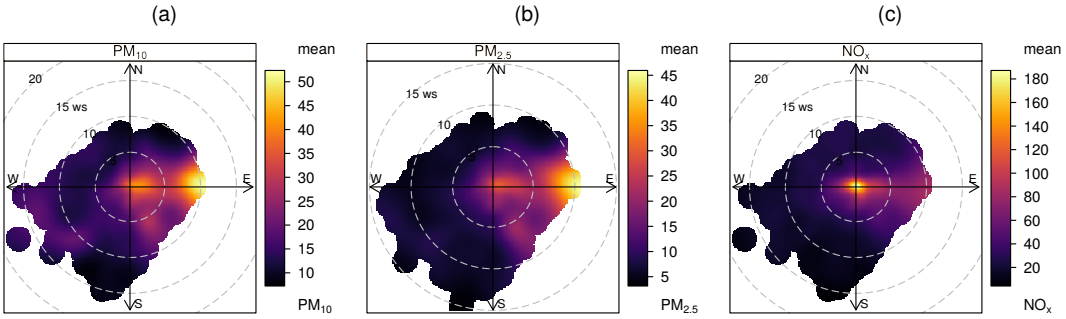

Figure 2.4: Polar plots of mean concentrations of PM_{10} (a), $PM_{2.5}$ (b), and NO_x (c) for 2013 at London North Kensington.

The correlation polar plot of $PM_{2.5}$ and PM_{10} demonstrates that during easterly winds, the London North Kensington $PM_{2.5}$ and PM_{10} concentrations were very highly correlated with $r \approx 0.9$ (Figure 2.5). The zone of high correlation is interpreted to be due to long-range transport which is characterised by the majority of PM_{10} being made up of $PM_{2.5}$. In London, and most areas of the UK, long-range transport is most important under easterly conditions where air-masses originate from continental Europe.[28–30] Under these conditions the concentrations of fine particulate sulphate and nitrate can dominate absolute particle concentrations. The surface of Figure 2.5 is also smooth and covers a wide range of wind speed and directions which indicates a general, and large-scale process which is being appropriately smoothed and represented by the weighting procedure (Section 2.3.1). Other monitoring locations, including London Marylebone Road that also measure $PM_{2.5}$ and PM_{10} showed similar easterly behaviour (not shown).

Previous studies such as Harrison et al. [11], Liu and Harrison [30], Querol et al. [31], and Charron and Harrison [32] have reported high $PM_{2.5}$–

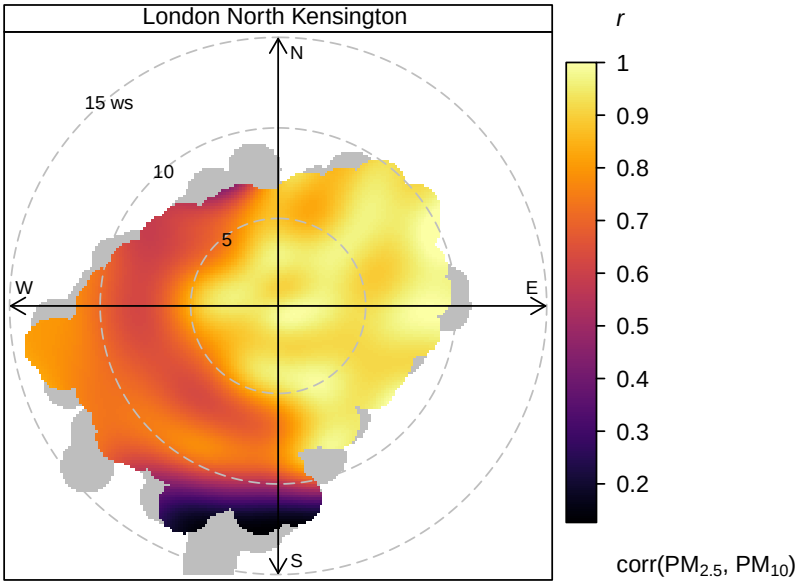

Figure 2.5: Polar plot of the correlation between $PM_{2.5}$ and PM_{10} for 2013 at London North Kensington. Grey areas indicate when fewer than two observations were present and were treated as missing due to the low number of samples.

PM_{10} ratios for European sourced particulate matter in the UK and the correlation presented in Figure 2.5 is consistent with these past studies which reported high $PM_{2.5}$–PM_{10} ratios. When HYSPLIT[33] back-trajectories for 2013 were clustered and joined to coincident pollutant observations, the cluster representing air-masses from Europe also had the highest $PM_{2.5}$–PM_{10} ratio of all clusters, consistent with the conclusions inferred from Figure 2.5.

The polar plot of the slope between $PM_{2.5}$ and PM_{10} at London North Kensington demonstrates a similar surface pattern as the correlation polar plot (Figure 2.6). The long-range sourced particulate from the east was indeed primarily composed of $PM_{2.5}$, as shown by a $PM_{2.5}$ to PM_{10} slope of about 90 %. For other wind directions, coarser particulate matter was a more important contributor to PM_{10} and the $PM_{2.5}$ contributions drop to approximately 30 % (Figure 2.6). This reduction of $PM_{2.5}$ to PM_{10} slope was most likely caused the local process of mechanical resuspension. Even

though the scatter plot of $PM_{2.5}$ and PM_{10} (Figure 2.3) does not indicate different source influences, it is clear from Figure 2.6 in particular that there are at least two major source types affecting particulate concentrations at the London North Kensington site. It should be noted that a careful wind speed, wind direction subset of the data shown in Figure 2.3 does confirm the behaviour seen in Figure 2.6 with a much lower $PM_{2.5}$ to PM_{10} slope for south-westerly winds above $5\,m\,s^{-1}$.

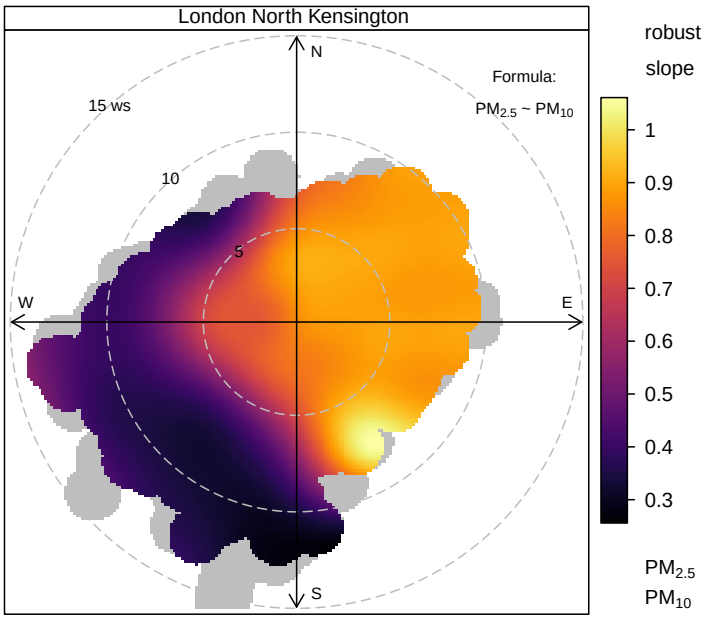

Figure 2.6: Polar plot of the robust slope between $PM_{2.5}$ and PM_{10} for 2013 at London North Kensington. Grey areas indicate when fewer than two observations were present and were treated as missing due to the low number of samples.

2.4.2 London Marylebone $PM_{2.5}$ and BC

Unlike PM_{10} and $PM_{2.5}$ at London North Kensington, the London Marylebone Road BC and $PM_{2.5}$ correlation was poor in 2013, as shown in Figure 2.7. Although BC exists primarily within the fine particle fraction[34,35] and would be expected to be an important component of $PM_{2.5}$ at a traffic-dominated location like London Marylebone Road, $PM_{2.5}$ also has a diverse

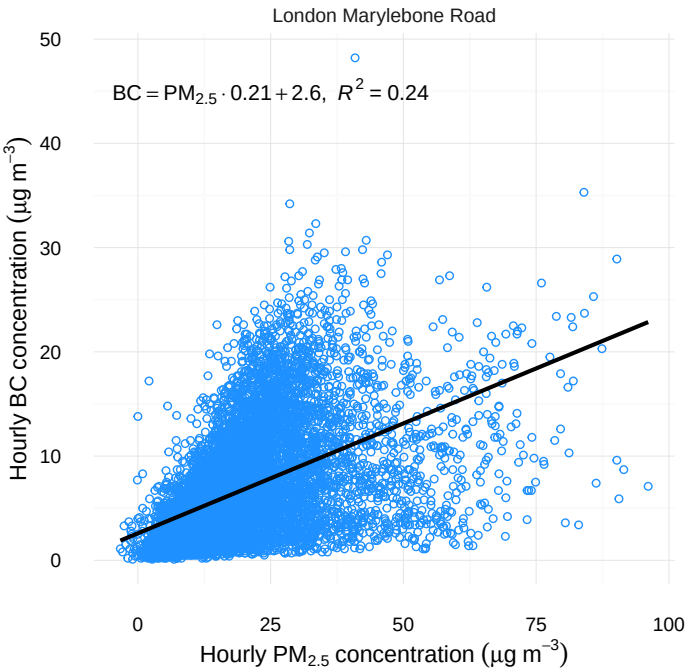

Figure 2.7: Simple *x*-*y* scatter plot of BC and $PM_{2.5}$ for 2013 at London Marylebone Road. Fitted line and equation represents the ordinary least-squared regression model.

number of other sources including secondary inorganic aerosol.[31] Therefore, at times, BC will be a major contributor to $PM_{2.5}$ while at others it will be a minor component depending on the strength of the various sources. Using a scatter plot to investigate this relationship is not immediately useful because the two variables do not follow a mean rate of change. Therefore, fitting a simple linear regression line to these data is not informative (Figure 2.7).

The robust regression slope of BC and $PM_{2.5}$ binned by wind speed and direction at London Marylebone Road demonstrated patterns that were not observed by the simple scatter plot alone (Figure 2.8a). Figure 2.8a shows that the ratio between BC and $PM_{2.5}$ was highly dependent on wind direction. Winds from the south and west at London Marylebone Road had a higher ratio of BC with $\approx 50\,\%$ of $PM_{2.5}$ being composed of BC. BC-$PM_{2.5}$ ratios are sparsely reported, however London Marylebone Road's ratio is con-

sistent with what Ruellan and Cachier [36] reported for a traffic-dominated monitoring location in Paris (Porte d'Auteuil) with ratios of 43 ± 20 %. When winds were from the north and westerly directions, the BC-$PM_{2.5}$ ratio was lower, usually under 20 %. Additionally, winds from the north were nearly completely free of BC particulate matter (Figure 2.8a).

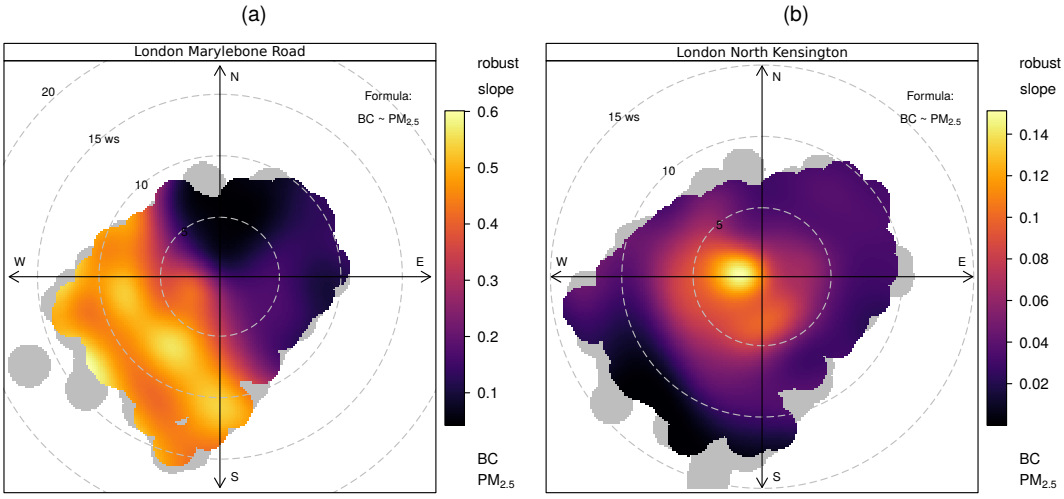

Figure 2.8: Polar plot of the robust slope between BC and $PM_{2.5}$ at London Marylebone Road (a) and London North Kensington (b). Grey areas indicate when fewer than two observations were present and were treated as missing due to the low number of samples.

The wind direction dependencies inferred from the polar plot are somewhat counter-intuitive given that the London Marylebone Road monitoring site is located one metre from the kerb on the south-side of an arterial road (Figure 2.9). However, the site is also within a significant street-canyon with a width of 40 m and a height of 41 m which is likely to lead to complex recirculation patterns at a range of wind speeds.[32,37] Based on this evidence, accumulation of pollutants on the buildings' lee-side (south) is an important process to consider at London Marylebone Road when interpreting source processes.

London North Kensington also measures BC and $PM_{2.5}$ and the slope of these two pollutants binned by wind speed is rather different compared with London Marylebone Road (Figure 2.8b). London North Kensington

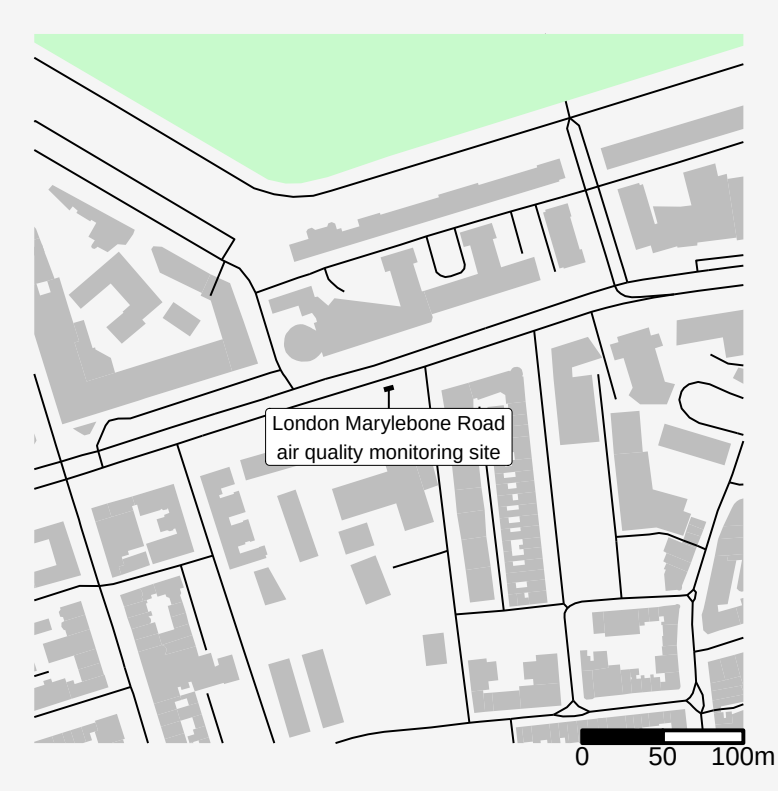

Figure 2.9: Location of the London Marylebone Road monitoring site in Central London and its surrounds. The lines represent public roads.

is an urban background site and lacks the large traffic source being in immediate proximity which London Marylebone Road experiences. Therefore, BC was a much smaller component of $PM_{2.5}$. In 2013, London North Kensington had a maximum contribution of $\approx$ 15% of BC to $PM_{2.5}$ (Figure 2.8b). However, this maximum contribution only occurred when wind speeds were low and suggests that this contribution is reached only when local traffic emissions influence the monitoring site.

Based on these results for the two monitoring sites, the clear and consistent BC-$PM_{2.5}$ ratio at London Marylebone Road of around 50% shown in Figure 2.8a in the south-west quadrant can be interpreted as a contribution dominated by local traffic sources. The lower ratio of between 10–20% mostly to the east is dominated by regional source contributions where the concentration of $PM_{2.5}$ is relatively high but where air masses contain very little BC.

2.4.3 Future directions

The examples presented for a single year of data for two air quality monitoring sites in London were the first steps for enhancing polar plots to include the functionality of pair-wise statistics. The enhancements were able to substantially improve the information content available from routinely monitored air pollutants where simple scatter plots and "standard" polar plots gave no suggestion of the processes subsequently illuminated by the correlation/slope polar plots.

The examples reported were for a few commonly measured species. However, it is expected that the use of polar plots using pair-wise statistics for multi-species data such as metal or VOC concentrations could be highly informative. Measurement of large numbers of metals and other species at higher time resolutions (hourly) is becoming more common. A correlation matrix of robust slope polar plots would potentially reveal more detailed information on common source origins.

The use of other statistics is another valuable future direction such as non-parametric measures of correlation such as Spearman. Other regression techniques such as quantile regression[38] could be implemented to provide slope information across a range of quantile levels, potentially providing more comprehensive information on the relationship between two pollutants and give further options when determining pollutant sources. The main advantage of quantile regression is likely to be related to resolving two or more sources that overlap and where there is not a single dominant slope caused by one source. In this case, considering the full distribution of slope values may help better resolve competing source contributions. Finally, the weighted statistics approach for paired statistics could usefully be extended to model evaluation where two sets of data are compared (observed and modelled). In this case, enhanced polar plot analyses could provide valuable information concerning where model agreement is good or poor and indicate more clearly the conditions under which model performance is acceptable and provide enhanced information on where model

performance is poor.

2.5 Conclusions

This paper outlined the development of enhanced bivariate polar plots to include pair-wise statistics to be used in the atmospheric sciences. Two groups of statistical techniques were implemented: correlation and regression. The new development brings together commonly used pair-wise statistics and relationships with wind speed and direction, which provides enhanced information on pollutant sources beyond currently used techniques.

Using a single year of data, in a single city, for routinely monitored pollutants demonstrated that the enhanced polar plots were capable of determining relationships and processes that were not suggested by simple scatter plots and the use of mean polar plots alone. Here we have reported that traffic dominated $PM_{2.5}$ is composed of 50 % BC at a London monitoring site. This is an important observation and ratios between other pollutants such as elemental carbon and organic carbon (EC and OC) is an obvious future application for the enhanced polar plots.

It is expected in the future that enhanced polar plots will be widely used for the investigation of ratios for pairs of pollutants and further extended to be a valuable tool for teasing apart pollutant sources and processes.

2.6 References

[1] Statheropoulos, M., Vassiliadis, N., and Pappa, A. Principal component and canonical correlation analysis for examining air pollution and meteorological data. *Atmospheric Environment* 32.6 (1998), pp. 1087–1095. URL: http://www.sciencedirect.com/science/article/pii/S1352231097003774.

[2] Elminir, H. K. Dependence of urban air pollutants on meteorology. *Science of The Total Environment* 350.1–3 (2005), pp. 225–237. URL: http://www.sciencedirect.com/science/article/pii/S0048969705000732.

[3] Henry, R. C., Chang, Y.-S., and Spiegelman, C. H. Locating nearby sources of air pollution by nonparametric regression of atmospheric concentrations on wind direction. *Atmospheric Environment* 36.13 (2002), pp. 2237–2244. URL: http://www.sciencedirect.com/science/article/pii/S1352231002001644.

[4] Westmoreland, E. J., Carslaw, N., Carslaw, D. C., Gillah, A., and Bates, E. Analysis of air quality within a street canyon using statistical and dispersion modelling techniques. *Atmospheric Environment* 41.39 (2007), pp. 9195–9205. URL: http://www.sciencedirect.com/science/article/pii/S1352231007006863.

[5] Bentley, S. Graphical techniques for constraining estimates of aerosol emissions from motor vehicles using air monitoring network data. *Atmospheric Environment* 38.10 (2004), pp. 1491–1500. URL: http://www.sciencedirect.com/science/article/pii/S1352231003010574.

[6] Cade, B. S. and Noon, B. R. A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment* 1.8 (2003), pp. 412–420. URL: http://dx.doi.org/10.1890/1540-9295(2003)001[0412:AGITQR]2.0.CO;2.

[7] Manoli, E., Voutsa, D., and Samara, C. Chemical characterization and source identification/apportionment of fine and coarse air particles in Thessaloniki, Greece. *Atmospheric Environment* 36.6 (2002), pp. 949–961. URL: http://www.sciencedirect.com/science/article/pii/S1352231001004861.

[8] Carslaw, D. C., Beevers, S. D., Ropkins, K., and Bell, M. C. Detecting and quantifying aircraft and other on-airport contributions to ambient nitrogen oxides in the vicinity of a large international airport. *Atmospheric Environment* 40.28 (2006), pp. 5424–5434. URL: http://www.sciencedirect.com/science/article/pii/S1352231006004250.

[9] Carslaw, D. C. and Beevers, S. D. Characterising and understanding emission sources using bivariate polar plots and k-means clustering. *Environmental Modelling & Software* 40 (2013), pp. 325–329. URL: http://www.sciencedirect.com/science/article/pii/S136481521200237X.

[10] Uria-Tellaetxe, I. and Carslaw, D. C. Conditional bivariate probability function for source identification. *Environmental Modelling & Software* 59.0 (2014), pp. 1–9. URL: http://www.sciencedirect.com/science/article/pii/S1364815214001339.

[11] Harrison, R. M., Yin, J., Mark, D., Stedman, J., Appleby, R. S., Booker, J., and Moorcroft, S. Studies of the coarse particle (2.5–10 μm) component in UK urban atmospheres. *Atmospheric Environment* 35.21 (2001), pp. 3667–3679. URL: http://www.sciencedirect.com/science/article/pii/S1352231000005264.

[12] Kassomenos, P., Vardoulakis, S., Chaloulakou, A., Grivas, G., Borge, R., and Lumbreras, J. Levels, sources and seasonality of coarse particles (PM_{10}–$PM_{2.5}$) in three European capitals — Implications for particulate pollution control. *Atmospheric Environment* 54.0 (2012), pp. 337–347. URL: http://www.sciencedirect.com/science/article/pii/S1352231012001665.

[13] Carslaw, D. C. and Ropkins, K. *openair* — An R package for air quality data analysis. *Environmental Modelling & Software* 27–28.0 (2012), pp. 52–61. URL: http://www.sciencedirect.com/science/article/pii/S1364815211002064.

[14] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2018. URL: https://www.R-project.org/.

[15] Henry, R., Norris, G. A., Vedantham, R., and Turner, J. R. Source Region Identification Using Kernel Smoothing. *Environmental Science & Technology* 43.11 (2009), pp. 4090–4097. DOI: 10.1021/es8011723. URL: http://dx.doi.org/10.1021/es8011723.

[16] Donnelly, A., Misstear, B., and Broderick, B. Application of nonparametric regression methods to study the relationship between NO_2 concentrations and local wind direction and speed at background sites. *Science of The Total Environment* 409.6 (2011), pp. 1134–1144. URL: http://www.sciencedirect.com/science/article/pii/S0048969710012726.

[17] Carslaw, D. and Grange, S. *polarplotr: Functions to plot polar-plots*. 2016. URL: https://github.com/davidcarslaw/polarplotr.

[18] Davison, A. C. and Hinkley, D. V. *Bootstrap Methods and Their Applications*. ISBN 0-521-57391-2. Cambridge: Cambridge University Press, 1997. URL: http://statwww.epfl.ch/davison/BMA/.

[19] Canty, A. and Ripley, B. D. *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-18. 2016.

[20] Kariya, T. and Kurata, H. *Generalized least squares*. John Wiley & Sons, 2004.

[21] Yohai, V. J. High Breakdown-Point and High Efficiency Robust Estimates for Regression. *The Annals of Statistics* 15.2 (1987), pp. 642–656. URL: http://www.jstor.org/stable/2241331.

[22] Huber, P. J. Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics* (1973), pp. 799–821.

[23] Venables, W. N. and Ripley, B. D. *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer, 2002. URL: http://www.stats.ox.ac.uk/pub/MASS4.

[24] Allen, G., Sioutas, C., Koutrakis, P., Reiss, R., Lurmann, F. W., and Roberts, P. T. Evaluation of the TEOM method for measurement of ambient particulate mass in urban areas. *Journal of the Air & Waste Management Association* 47.6 (1997), pp. 682–689. DOI: 10.1080/10473289.1997.10463923. URL: http://dx.doi.org/10.1080/10473289.1997.10463923.

[25] Green, D. C., Fuller, G. W., and Baker, T. Development and validation of the volatile correction model for PM_{10}—An empirical method for adjusting TEOM measurements for their loss of volatile particulate matter. *Atmospheric Environment* 43.13 (2009), pp. 2132–2141. URL: http://www.sciencedirect.com/science/article/pii/S1352231009000557.

[26] NOAA. Integrated Surface Database (ISD). 2016. URL: https://www.ncdc.noaa.gov/isd.

[27] Carslaw, D. ***worldmet**: Import Surface Meteorological Data from NOAA Integrated Surface Database (ISD)*. R package version 0.7.5. 2017. URL: http://github.com/davidcarslaw/worldmet.

[28] Buchanan, C., Beverland, I., and Heal, M. The influence of weather-type and long-range transport on airborne particle concentrations in Edinburgh, UK. *Atmospheric Environment* 36.34 (2002), pp. 5343–5354. URL: http://www.sciencedirect.com/science/article/pii/S1352231002005794.

[29] Abdalmogith, S. S. and Harrison, R. M. The use of trajectory cluster analysis to examine the long-range transport of secondary inorganic aerosol in the UK. *Atmospheric Environment* 39.35 (2005), pp. 6686–6695. URL: http://www.sciencedirect.com/science/article/pii/S1352231005006825.

[30] Liu, Y.-J. and Harrison, R. M. Properties of coarse particles in the atmosphere of the United Kingdom. *Atmospheric Environment* 45.19 (2011), pp. 3267–3276. DOI: 10.1016/j.atmosenv.2011.03.039. URL: http://www.sciencedirect.com/science/article/B6VH3-52G8GT9-1/2/1c4d705b78225fca7c930197cd80c35a.

[31] Querol, X., Alastuey, A., Ruiz, C., Artiñano, B., Hansson, H., Harrison, R., Buringh, E., Brink, H. ten, Lutz, M., Bruckmann, P., Straehl, P., and Schneider, J. Speciation and origin of PM10 and PM2.5 in selected European cities. *Atmospheric Environment* 38.38 (2004), pp. 6547–6555. URL: http://www.sciencedirect.com/science/article/pii/S1352231004008143.

[32] Charron, A. and Harrison, R. M. Fine ($PM_{2.5}$) and Coarse ($PM_{2.5–10}$) Particulate Matter on A Heavily Trafficked London Highway: Sources and Processes. *Environmental Science & Technology* 39.20 (2005), pp. 7768–7776. DOI: 10.1021/es050462i. URL: http://dx.doi.org/10.1021/es050462i.

[33] Stein, A. F., Draxler, R. R., Rolph, G. D., Stunder, B. J. B., Cohen, M. D., and Ngan, F. NOAA's HYSPLIT Atmospheric Transport and Dispersion Modeling System. *Bulletin of the American Meteorological Society* 96.12 (2015), pp. 2059–2077. DOI: 10.1175/BAMS-D-14-00110.1. URL: http://dx.doi.org/10.1175/BAMS-D-14-00110.1.

[34] Petzold, A., Kopp, C., and Niessner, R. The dependence of the specific attenuation cross-section on black carbon mass fraction and particle size. *Atmospheric Environment* 31.5 (1997), pp. 661–672. URL: http://www.sciencedirect.com/science/article/pii/S1352231096002452.

[35] Viidanoja, J., Sillanpää, M., Laakia, J., Kerminen, V.-M., Hillamo, R., Aarnio, P., and Koskentalo, T. Organic and black carbon in $PM_{2.5}$ and PM_{10}: 1 year of data from an urban site in Helsinki, Finland. *Atmospheric Environment* 36.19 (2002), pp. 3183–3193. URL: http://www.sciencedirect.com/science/article/pii/S1352231002002054.

[36] Ruellan, S. and Cachier, H. Characterisation of fresh particulate vehicular exhausts near a Paris high flow road. *Atmospheric Environment* 35.2 (2001), pp. 453–468. URL: http://www.sciencedirect.com/science/article/pii/S1352231000001102.

[37] Giorio, C., Tapparo, A., Dall'osto, M., Beddows, D. C. S., Esser-Gietl, J. K., Healy, R. M., and Harrison, R. M. Local and Regional Components of Aerosol in a Heavily Trafficked Street Canyon in Central London Derived from PMF and Cluster Analysis of Single-Particle ATOFMS Spectra. *Environmental Science & Technology* 49.6 (2015), pp. 3330–3340. DOI: 10.1021/es506249z. URL: http://dx.doi.org/10.1021/es506249z.

[38] Koenker, R. and Bassett Jr, G. Regression quantiles. *Econometrica: journal of the Econometric Society* 46.1 (1978), pp. 33–50.

Chapter 3

Air quality data — smonitor Europe

3.1 Introduction

Ambient air quality monitoring data for all European Union (EU) member states and many other cooperating countries are freely available to the public. There are two main repositories containing these data which are maintained by the European Environment Agency (EEA) for the compliance to the air quality reporting directives 2004/107/EC and 2008/50/EC.[1,2] The first repository is called AirBase, which contains data between 1969 and 2012 inclusive.[3,4] From 2013 onwards, the AirBase reporting system was superseded by what is known as Air Quality e-Reporting (AQER).[5] AQER forms a small piece of the larger EEA central data repository (CDR) which contains data and information for a very wide range of topics for the EU and its member states. Although these data are freely available, it can be difficult for data users to access the air quality data quickly and efficiently. This is especially true for the newer AQER system because of the complicated and convoluted data models and the XML file format which has been chosen for transmission.

The lack of ease of access resulted in the development of **smonitor** Europe, a user-focused relational database which allows access to European

air quality data in a quick and efficient manner.[6] The current database contains 12 800 monitoring sites, approximately 170 000 unique time series, 4.1×10^9 observations, and consumes 413 GB of disc space (including indices). An interactive map of the sites contained in the database is maintained for public viewing[7] and the sites are also mapped in Figure 3.1. This chapter documents the database's components and data sources.

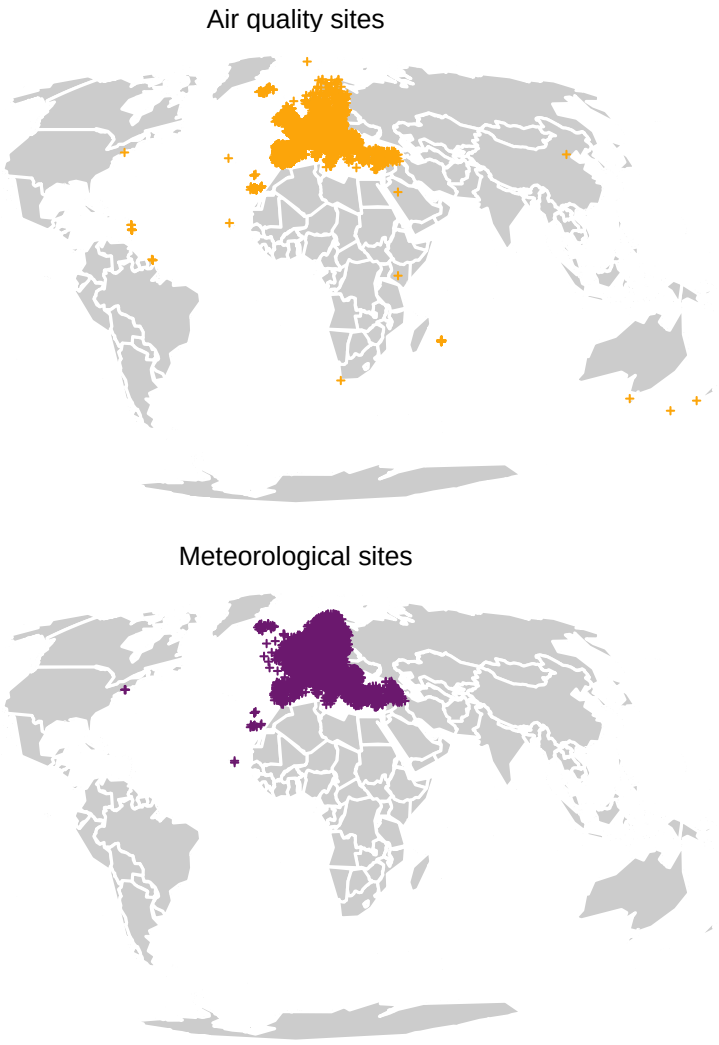

Figure 3.1: **smonitor** Europe's monitoring sites map split by their data type. As the database's name suggests, most data are from Europe, but there are data from monitoring sites located in other regions.

3.2 The smonitor framework

smonitor refers to a framework, a data model, and a collection of functions which operate on this specific data model.[6] The **smonitor** data model has been designed for simplicity so getting data into and out of the database is as easy as possible. Despite the simplicity, all air quality observations can be related back to a data source and all of the granularity in the AirBase and AQER data models are preserved so very little data is lost. PostgreSQL is the database technology used for **smonitor** Europe and the functions are written in R and contained within a portable R package.[6,8]

The core data model of **smonitor** is formed by five database tables: `sites`, `processes`, `observations`, `aggregations`, and `summaries`. However, in the case of **smonitor** Europe, the `aggregations` and `summaries` tables are not used because these tables contain data on how to calculate and perform aggregations on observations once they have been inserted. These actions are not applicable in the European case because, currently, the database only serves a storage function. The tables are self explanatory with the `sites` table containing information about air quality monitoring sites/stations/facilities, `processes` contains the details about the time series associated at the sites, and the `observations` table contains the dates, validity, and values of the observations/measurements.

In the **smonitor** nomenclature, a process is an important identifier to understand. A process is best defined as a unique, and usually, an uninterrupted time series. For any given site, there will generally be a number of processes representing the different variables (pollutants or surface meteorological measures) monitored at the location. The simplest definition of a process could therefore be a location-variable pair. However, there can be further granulation of a single variable based on different data sources, aggregation methods, or in the case of AQER, different combinations of inlets, sampling points, and sampling point processes. An example of processes in **smonitor** Europe can be displayed by using an example site: `ch0010a`, a

monitoring site located in Zürich city (Switzerland) called Zürich-Kaserne. The site's entry in the `sites` table includes what is displayed in Table 3.1 (there are however more variables contained in the database).

Table 3.1: A selection of site information in **smonitor** Europe for the ch0010a (Zürich-Kaserne) monitoring site.

site	site_name	latitude	longitude	site_type	date_start	date_end
ch0010a	Zürich-Kaserne	47.38	8.53	background	1992-01-01	2017-12-31

This site has data from 1992 to the end of 2017 which means it will be present in both the AirBase and AQER data repositories. For this site, the `processes` table includes the processes displayed in Table 3.2. There are a total of 73 processes associated with this particular site (not all are shown) which are referenced by integer keys. The variables are replicated because there are multiple aggregation periods and there are observations from two data sources (`airbase` and `aqer`). Unlike the AirBase data, the AQER data no longer contains the aggregations of observations because in the more recent years only the "primary" measurements are reported.

The different periods forming separate processes for a single variable is non-standard in the **smonitor** framework. This is because multiple aggregation methods/summaries usually hang off a single process. In the **smonitor** Europe case, using multiple processes is an easier way to handle the many aggregations and makes the database "flatter". There are no duplicate sites in **smonitor** Europe, but there has been no attempt to map the two different data sources together to collapse the multiple processes for one variable, for one aggregation period, for a site. It is more appropriate to keep the two data sources separate and allow for the importing functions to abstract this feature from the user.

The table which contains the observations/measurements is named `observations`. The `observations` table is very simple and contains seven columns, all of which are integer or numeric data types. There are three date variables: `date_insert` (observation insertion date), `date` (start date of

Table 3.2: A select number of processes in **smonitor** Europe for ch0010a (Zürich-Kaserne) monitoring site.

process	site	variable	period	date_start	date_end	data_source
14217	ch0010a	as_in_pm10	year	2006-01-01 00:00:00	2012-01-01 00:00:00	airbase
14218	ch0010a	bap_in_pm10	year	2006-01-01 00:00:00	2012-01-01 00:00:00	airbase
14221	ch0010a	cd_in_pm10	year	1997-01-01 00:00:00	2012-01-01 00:00:00	airbase
14226	ch0010a	ni_in_pm10	year	2006-01-01 00:00:00	2012-01-01 00:00:00	airbase
14229	ch0010a	no2	day	1992-01-01 00:00:00	2012-12-31 00:00:00	airbase
14230	ch0010a	no2	hour	1992-01-01 00:00:00	2012-12-31 23:00:00	airbase
14231	ch0010a	nox	day	1992-01-01 00:00:00	2012-12-31 00:00:00	airbase
14232	ch0010a	nox	hour	1992-01-01 00:00:00	2012-12-31 23:00:00	airbase
14233	ch0010a	o3	day	1992-01-01 00:00:00	2012-12-31 00:00:00	airbase
14234	ch0010a	o3	dymax	1992-01-01 00:00:00	2012-12-31 00:00:00	airbase
14235	ch0010a	o3	hour	1992-01-01 00:00:00	2012-12-31 23:00:00	airbase
14236	ch0010a	o3	hour8	1992-01-01 00:00:00	2012-12-31 23:00:00	airbase
14237	ch0010a	pb_in_pm10	year	1997-01-01 00:00:00	2012-01-01 00:00:00	airbase
14238	ch0010a	pm10	day	1997-01-01 00:00:00	2012-12-31 00:00:00	airbase
14239	ch0010a	pm2.5	day	1998-01-01 00:00:00	2012-12-31 00:00:00	airbase
14242	ch0010a	spm	day	1992-01-01 00:00:00	1996-12-31 00:00:00	airbase
14874	ch0010a	as_in_pm10	var	2012-12-31 23:00:00	2012-12-31 23:00:00	aqer
14875	ch0010a	bap_in_pm10	var	2012-12-31 23:00:00	2012-12-31 23:00:00	aqer
14877	ch0010a	cd_in_pm10	var	2012-12-31 23:00:00	2012-12-31 23:00:00	aqer
14879	ch0010a	ni_in_pm10	var	2012-12-31 23:00:00	2012-12-31 23:00:00	aqer
14881	ch0010a	no2	hour	2012-12-31 23:00:00	2013-12-31 22:00:00	aqer
14882	ch0010a	nox	hour	2012-12-31 23:00:00	2013-12-31 22:00:00	aqer
14883	ch0010a	o3	hour	2012-12-31 23:00:00	2013-12-31 22:00:00	aqer
14884	ch0010a	pb_in_pm10	var	2012-12-31 23:00:00	2012-12-31 23:00:00	aqer
14885	ch0010a	pm10	day	2012-12-31 23:00:00	2013-12-30 23:00:00	aqer
14886	ch0010a	pm2.5	day	2013-01-03 23:00:00	2013-12-29 23:00:00	aqer
15171	ch0010a	as_in_pm10	year	2013-12-31 23:00:00	2015-12-31 23:00:00	aqer
15172	ch0010a	bap_in_pm10	year	2013-12-31 23:00:00	2015-12-31 23:00:00	aqer
15174	ch0010a	cd_in_pm10	year	2013-12-31 23:00:00	2015-12-31 23:00:00	aqer
15175	ch0010a	ni_in_pm10	year	2013-12-31 23:00:00	2015-12-31 23:00:00	aqer
15177	ch0010a	no2	hour	2013-12-31 23:00:00	2017-12-31 22:00:00	aqer
15178	ch0010a	nox	hour	2013-12-31 23:00:00	2017-12-31 22:00:00	aqer
15179	ch0010a	o3	hour	2013-12-31 23:00:00	2017-12-31 22:00:00	aqer
15180	ch0010a	pb_in_pm10	year	2013-12-31 23:00:00	2015-12-31 23:00:00	aqer
15181	ch0010a	pm10	day	2013-12-31 23:00:00	2015-12-30 23:00:00	aqer
15182	ch0010a	pm2.5	day	2014-01-02 23:00:00	2017-12-30 23:00:00	aqer
15291	ch0010a	pm10	day	2015-12-31 23:00:00	2016-12-30 23:00:00	aqer
15374	ch0010a	as_in_pm10	year	2016-12-31 23:00:00	2016-12-31 23:00:00	aqer
15375	ch0010a	bap_in_pm10	year	2016-12-31 23:00:00	2016-12-31 23:00:00	aqer
15377	ch0010a	cd_in_pm10	year	2016-12-31 23:00:00	2016-12-31 23:00:00	aqer
15378	ch0010a	ni_in_pm10	year	2016-12-31 23:00:00	2016-12-31 23:00:00	aqer
15379	ch0010a	pb_in_pm10	year	2016-12-31 23:00:00	2016-12-31 23:00:00	aqer
15380	ch0010a	pm10	day	2016-12-31 23:00:00	2017-12-30 23:00:00	aqer

observation), and `date_end` (end date of observation) which are all stored in Unix time (number of seconds since January 1, 1970). The integer date formats are time zone agnostic and makes `SQL BETWEEN` clauses efficient when used within `SELECT` statements. The conversion of the Unix time variables to date-time data classes is handled by the importing functions and allow for time zone logic to be applied by the user and client during import. A limitation of the **smonitor** data model is that values can only take numeric values. This has not been an issue to-date, but an additional table and relationship will need to be designed in the future to be able to store non-numeric values if this is desired.

Table 3.3: Ten rows of the **smonitor** Europe `observations` table. The process is `ch0010a`'s (Zürich-Kaserne's) hourly O_3 process sourced from the AQER repository (Table 3.2).

date_insert	date	date_end	process	summary	validity	value
1536241368	1388530800	1388534400	15179	1	1	2.88
1536241368	1388534400	1388538000	15179	1	1	2.04
1536241368	1388538000	1388541600	15179	1	1	2.06
1536241368	1388541600	1388545200	15179	1	1	32.28
1536241368	1388545200	1388548800	15179	1	1	31.72
1536241368	1388548800	1388552400	15179	1	1	15.63
1536241368	1388552400	1388556000	15179	1	1	20.57
1536241368	1388556000	1388559600	15179	1	1	12.65
1536241368	1388559600	1388563200	15179	1	1	16.60
1536241368	1388563200	1388566800	15179	1	1	21.12

3.3 Importing data

The **smonitor** R package contains two main importing functions which are used to fetch data from the database and make them available for an R user.[6] These functions are `import_by_process` and `import_by_site`. Both of these functions take arguments which are translated to SQL statements

which are sent to the database service and R's equivalent of a database table, a data frame, is returned to the user. The `import_by_process` function allows for more flexibility, but generally the `import_by_site` is used by a user. An example using the Zürich-Kaserne (ch0010a) site is shown in Listing 3.1.

Listing 3.1: An example of using **smonitor** Europe to import a site's hourly monitoring data. The con object is a previously created database connection.

```
1 # Get all hourly data for the Zurich-Kaserne site
2 data_zurich <- import_by_site(
3   con,
4   site = "ch0010a",
5   period = "hour"
6 )
```

This function call returns 1 857 985 observations and on my system using the remote PostgreSQL database server, it takes about 15 seconds to run. `import_by_site` allows for many other options such as returning only specific periods of data between dates, selected aggregation periods, is able to return multiple sites with one function call, and can return reshaped data which are ready for direct use in data analysis packages such as **openair**.[9] An example of such a use with some optional arguments is displayed in Listing 3.2.

Listing 3.2: An example of using **smonitor** Europe to import two sites' hourly NO_x monitoring data with the use of many optional arguments.

```
# Get hourly data for the Zurich-Kaserne and
# Zurich-Stampfenbachstrasse sites
data_zurich_two <- import_by_site(
  con,
  site = c("ch0010a", "ch0013a"),
  variable = c("no", "no2", "nox"),
  period = "hour",
  spread = TRUE,
  tz = "Etc/GMT-1",
  start = 2010,
  end = 2018
)
```

3.4 Data sources

3.4.1 AirBase

AirBase Version 8 data were downloaded in bulk from the European Environment Agency data portal and the observations and aggregations are available per country/state as `.zip` archives.[3] The files containing the observational data are supplied in simple, tabular text files which were easy to clean and prepare for database inserts. The file names were decoded because these contained the site, observed property (an integer which represents a pollutant), and averaging period information which was not part of the tabular data proper. The different averaging period files have different formats and this is the only inconsistency across the data files. Two helper tables were needed too, the station and measurement configurations files which are approximately equivalent to the **smonitor** `sites` and `processes` tables. These metadata tables contained the extra variables needed to build a complete data model. Because the AirBase data repository has

been replaced with AQER, it is likely that these data will remain static into the future unless a large systematic error is discovered and needs to be addressed. There are a few known outstanding issues with the AirBase data. The two most notable is that for a few member states, the 2012 year is completely absent from the repository, for example Germany and France and the time zone used for each member state is undocumented.

3.4.2 Air Quality e-Reporting (AQER)

The AQER data were a far greater challenge to prepare for database insertion compared to the AirBase data files. The AQER data model is far more comprehensive and convoluted than the AirBase system and many issues were encountered with the data files across the many member states. Many of the issues found were due to the challenges of representing normalised and tabular data in deeply nested XML files which despite strict schemas being defined, much diversity was encountered which required significant software development to handle correctly.

The AQER system is composed of a large number of observational units which are termed "data flows". For example Data Flow B contains information about zones and agglomerations, Data Flow D contains data about monitoring activities, and Data Flow E1a contains the observational data. All XML files representing the data flows contain at least two observational units; a header and a data unit. For example, a Data Flow D document containing metadata of monitoring activities contains more than two units and usually contains these six components:

Header A document preamble indicating what the other units contain and what schema is being used.

Network Information about the monitoring network.

Station Information about the monitoring stations/sites/facilities.

Sampling point Information about the variables/pollutants being monitored.

Sample Information about the inlets used for the sampling points.

Sampling point process Information about the monitors/analyzers used for the sampling points.

Functions and programmes were developed to transform all member states' AQER XML documents within the B (zones), D (monitoring activities), and E1a (observations) data flows into tabular data. The zones files are spatial of nature and many member states have chosen not to use XML (or GML) for their submissions and use other spatial data files such as shapefiles in their place until the tools exist to easily generate the custom XML files.

The downloading and selection of the AQER XML documents is an imperfect process. Currently, a helper table is manually maintained which contains the URLs of the documents and some metadata about the files which are used to populate **smonitor** Europe. Often, multiple revisions of the same XML documents are uploaded to the CDR repository due to mistakes or errors which are found by the member state or the EEA checking processes. Unfortunately, there is little clarity on what has been changed with the new revisions and what document the revision supersedes. Sometimes, only pieces of the documents are revised which makes it extremely difficult to understand what needs to be done to gain a complete and correct data set. Examples of this is the 2015 Belgian E1a submission which contains three revisions and the 2015 Polish E1a submission which at first contained one XML document, but the revision contained multiple documents. Despite these issues, all efforts are taken to create a correct source data table which can be modified if improvements need to be made in the future.

When it is decided which AQER XML files are necessary, they are downloaded and the different observational units are parsed, cleaned, and trans-

formed into tabular data with R functions. After the tables are created, a significant challenge is determining the linkage between the station document within the measures (D) data flow and what is contained within the observation (E1a) documents. The station document's primary identifier is a prefixed code. For example, a station in Macedonia is represented as `STA-MK0001A`. There are some differences among the member states on the format and what the prefix is, but they are minor. However, in the observation documents, with the exception of the United Kingdom, the station is not directly referenced. To create the key to link the station and observation documents therefore requires parsing of the `observation_id`, `feature_of_interest`, `procedure`, or `sampling_point` XML nodes. Among the member states, there is much diversity and at times, for example Austria, the format of the observation XML nodes changes over time. In one example, the 2015 Czech Republic observation document, it was not possible to link the observations with the stations in the 2015 measures data flow which is clearly an error and therefore the observations could not be used at the time. This seems to be an oversight by the developers of the documents' schema because when building a relational data model, a critical component is how the observational units are linked and it should be simple for a data user to join the units together with a common variable (a key).

Once these issues are addressed, it becomes trivial to use these data in the **smonitor** framework. Site keys are assigned and the unsuffixed lower case version of the primary station identifier was used for this, for example see Table 3.1. For the insertion of observations, the process integer keys need to be joined to the tabular data and then inserted into the database. The process keys used for observations are constrained so they must be present in the `processes` table before they can be used in the `observations` table. When new AQER data becomes available, new sites and processes are created if necessary before the observations are inserted. The programmes which handle the transformations and cleaning contain a number of tests for these requirements and the database also enforces constraints to ensure

the **smonitor** data model's integrity. Most notably, the `site` → `process` → `observation` relationship must be correct to ensure the data model is valid.

3.4.3 NOAA's Integrated Surface Database

Many air quality data analysis situations require meteorological data as well as pollutant concentrations. To ensure surface meteorological data can be accessed alongside air quality observations, all Integrated Surface Database (ISD)[10] sites within an approximate European boundary which have air temperature, atmospheric pressure, relative humidity, wind speed, and wind direction observations have been inserted in **smonitor** Europe. There are 2 900 ISD sites serviced in **smonitor** Europe (Figure 3.1). The country naming convention used in the ISD do not conform to the ISO country codes or names and therefore new site codes were generated to maintain consistency with the European data sources.

3.4.4 Other data sources

Other people who have found **smonitor** Europe useful have often requested other data sources to be included so they can have a consistent data access API (application programming interface). I have been somewhat reluctant to include other data sources because of the effort required to maintain the various data sources by one person. However, the data cleaning, formatting and inserting functions have been written for the Centre for Environmental Data Analysis (CEDA)[11], EBAS[12] the World Data Centre for Greenhouse Gases (WDCGG)[13], Airparif (Paris's monitoring network)[14], and **openair**'s[15] access to monitoring data outside United Kingdom's Automatic Urban and Rural Network (AURN) data portals or products. There are a small amount of data from these data sources and about a few hundred other sites not contained in the European or ISD data repositories because of this (see Figure 3.1).

3.5 Outstanding issues

There are a handful of outstanding issues with **smonitor** Europe.

3.5.1 Source data issues

AirBase time zones The time zones used for the member states in the AirBase repository are undocumented. UTC has been used but this is likely to be incorrect for some member states' data and is rather difficult to test.

AirBase 2012 data Some member states observational data for 2012 are absent from the AirBase repository, for example Germany and France.

AQER E1a file history A major source of frustration is how the revisions of the E1a (observational) documents are handled by the member states. It is unclear in many situations what the revisions are and what data they replace. Decisions have been made to populate **smonitor** Europe, but it would be preferable to have a new, single file which completely supersedes the incorrect submissions when a revision is made.

NO_x, NO_2, and NO reporting Many member states no longer report the full complement of NO_x, NO_2, and NO. This is frustrating to air quality scientists because at least two of these need to be present to conduct many analyses.

Polish site names The special characters in the Polish site names have not been encoded correctly and therefore are incorrectly represented.

AQER zones (Data Flow B) Many of the member states' zones data files are invalid.

3.5.2 smonitor database issues

Validity The **smonitor** framework has a binary system to classify the validity of an observation: an observation is either valid or it is not. This

results in the more granular validity classifications such as "valid but below detection limit" used by AirBase and AQER to be lost in **smonitor** Europe.

Extra identifiers from AQER's Data Flow D Although the AQER sampling point, feature of interest, and sampling point process identifiers are present in **smonitor** Europe `processes` table, the extra observational units (tables) have not been inserted.

3.6 Sharing data and the europeimportr package

The entire PostgreSQL **smonitor** Europe database could be exported or dumped and moved to another location for others to replicate and use. However, arguably, a better way to share these data is to provide pre-prepared data objects on a web server which can be retrieved and used quickly and easily. An example of this has been provided Ricardo Energy & Environment in the form of an R package called **europeimportr**.[16]

The **europeimportr** interface contains two components. The first, is a directory containing a large number of `.rds` files, a native R data format, which have been exported from **smonitor** Europe and have been uploaded to a web server. Observations have been exported for every site in the database, some aggregations have been calculated and are available, as are the `sites`, `processes`, and `zones` (Data Flow B) tables. The second component is a R package called **europeimportr** which contains a handful of simple functions which download, load, and return data to an R user. The **europeimportr** package makes it very easy and fast to access European air quality monitoring data and a similar interface may be released in the future to the public.

3.7 References

[1] European Parliament. Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. *Official Journal of the European Union* (2008). URL: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32008L0050.

[2] European Parliament. 2011/850/EU: Commission Implementing Decision of 12 December 2011 laying down rules for Directives 2004/107/EC and 2008/50/EC of the European Parliament and of the Council as regards the reciprocal exchange of information and reporting on ambient air quality (notified under document C(2011) 9068). *Official Journal of the European Union* (2011). Document 32011D0850. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32011D0850.

[3] European Environment Agency. *AirBase – The European air quality database (Version 8)*. 2014. URL: https://www.eea.europa.eu/data-and-maps/data/airbase-the-european-air-quality-database-8.

[4] Simoens, D. AirBase version 8 data products on EEA data service. European Environment Agency. 2014. URL: http://www.eea.europa.eu/data-and-maps/data/airbase-the-european-air-quality-database-8.

[5] European Environment Agency. *Eionet Central Data Repository*. 2017. URL: http://cdr.eionet.europa.eu/.

[6] Grange, S. K. ***smonitor**: A framework and a collection of functions to allow for maintenance of air quality monitoring data*. 2018. URL: https://github.com/skgrange/smonitor.

[7] Grange, S. K. smonitor Europe air quality monitoring sites. 2018. URL: http://skgrange.github.io/www/maps/smonitor_europe_sites/smonitor_europe_sites.html.

[8] PostgreSQL Global Development Group. PostgreSQL. Version 9.5. URL: https://www.postgresql.org/.

[9] Carslaw, D. C. and Ropkins, K. *openair — An R package for air quality data analysis*. *Environmental Modelling & Software* 27–28.0 (2012), pp. 52–61. URL: http://www.sciencedirect.com/science/article/pii/S1364815211002064.

[10] NOAA. Integrated Surface Database (ISD). 2016. URL: https://www.ncdc.noaa.gov/isd.

[11] Centre for Environmental Data Analysis. Centre for Environmental Data Analysis (CEDA). Data and information services for environmental science. 2017. URL: http://www.ceda.ac.uk/.

[12] Norwegian Institute for Air Research. EBAS. 2017. URL: http://ebas.nilu.no/.

[13] Japan Meteorological Agency. World Data Centre for Greenhouse Gases (WDCGG). 2017. URL: http://ds.data.jma.go.jp/gmd/wdcgg/.

[14] Airparif. *Association de surveillance de la qualité de l'air en Île-de-France*. http://www.airparif.asso.fr/.

[15] Carslaw, D. and Ropkins, K. *openair: Open-source tools for the analysis of air pollution data*. 2015.

[16] Grange, S. *europeimportr: Tools to Import European Air Auality Monitoring Data Sourced From the AirBase and Air Quality e-Reporting Repositories*. 2017.

Chapter 4

European vehicular primary NO_2 trends

This work was originally published in *Nature Geoscience* on 27 November, 2017 and was featured on the cover of the journal's issue (Figure 4.1).[†]

Figure 4.1: *Nature Geoscience* Volume 10 Number 12 cover.

[†]https://doi.org/10.1038/s41561-017-0009-0

4.1 Abstract

Many European countries do not currently meet legal air quality standards for ambient nitrogen dioxide (NO_2) near roads; a problem that has been forecast to persist to 2030. Whereas European air quality standards regulate NO_2 concentrations, emissions standards for new vehicles instead set limits for NO_x – the combination of nitric oxide (NO) and NO_2. From around 1990 onwards, total emissions of NO_x declined significantly in Europe, but roadside concentrations of NO_2 – a regulated species – declined much less than expected. This discrepancy has been attributed largely to the increasing usage of diesel vehicles in Europe and more directly-emitted tailpipe NO_2. Here we apply a data filtering technique to 130 million hourly measurements of NO_x, NO_2 and ozone (O_3) from roadside monitoring stations across 61 urban areas in Europe over the period 1990 to 2015 to estimate the continent-wide trends of directly emitted NO_2. We find that the ratio of NO_2 to NO_x emissions increased from 1995 to around 2010 but has since stabilised at a level that is substantially lower than is assumed in some key emissions inventories. The proportion of NO_x now being emitted directly from road transport as NO_2 is up to a factor of two smaller than the estimates used in policy projections. We therefore conclude that there may be a faster attainment of roadside NO_2 air quality standards across Europe than is currently expected.

4.2 Introduction

Since the mid-1990s the European vehicle fleet has undergone considerable dieselisation[1–4] with incentivisation over other fuels and technologies on the basis of predicted fuel efficiency, lower CO_2 emissions, and increased driving performance.[5–7] By 2014 diesel vehicles accounted for an average of 53 % of new European passenger vehicle sales compared to 14 % in 1990, in contrast to little increase in their adoption into US fleets.[3,4]

The proportion of diesel powered vehicles across Europe has contributed to widely published problems where legal ambient air quality standards are breached, usually near roads. Of particular concern in recent years is nitrogen dioxide (NO_2) although particulate matter (PM) is also important.[8] Many European Union (EU) member states are struggling to comply with the 2008/50/EC Air Quality Directive which sets legal limits for hourly and annual average NO_2 concentrations.[8–10] While total national emissions of NO_x (NO + NO_2) have shown reductions in Europe, urban concentrations of NO_2 have decreased less than expected and this has been attributed to the growth in diesel fuelled vehicles.[11–19]

The impacts on public health of NO_2 are significant both through direct harm on inhalation and as a precursor to secondary pollutants ozone (O_3) and PM.[20] Published estimates of premature deaths due to NO_2 in 28 EU countries were reported to be 72 000 annually, based on a 2012 analysis year.[21] Roadside locations are perhaps the most important places where NO_2 must be controlled because this is where human exposure is at its highest. These are challenging locations from a legal compliance perspective — of all the reported exceedances of EU hourly and annual limit values in 2016, 94 % of those occurred at roadside monitoring locations.[22]

NO_2 concentrations at roadside locations are primarily controlled by local road transport and are influenced by, firstly, the total amount of NO_x emitted and then the fraction of that NO_x that is directly emitted as NO_2.[23] A shift towards higher NO_2/NO_x emissions from road transport can lead to a counter intuitive situation where total NO_x emissions can fall over time, yet roadside concentrations of NO_2 do not decline. The influence of this key ratio in driving trends and forecasts has already been shown in central London.[16] Predictions of future NO_2 concentrations in Europe must make assumptions about this NO_2/NO_x ratio, and predicted increases in this ratio are in part, behind a predicted lack of air quality standard attainment in many cities until 2025–2030.[15] Despite the critical importance of the NO_2/NO_x ratio in controlling urban roadside concentrations, specific lim-

its do not exist as part of European vehicular emission standards tests. New European vehicle tests report only total NO_x ($NO + NO_2$) in exhaust gases and whilst emission standards set limits for total NO_x they do not speciate between NO and NO_2. Beyond initial new vehicle tests little is known about how technologies such as diesel oxidation catalysts (DOC) and diesel particulate filters (DPF) influence this ratio in the real-world, despite the high profile given to the topic since the Volkswagen (VW) emissions scandal.[7,24] The implications of not correctly estimating NO_2/NO_x ratios in policy support tools such as COPERT and HBEFA (Handbook Emission Factors for Road Transport) have been described by others.[25–28]

Although recent NO_x emission underestimates from passenger cars have received most media attention, other vehicles such as heavy duty vehicles (HDVs) and buses are also important in controlling roadside NO_2 because they are predominately diesel fuelled. In this study, which focuses on NO_2 trends in urban areas, it is expected that light duty vehicles (LDVs) and urban buses will make significant contributions to vehicle emissions. It should also be noted that in terms of emissions data availability there is considerably more information available on passenger cars compared with other types of vehicles. As a consequence, there is uncertainty in both the absolute and relative contributions to NO_x and NO_2 from these additional transport sources.

The NO_2/NO_x ratio from diesel vehicles is controlled by both engine and exhaust control technologies that have advanced in response to the "Euro" series of emissions standards. The introduction of Euro 3 in 2000 saw the introduction of DOC into passenger vehicles; where in the presence of excess oxygen, NO can be oxidised to NO_2 over DOC metal catalysts resulting in more direct NO_2 being emitted.[16,29,30] The introduction of DPF in 2009 for compliance with the Euro 5 emission standards introduced a further technology that could lead to additional direct tailpipe NO_2.[31] However, as each progressive Euro standard has been introduced there have been no systemic observations of how new exhaust technologies might affect the

NO_2/NO_x ratio in real world emissions, or evaluation of whether the emissions inventories that need this ratio for forecasts, and that unpin policy, are preforming well.

4.3 Methods

4.3.1 Data

The primary data sources for the air quality data used in this study were the European Environment Agency (EEA) AirBase and air quality e-Reporting (AQER) data repositories.[32,33] These two repositories cover all European Union (EU) member states and other cooperating countries such as those in the European Economic Area (EEA) and Switzerland. The AirBase repository contains observational data during 1969–2012 but from 2013 onwards, the AirBase system was superseded with the more comprehensive AQER reporting system. AQER uses new data vocabulary, file formats, and requires EEA member states to report a range of observational units called "data flows" which were not required for AirBase. The AQER system uses the XML (Extensible Markup Language) file format to transfer data but it is common for other file formats to be used alongside XML for some data flows.

The AirBase and AQER data were cleaned and inserted into a single database with a simple data model.[34] The AirBase data are available in well-formatted tabular text files which only required decoding of their file names to be used. However, the AQER XML, documents were a far greater challenge due to the need to parse different observational units to create a coherent and decoded data model. Despite AQER formalising XML schemas, many variations were found across the member states' files which required significant development to ensure that the variations were handled correctly.

The database was also supplemented with other data where available.

London for example, has a much larger air quality monitoring network which is not represented by AirBase and the AQER repositories because these monitoring activities are coordinated by other bodies and do not form part of the national network. Therefore, these additional sites and data were accessed using **openair**, which accesses data from King's College London.[35,36] These additional sites follow equivalent quality assurance and quality procedures as the national network. Many countries have not reported the full complement of NO, NO_2, and NO_x presumably due to a lack of a legal obligation and file size concerns. The analysis reported here required both hourly NO_2 and NO_x to be present for a monitoring site and therefore the missing variables were derived from the other components if possible. In the case of Paris, the additional NO_x was accessed through the Airparif web portal.[37] Once the cleaning and tidying was complete, the database contained 2.7×10^9 observations from 8 400 air quality monitoring sites.[34,38]

The data import, transformation, and tidying was conducted with R and the database technology used was PostgreSQL.[39,40] NO_x data spanned from 1973 to 2015, but the analysis focused on years between 1990 and 2015 when the operation of chemiluminescent NO_x instrumentation was wide-spread throughout Europe.

4.3.2 NO_x filtering method

To isolate the primary NO_2 component, a multi-step filtering process was conducted which was similar to past calculation of CO/NO_x ratios by other authors (for examples see Parrish et al. [41] and Hassler et al. [42]). The first step was to choose urban areas and these were generally identified by the European Commission's Functional Urban Area definition.[43] A Functional Urban Area includes a city and their commuting zones, which is approximately equivalent to a metropolitan area. The spatial boundaries (polygons) for these urban areas were obtained from the AQER zones data flow which form the official EU air quality management zones. When the polygons were

not available or not suitable for use in the AQER repository, the appropriate administrative boundaries were scraped from OpenStreetMap.[44,45] These polygons were then used as a spatial boundary for an urban area and only monitoring sites within the boundary were selected and used. Seventy-six urban areas were identified and used but after the filtering process, 61 urban areas had the variables and volume of data needed for the analysis. An European urban area map can be found in Figure 4.2.

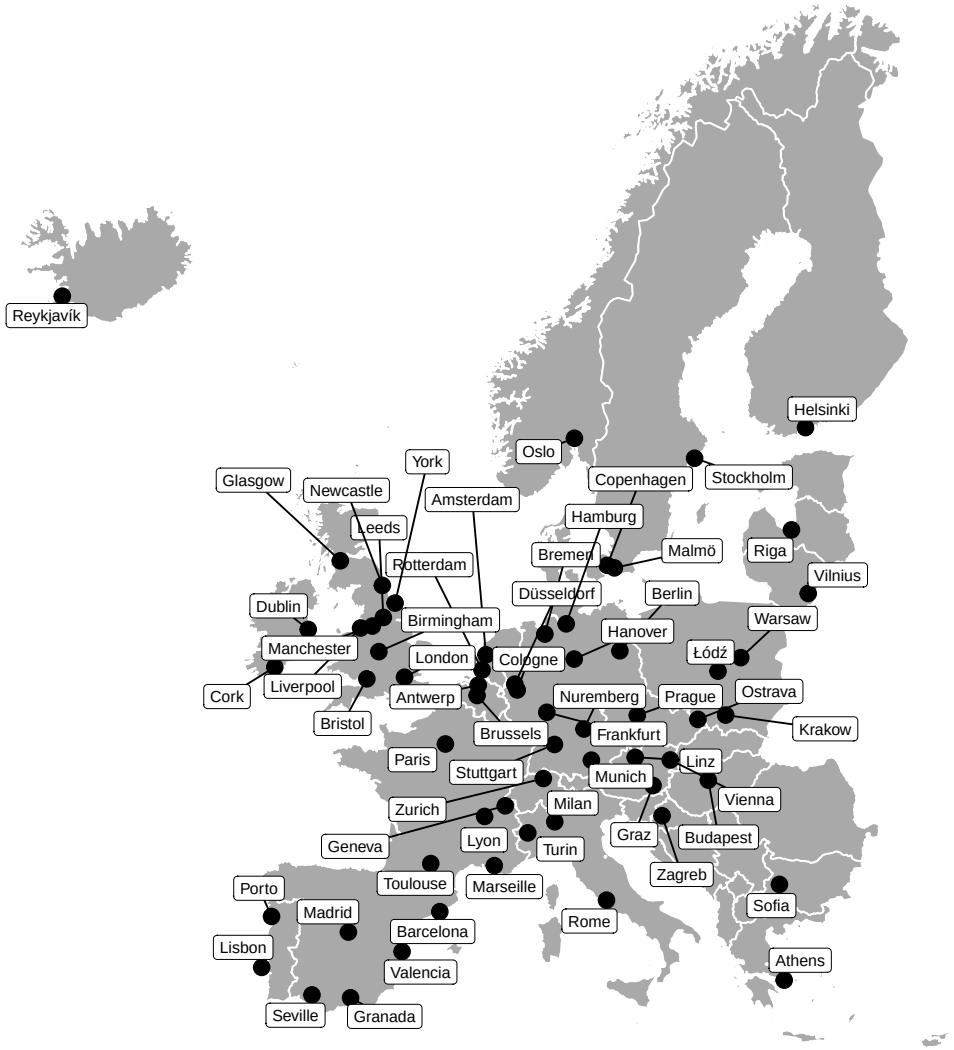

Figure 4.2: The 61 European urban areas which were tested for changes in their NO_x/NO_2 ratios using filtering methods. Internal polygons/lines indicate country boundaries.

For each urban area that was defined with a boundary, a representative

Table 4.1: Details for the 61 European urban areas analysed in this analysis.

Country	Urban area	Population	Latitude	Longitude	O_3 site
Austria	Vienna	2179769	48.16	16.53	at90lob
Austria	Linz	193814	48.27	14.31	at4s416
Austria	Graz	269997	47.08	15.44	at60018
Belgium	Antwerp	1200000	51.21	4.43	betr801
Belgium	Brussels	1800663	50.80	4.36	betr012
Bulgaria	Sofia	1263807	42.67	23.40	bg0052a
Croatia	Zagreb	792875	45.76	16.01	hr0009a
Czech Republic	Prague	1964750	50.01	14.45	cz0alib
Czech Republic	Ostrava	1153876	49.84	18.26	cz0toff
Denmark	Copenhagen	1806667	55.70	12.55	dk0030a; dk0045a
Finland	Helsinki	1224107	60.29	25.04	fi00208; fi00425
France	Paris	11089124	48.83	2.36	fr04037
France	Marseille	1831500	43.53	5.44	fr03029
France	Lyon	1717300	45.74	4.83	fr20017
France	Toulouse	1052497	43.62	1.44	fr12021
Germany	Hamburg	3134620	53.56	9.97	dehh008
Germany	Berlin	4971331	52.54	13.35	debe010
Germany	Munich	2531706	48.15	11.55	deby039
Germany	Cologne	1873580	51.02	6.88	denw053
Germany	Frankfurt	2517561	50.13	8.75	dehe008
Germany	Stuttgart	2663660	48.81	9.23	debw013
Germany	Hanover	1294447	52.23	10.47	deni011
Germany	Bremen	1249291	53.18	8.63	dehb004
Germany	Dusseldorf	1525029	51.25	6.73	denw071
Germany	Nuremberg	1288797	49.45	11.09	deby053
Greece	Athens	4013368	38.08	23.70	gr0027a
Hungary	Budapest	2393846	47.31	18.92	hu0032a
Iceland	Reykjavik	130345	64.14	-21.77	is0004a; is0005a
Ireland	Dublin	1535446	53.35	-6.28	ie0028a; ie0140a
Ireland	Cork	399216	51.92	-8.38	ie0091a
Italy	Rome	3457690	41.93	12.51	it0953a
Italy	Milan	3076643	45.50	9.25	it1017a; it1650a
Italy	Turin	1745221	45.01	7.55	it1120a
Latvia	Riga	1003949	56.94	24.16	lv0rke2
Lithuania	Vilnius	806404	54.69	25.21	lt00002
Netherlands	Amsterdam	1443258	52.39	4.92	nl00520; nl00564
Netherlands	Rotterdam	1186818	51.81	4.67	nl00441; nl00418
Norway	Oslo	1090513	59.95	10.49	no0081a
Poland	Warsaw	2660406	52.28	20.96	pl0044a
Poland	Lodz	1163516	51.76	19.53	pl0096a
Poland	Krakow	1725894	50.01	20.01	pl0011a ; pl0038a; pl0501a
Portugal	Lisbon	2435837	38.70	-9.21	pt03087
Portugal	Porto	1099040	41.16	-8.59	pt01028; pt01050
Spain	Madrid	5804829	40.42	-3.75	es1193a
Spain	Barcelona	4233638	41.39	2.01	es0694a; es1679a
Spain	Valencia	1564145	39.48	-0.37	es1619a; es1183a
Spain	Seville	1249346	37.34	-6.04	es1630a
Spain	Granada	237540	37.20	-3.61	es1560a; es1394a; es1924a
Sweden	Stockholm	1860872	59.32	18.06	se0022a
Sweden	Malmo	687481	55.61	13.00	se0001a
Switzerland	Zurich	1110478	47.38	8.53	ch0010a
Switzerland	Geneva	197376	46.20	6.13	ch0051a
United Kingdom	London	11917000	51.52	-0.21	gb0620a
United Kingdom	Birmingham	2357100	52.51	-1.83	gb0851a; gb0569a
United Kingdom	Leeds	2393300	53.80	-1.55	gb0584a
United Kingdom	Glasgow	1747100	55.86	-4.26	gb0641a; gb1028a
United Kingdom	Manchester	2539100	53.48	-2.24	gb0613a
United Kingdom	Bristol	1006600	51.46	-2.59	gb0585a; gb0884a
United Kingdom	Liverpool	1365900	53.41	-2.98	gb0594a; gb0777a
United Kingdom	Newcastle	1055600	54.98	-1.61	gb0568a
United Kingdom	York	204439	54.33	-0.81	gb0014r

ozone (O_3) background site was identified. The representative O_3 site had the requirements of having a continuous monitoring operation, *i.e.* not a seasonal site and having an hourly time series of at least five years. These O_3 time series were used to represent the typical urban background concentrations of O_3 for each urban area. In some situations, an unbroken time series was unavailable, usually due to monitoring site closures, therefore more than one representative O_3 site was used to gain a minimum of five years of O_3 data (Table 4.1). No data capture filters were applied to the observations. Sites classified as urban background were prioritised over other site types but for seven urban areas this was not possible and an industrial or roadside site was used. One-hundred and thirty million hourly measurements of NO_2, NO_x, and O_3 were evaluated from 488 sites. Details on the urban areas and the O_3 monitoring sites can be found in Table 4.1.

After a representative O_3 site was identified for an urban area, hourly NO_2 and NO_x observations from traffic, roadside, and kerbside sites where filtered to include only traffic-dominated periods between 06:00–18:00 (Co-ordinated Universal Time, Eastern European Time, or Central European Time depending on location; Table 3) for weekdays (Monday–Friday), and when the representative O_3 background concentrations were low. Low-O_3 conditions were considered when hourly concentrations were $\leq 10\,\mu g\,m^{-3}$ (5 ppb). The low-O_3 threshold was varied to determine the effect on the calculated ratio of NO_2 to NO_x. Varying the absolute value of the threshold between 5 and $30\,\mu g\,m^{-3}$ did not alter the patterns which were determined, only the absolute values of the NO_2/NO_x ratio due to an increase of contamination of non-primary NO_2 (Figure 4.3). The $10\,\mu g\,m^{-3}$ threshold allowed for more recent years with higher urban O_3 concentrations when compared to earlier time periods to have an adequate number of observations which could be used to estimate the NO_2/NO_x ratio which was not the case for the $5\,\mu g\,m^{-3}$ threshold.

The filtering process removed many of the total NO_2 and NO_x observations but had the goal of isolating the times when the influence of the NO

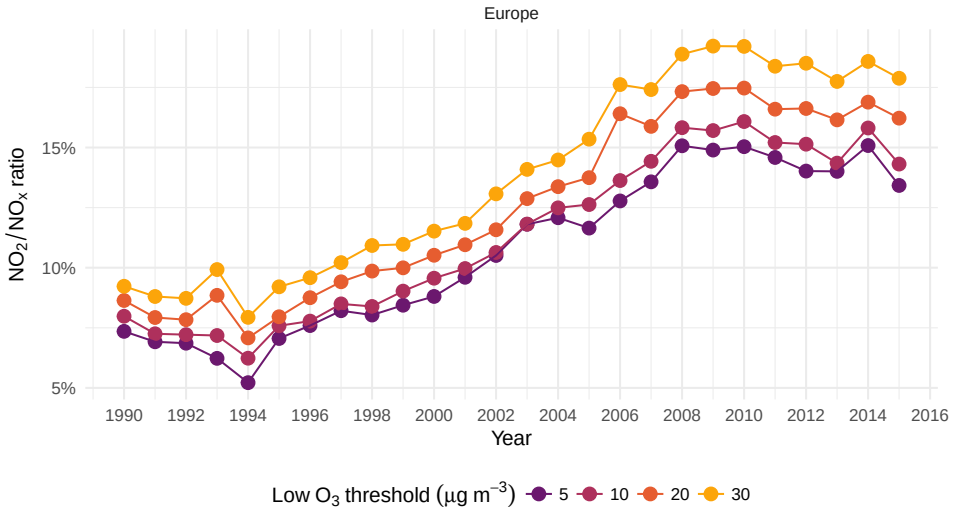

Figure 4.3: The influence of different low-O_3 thresholds on mean European NO_2/NO_x ratios. The pattern remains identical but the offset increases due to increased contamination of secondary NO_2 into the filtered data. The analysis reported used $10\,\mu g\,m^{-3}$ as the low-O_3 threshold.

+ O_3 reaction was negligible. These conditions would therefore represent those when the roadside increment in NO_2 above background would be dominated by primary NO_2 emissions from vehicles using the road. A potential source of uncertainty is the use of chemiluminescent NO_x analysers with molybdenum catalysts in most analysers for compliance monitoring. These instruments are affected by interference due to NO_y species, which are detected as NO_2. However, at roadside locations, and in particular for increments above local background concentrations with very little ageing of the airmass, the influence of NO_y species is expected to be negligible.[46] A potentially more important interferent is the direct emission of nitrous acid (HONO), which would also be detected as NO_2 in these instruments. Measurements of HONO in vehicle exhausts suggests only low amounts are emitted and its effect would be small. For example, Kirchstetter et al. [47] measured a $HONO/NO_x$ ratio of $2.9 \pm 0.5 \times 10^{-3}$.

4.3.3 NO_2/NO_x ratio estimation

After the filters had been applied, for each site and year combination, the NO_2/NO_x ratio was calculated with robust linear regression with an *MM*-estimator. The use of the linear model in this way allowed for the slope to be estimated, which represents an estimate of the the primary NO_2/NO_x ratio. The robust linear regression functions were provided with the **MASS** R package.[48] The robust regression technique is hardened against outliers by a high breakdown point which helped handle noisy observations before 2000 in some locations. When ratios were sequentially aggregated to urban area, country, and European level the arithmetic mean was used as the summary function. After the NO_2/NO_x ratio estimates were aggregated to European level, the trend was non-monotonic. The breakpoints in the trend were identified with the **segmented** R package and three linear least squares regression models were calculated to represent the pieces of the trend.[49,50]

4.3.4 Method validation

The filtering method employed was tested with a total oxidant ($OX = NO_2 + O_3$) method reported by Jenkin [51]. OX can be thought of as the sum of regional and local oxidant contributions at a monitoring site. Like the filtering method, if the OX method is applied to a roadside site, the local oxidant component can provide an estimate of the primary NO_2/NO_x ratio. Therefore the estimates of the filtering and OX methods can be directly compared. The OX method has the limitation of requiring O_3 observations as well as NO_x observations. However, the measurement of O_3 at roadside sites is uncommon. The two methods showed very good agreement and for London Marylebone Road, a monitoring site reported by Jenkin [51], the methods demonstrated near-equivalence for the years 1997–2014 (Figure 4.4).

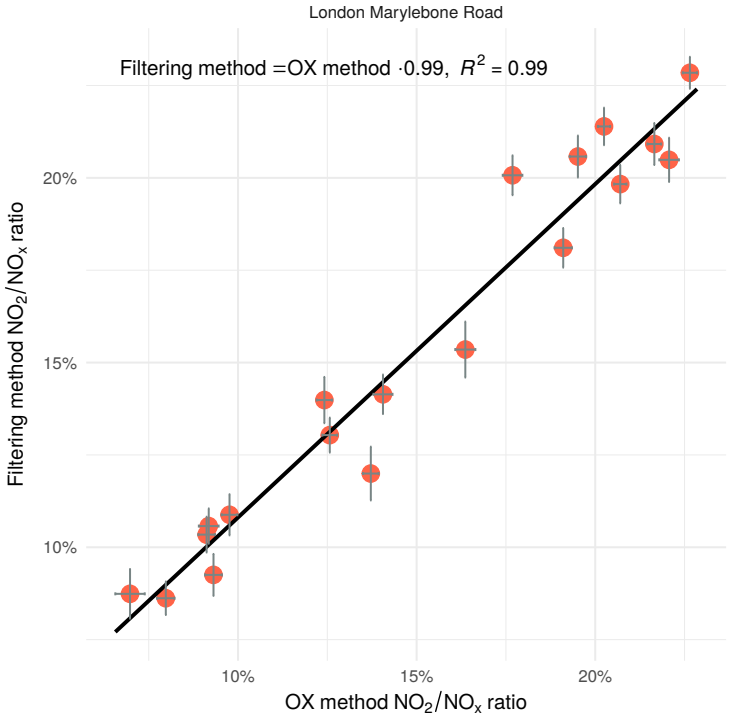

Figure 4.4: Relationship between two methods which estimate the annual NO_x/NO_2 ratio at London Marylebone Road between 1997 and 2015. The two methods show near-equivalence at this site, the fitted line represents a simple linear regression model with a forced zero intercept, and the error bars are the 95 % confidence intervals of the estimates.

4.4 Ambient observations to determine the NO_2/NO_x trend

Using the measured roadside atmospheric ratio of NO_2 to NO_x (NO_2/NO_x ratio, expressed as a molar volume ratio) is one effective way of determining the influence on NO_2 of increased proportions of diesel vehicles in a fleet, as well as a method to detect change in after treatment technologies resulting from progressive tightening of the Euro standards. Since there is no systematic set of vehicle exhaust measurements that show NO_2/NO_x trends we look instead at the combined national data sets of ambient monitoring information which measure NO and NO_2 in air. We carefully filter these datasets for roadside locations where the ratio of these two species can be taken as a proxy for the exhaust emission ratio. We note that there is considerable diversity in the penetration and uptake of diesel vehicles, typical vehicle lifespans, and climates when considering Europe as a whole. The analysis in this section uses data from roadside monitoring sites across 61 European urban areas between 1990 and 2015. The combined European trend (Figure 4.5) for the 61 areas demonstrates a clear increase in annual mean NO_2/NO_x ratio between 1995 and 2010. The aggregation was performed on the mean for each city in each year to ensure the results were not biased towards cities with more measurement locations, such as London.

Figure 4.5 shows three distinct periods where NO_2/NO_x ratio behaviour differed. The first, from 1990 to 1994 coincides with a pre-Euro 3 fleet that did not use diesel oxidation catalysts (DOCs) and the ratio was stable within the uncertainty of the slope estimate and less than 10 % (Table 4.2). The second period from 1995 to 2008 is a period when there was a clear, sustained, and significant increase in the NO_2/NO_x ratio corresponding to a period of growth in diesel passenger cars numbers and the introduction of DOC to new vehicles via Euro 3 and Euro 4. Over this period the ratio increased to a peak value of approximately 16 % in 2010. The third period is characterized by a stabilisation in the NO_2/NO_x ratio and coincides with the

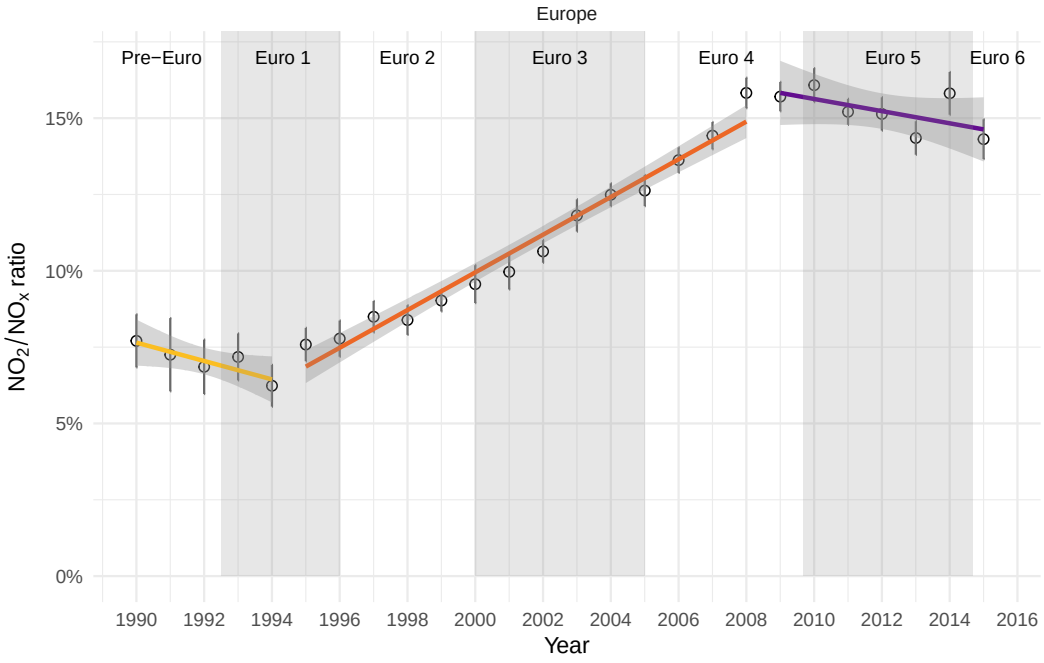

Figure 4.5: Mean NO_2/NO_x ratio for all roadside monitoring sites for the 61 European urban areas analysed between 1990 and 2015. The error bars represent the 95% confidence intervals of the slope estimates based on the number of samples. Linear regression models were applied to three separate periods: 1990–1994, 1995–2008, and 2009–2015 identified by segmented regression.

introduction of Euro 5 vehicles fitted with diesel particle filters (DPFs). The second period is the only period that shows a statistically significant change NO_2/NO_x ratio (Table 4.3). The trends shown in Figure 4.5 broadly follow the pattern of reported changes in emissions seen from sporadic remote sensing measurements of almost 70 000 vehicles in London (during 2012), with a progressive increase in NO_2/NO_x ratio for diesel passenger cars and light vans from pre-Euro to Euro 5.[52]

Table 4.2: Extra summary statistics for Figure 4.5. The error is the aggregated standard error of the slope estimates weighted by number of valid slope estimates ($\frac{\varepsilon}{\sqrt{n}}$). The 95% confidence interval (CI) is the error × 1.96.

Year	Slope	Count of sites	Count of urban areas	Error	CI lower (95%)	CI upper (95%)
1990	0.077058	9	5	0.008685	0.060034	0.094081
1991	0.072517	20	7	0.011990	0.049016	0.096018
1992	0.068564	27	9	0.008907	0.051107	0.086021
1993	0.071794	28	9	0.007688	0.056725	0.086863
1994	0.062352	29	10	0.006827	0.048971	0.075733
1995	0.075844	30	11	0.005465	0.065134	0.086555
1996	0.077790	41	13	0.005923	0.066181	0.089398
1997	0.084951	61	17	0.005150	0.074857	0.095045
1998	0.083878	93	24	0.004851	0.074370	0.093386
1999	0.090260	121	29	0.003579	0.083245	0.097276
2000	0.095636	157	34	0.006192	0.083501	0.107772
2001	0.099663	163	36	0.005772	0.088350	0.110976
2002	0.106345	178	40	0.003690	0.099113	0.113577
2003	0.118140	185	42	0.005280	0.107791	0.128490
2004	0.124910	217	49	0.003740	0.117579	0.132241
2005	0.126285	215	48	0.005085	0.116318	0.136252
2006	0.136261	217	52	0.004156	0.128116	0.144407
2007	0.144266	213	50	0.004464	0.135516	0.153015
2008	0.158265	217	56	0.004988	0.148488	0.168042
2009	0.157058	219	55	0.004810	0.147630	0.166486
2010	0.160809	235	56	0.005563	0.149905	0.171712
2011	0.152091	249	60	0.004245	0.143771	0.160410
2012	0.151386	202	43	0.005455	0.140695	0.162078
2013	0.143514	187	45	0.005484	0.132764	0.154263
2014	0.158128	198	44	0.006992	0.144424	0.171831
2015	0.143121	168	37	0.006510	0.130361	0.155881

Although the ambient derived NO_2/NO_x ratio turning points in Figure 4.5 broadly coincide with identifiable regulatory landmarks, the changes

Table 4.3: Linear regression model summaries for Figure 4.5. Period represent the three models: 1990–1994, 1995–2008, and 2009–2015.

Period	n	P-value	Slope	R^2
1	5	0.0527	-0.0030	0.7639
2	14	< 0.001	0.0062	0.9681
3	7	0.1397	-0.0020	0.3810

are more complex than they would first appear. First, when a new Euro class is introduced, it takes time for those new vehicles to significantly penetrate the vehicle fleet and affect overall emissions. Second, the emissions characteristics of vehicles will be expected to change as they age. For example, a Euro 3 car introduced in year 2000 will be $\approx$ 5–6 years old at the end of the Euro 3 period. Analysis of vehicle emission remote sensing data has shown that vehicle ageing tends to decrease the NO_2/NO_x ratio of diesel passenger cars (and likely other types of vehicles fitted with DOC).[16,53] All these influences, as well as other local effects, contribute to the overall pattern seen in Figure 4.5. Nevertheless, it is clear that on average, across Europe, the ratio has not continued to increase after 2010 and is now declining.

At an European level, mean annual roadside NO_x concentrations demonstrated an overall decrease from 1998 to 2015 with mean NO_x concentrations reducing from 338 to 228 $\mu g\,m^{-3}$ (Figure 4.6). Before 1998, the NO_x means are scattered due to fewer sites and observations and larger uncertainties concerning the quality of the measurements. This decrease can be attributed to improved vehicular NO_x emission control during this period. Figure 4.6 shows that mean NO_x concentrations have remained stable since 2010, however, the trend in NO_2 concentrations (the regulated species of NO_x) differs from total NO_x in several important ways. First, NO_2 concentrations tended to increase over the period from around 1997 to 2009 (despite concentrations of NO_x decreasing). Second, concentrations of NO_2 have tended to decrease from around 2009 at a time when concentrations of NO_x have been stable. These changes in concentrations are consistent with

the changes calculated for the NO_2/NO_x ratio, shown in Figure 4.5.

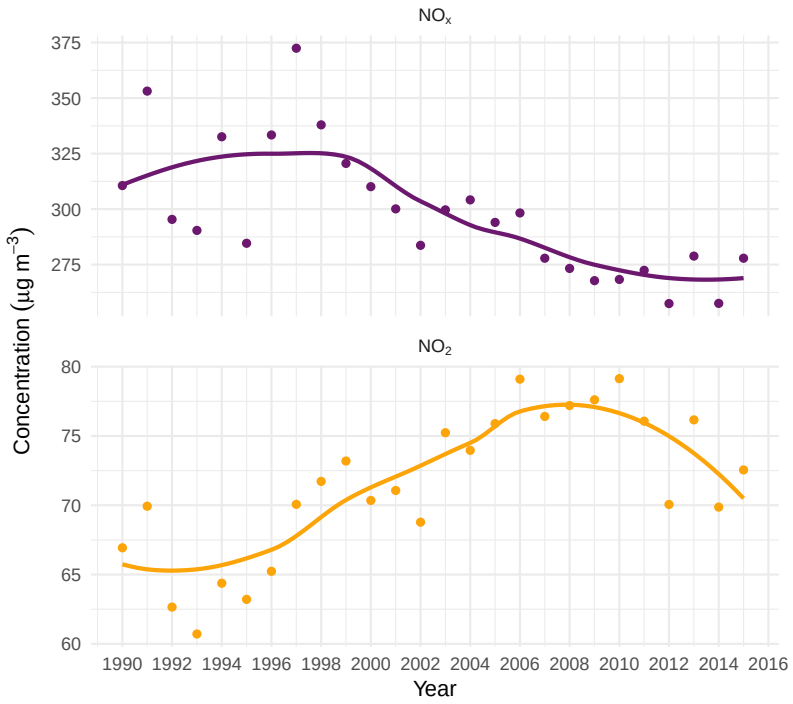

Figure 4.6: Mean NO_x and NO_2 concentrations after the filtering method was applied for all roadside monitoring sites for the 61 European urban areas analysed between 1990 and 2015. These concentration data were used for the calculation of the NO_2/NO_x ratio displayed in Figure 4.5. The smoothed lines are loess (local regression) fits.

4.4.1 Spatial analysis of roadside NO_2/NO_x over Europe

The Europe-wide aggregation displayed in Figure 4.5 hides the diversity of trends in the NO_2/NO_x ratio across European roadside monitoring sites, urban areas, and countries. When estimates of the NO_2/NO_x ratio were aggregated at an urban level, a peak ratio was observed at or near 2010 in most European urban areas (Figure 4.7). The trends in NO_2/NO_x ratio are shown for two periods 2005 to 2010 and 2010 to 2015. Over the first period most urban areas showed an increase in NO_2/NO_x, most pronounced in western and central Europe. For the later period the majority of regions

showed a declining trend in NO_2/NO_x albeit generally smaller than the earlier increases.

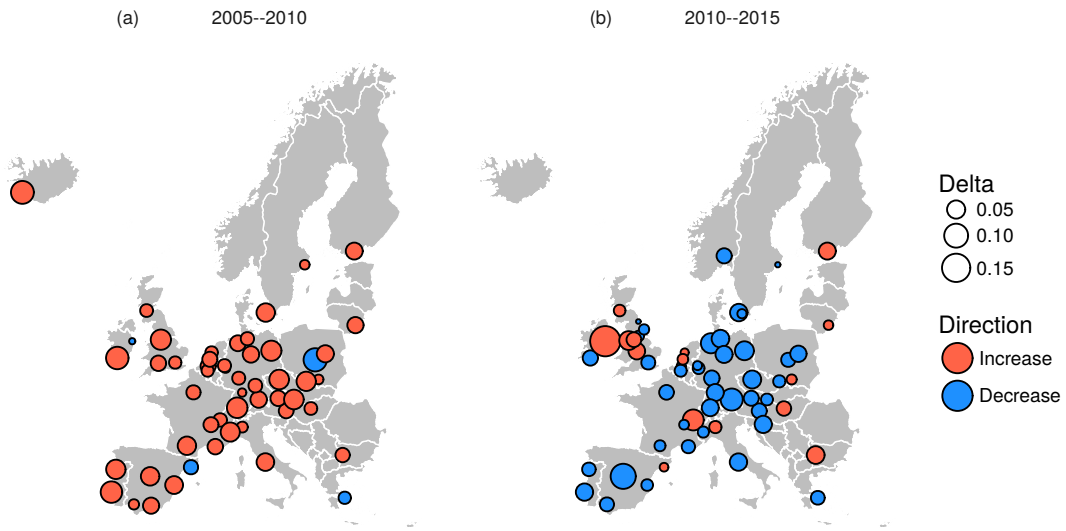

Figure 4.7: The change in the NO_2/NO_x ratio for each urban area for two time periods, the five years leading up to 2010, and the five years after 2010 (2010 is the year with the highest NO_2/NO_x ratio). Plot (a) shows the change in the NO_2/NO_x ratios from 2005 to 2010 and the plot (b) displays the change in ratio from 2010 to 2015. The size of the dots indicates the magnitude of the change.

Seven percent of the urban areas however showed opposing trends most likely reflecting unique and localised site or urban area conditions. Some of these urban areas including Amsterdam (Netherlands), Barcelona (Spain), Milan (Italy), and Krakow (Poland) demonstrate a levelling-off of the NO_2/NO_x ratio but had not shown decreasing trends by 2015. Other urban areas such as Dublin (Ireland which had the largest delta), Rotterdam (Netherlands), some urban areas in central United Kingdom, and Helsinki (Finland) showed further increases in NO_2/NO_x by 2015. Some urban areas, most conspicuously in Reykjavík (Iceland), are not shown in the 2010–2015 panel (b) in Figure 4.7. This was due to the absence of more-recent observations, usually due to O_3 or NO_x monitoring site closures or when the EU member state stopped reporting NO_x and NO alongside NO_2. It is very dif-

ficult to attempt attribute the underlying causes of the 7 % outliers; it may be associated with fleet make-up or indeed other local factors such as changing road layouts, new sources and urban infrastructure. In the absence of consistent information across Europe on these factors we do not speculate further.

The overwhelming consistency seen in the 93 % of urban areas and across the whole of the continent is however strongly suggestive of a European-scale influence on primary NO_2, not that this change in NO_2/NO_x is a result of a series of uncoordinated local factors. These changes are consistent with a steady evolution of the European fleet as a whole, for example, the effect of Euro standards and technologies, rather than trends driven by city or country specific interventions such as changes to local urban public transport fleets, introduction of congestion zones, and so on.

4.4.2 Potential factors controlling recent declines in NO_2/NO_x

Whilst the periods of increase in the NO_2/NO_x ratio can be rationalised based on previous evidence, the recent declines in ratio from around 2010 are more difficult to understand because diesel vehicles continue to use DOC with DPF. We raise here some potential factors that could explain this result. Remote sensing measurement of selected vehicles has showed that selective catalytic reduction (SCR) control systems introduced on heavy duty vehicles have improved, resulting in both lower overall emissions of NO_x and a better control of NO_2.[16] Although the numbers of heavy duty vehicles passing each monitor is unknown across Europe, this technology working on part of the fleet may have contributed to the ratio declining. A second potential factor is the ageing of exhaust control systems themselves, and an engineering shift towards "catalytic thrifting". This refers to vehicle manufacturers and catalyst developers progressively reducing the amount of platinum group metals used in exhaust systems which in turn has a consequence of reducing the amount of NO_2 generated. Finally, evidence from vehicle emission remote sensing shows that as light duty diesel vehicles age,

the NO_2/NO_x ratio does decrease over time although the extent of this is uncertain.[16] It would seem plausible that all of these poorly understood factors could, in combination, contribute to the stabilisation and decline seen in NO_2/NO_x ratio since 2010. However, with ambient data alone, it is impossible to quantify the individual contributions robustly.

4.5 Comparisons to emissions inventories

The Europe-wide primary NO_2/NO_x estimated by the observational filtering method here differs substantially from previous works which report roadside NO_2/NO_x ratio trends. Other inventories estimate higher NO_2/NO_x than what we see in the real world. A modelled estimate of traffic emissions at a national and European level in five year intervals between 2000 and 2030[15] predicted NO_2/NO_x to increase $\approx$ 25 % by 2020 and stay at this level until 2030 (Figure 4.8). Using these model estimates of NO_2/NO_x around 30 monitoring areas were then forecast to still be in breach of the European NO_2 air quality standard in 2030. The current United Kingdom (UK) vehicular primary NO_2 emission factors are also predicted up to 2030 in the National Atmospheric Emissions Inventory (NAEI).[54] The UK emission factors are derived from the COPERT database with modelling of predicted fleet changes in the future. The UK primary NO_2 emission factors for all UK urban areas are currently predicted to reach a peak NO_2/NO_x ratio in 2015 at 23 % (Figure 4.8). After 2015, the UK emission factors decrease until 2030 to a minimum ratio of 17 %.

Both emission estimates appear to substantially overstate the current fraction of emissions that is directly released as NO_2, in one case by nearly a factor two for the year 2015, and the measured vs. modelled trends are currently diverging further from one another. If primary NO_2 emissions remain similar or even further decreases as the current analysis suggests, the use of these inventory estimates for air quality modelling purposes would result in overly pessimistic future predictions of compliance with European

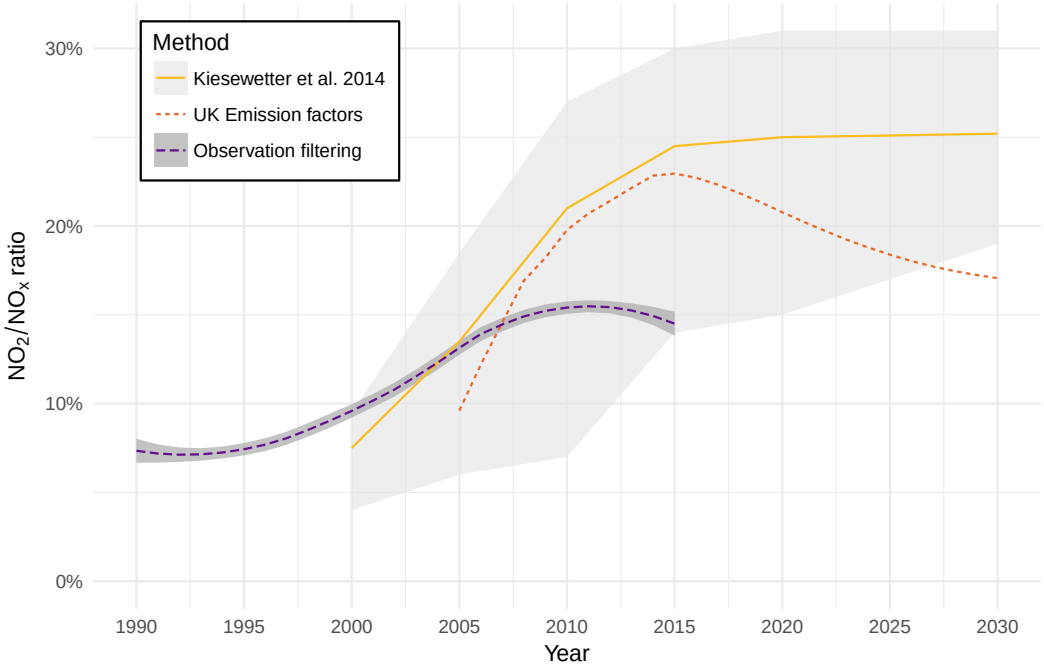

Figure 4.8: Comparison of three methods which estimate roadside primary NO_2 as a NO_2/NO_x ratio and forecasts from two other sources.[15,54] Shaded zones are the individual EU member state range in Kiesewetter et al. [15] and the 95% confidence interval of the observation filtering method's loess fit.

NO_2 ambient air quality standards.

4.6 Impact on the attainment of air quality standards

Policy projections of air quality that use too high a value for the NO_2/NO_x ratio will predict higher concentrations of roadside NO_2 than may actually occur for the same total amount of NO_x emitted. As an example of the potential changes brought about by using different NO_2/NO_x ratios, we compare how ambient concentrations would vary based on the current range of estimates. The most recent ratio reported here by the filtering method was 14.5 % in 2015 while the other reported estimates ranged from 25 to 22 % (Figure 4.8). To estimate the influence of differing primary NO_2 assumptions on roadside annual mean NO_2 concentrations, we have considered the roadside increment of NO_x concentration at each measurement site *i.e.* the increment in NO_x concentration above urban background values of NO_2. Two scenarios have been considered: first, that the roadside NO_x increment is associated with a NO_2/NO_x ratio of 14.5 % and second, that it is associated with a ratio of 23 %. Considering all European roadside sites, the mean difference in NO_2 concentration between these two scenarios is $6.6\,\mu g\,m^{-3}$. The current analysis, which applies data filtering techniques, is not strictly consistent with the changes expected to annual mean NO_2 concentrations because only a subset of data have been analysed. However, the changes in the NO_2/NO_x ratio identified will have a strong influence on annual mean NO_2 concentrations close to roads.

The impact of differing primary NO_2 assumptions will clearly vary depending on individual sites. However, for the most polluted NO_2 sites in Europe, examples being Brixton Road and Farringdon Street in London, the annual mean difference in NO_2 from the traffic contribution could be as much as $19\,\mu g\,m^{-3}$. Differences in projected NO_2 of this kind of magnitude are highly significant when compared against targets for compliance with

the European annual NO_2 ambient standard which is currently $40\,\mu g\,m^{-3}$. In this respect, current air quality modelling of roadside NO_2 that uses these unrealistically high NO_2/NO_x ratios for the future will tend to also be overly pessimistic. Should NO_2/NO_x ratios of the kind now being observed across Europe be projected forward for the next decade then attainment of annual roadside NO_2 standards in many places might be achieved sooner than is currently predicted.

We note however the substantial disconnections that still exist between the legislative controls being placed on reporting vehicle emissions and air quality standards designed to protect public health. By only requiring the reporting of total NO_x from new vehicles, and not NO and NO_2 as separate quantities, the later impacts of those vehicles, and how they influence the regulate pollutant NO_2, cannot be assessed. The continued lack of any systematic collection of information on changes to NO and NO_2 emissions as vehicles age is a further gap in evidence that if filled would greatly improve the reliability of future forecasts of air quality in cities.

4.7 Post-publication: Updates when additional observations were delivered

The original analysis used observations between 1990 and 2015. However, since publication two additional years of data have been reported and are available to update the analysis. The trends discussed have continued to follow the patterns outlined in the original analysis (Figure 4.9, Figure 4.10, and Figure 4.11). Figure 4.10 demonstrates that the European NO_2/NO_x emission ratio determined by the observational record is now outside the uncertainty of the European model presented in Kiesewetter et al. [15]. Therefore, it seems there is little doubt that the flattening off or decrease of the NO_2/NO_x emission ratio at a European level will continue for the next several years.

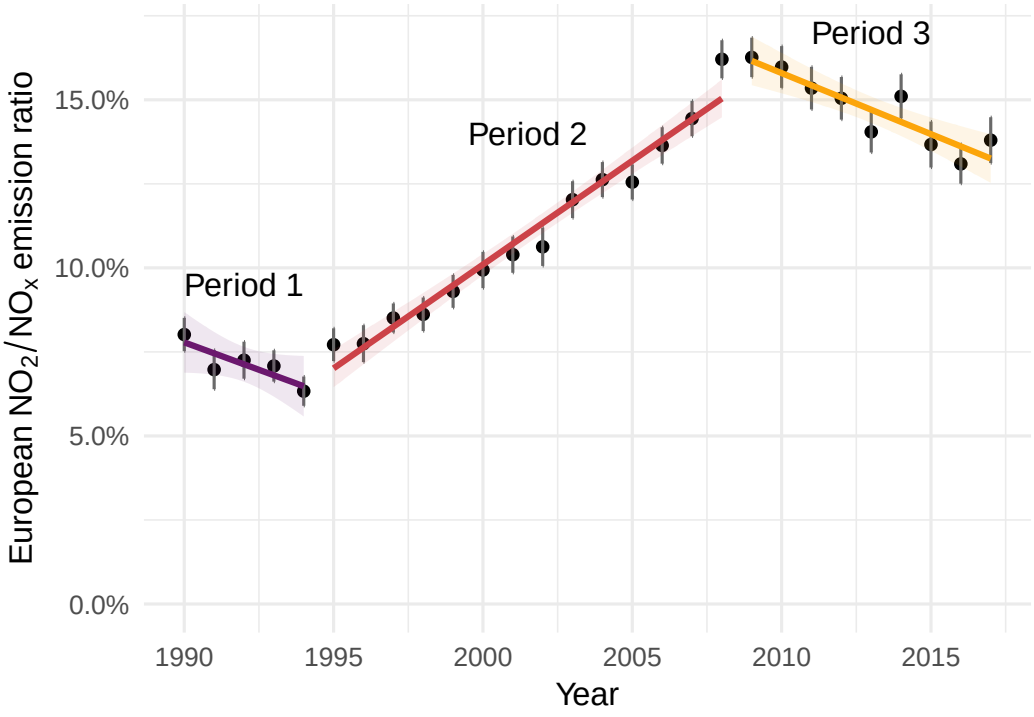

Figure 4.9: Updated European primary NO_2/NO_x emission ratio. The original data set ended at the end of 2015, data for 2016 and 2017 are now available.

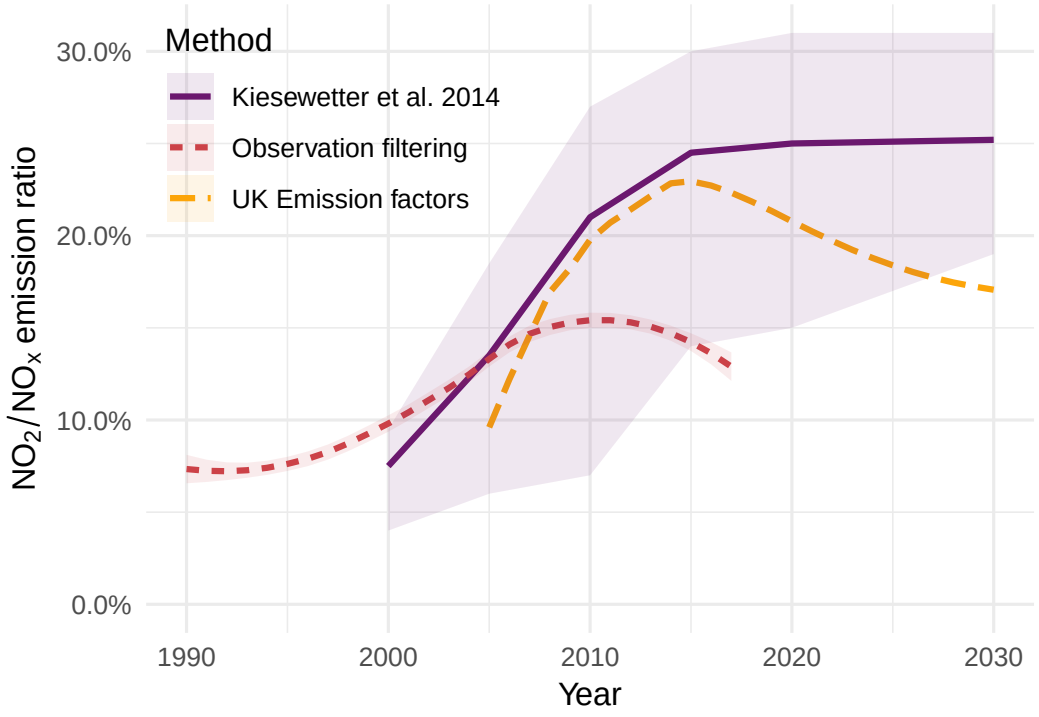

Figure 4.10: Updated comparison of three methods which estimate roadside NO_2/NO_x emission ratios. The original data set ended at the end of 2015, data for 2016 and 2017 are now available.

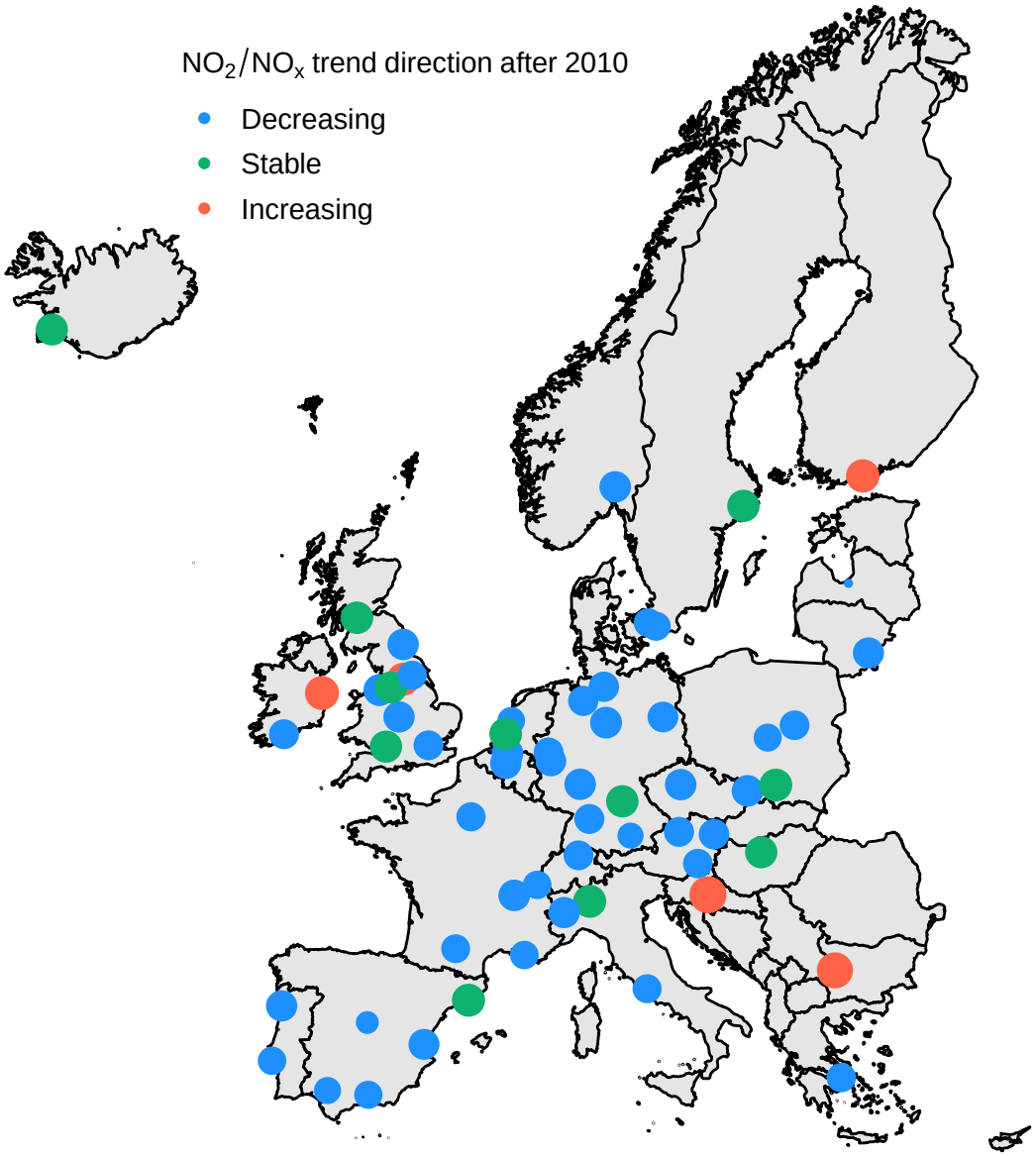

Figure 4.11: Updated change in the NO_2/NO_x emission ratios for each urban area. The direction is calculated from 2010 (the peak emission ratio year) until 2017, the final year of data included in the analysis.

4.8 References

[1] Cames, M. and Helmers, E. Critical evaluation of the European diesel car boom - global comparison, environmental effects and various national strategies. *Environmental Sciences Europe* 25.1 (2013), pp. 1–22. DOI: 10.1186/2190-4715-25-15. URL: http://dx.doi.org/10.1186/2190-4715-25-15.

[2] European Environment Agency. *Dieselisation in the EEA*. 2015. URL: http://www.eea.europa.eu/data-and-maps/figures/dieselisation-in-the-eea.

[3] International Council on Clean Transportation Europe. *European vehicle market statistics, 2015/2016*. Pocketbook. 2015. URL: http://www.theicct.org/european-vehicle-market-statistics-2015-2016.

[4] The European Automobile Manufacturers' Association. Share of Diesel in New Passenger Cars. European Automobile Manufacturers' Association. 2016. URL: http://www.acea.be/statistics/tag/category/share-of-diesel-in-new-passenger-cars.

[5] Koetse, M. J. and Hoen, A. Preferences for alternative fuel vehicles of company car drivers. *Resource and Energy Economics* 37 (2014), pp. 279–301. URL: http://www.sciencedirect.com/science/article/pii/S0928765514000049.

[6] European Automobile Manufacturers' Association. *ACEA Tax Guide*. 2016. URL: http://www.acea.be/uploads/news_documents/ACEA_TAX_GUIDE_2016.pdf.

[7] Schmidt, C. W. Beyond a One-Time Scandal: Europe's Onging Diesel Pollution Problem. *Environmental Health Perspectives* 124.1 (2016), A19–A22. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4710587/.

[8] Weiss, M., Bonnel, P., Kühlwein, J., Provenza, A., Lambrecht, U., Alessandrini, S., Carriero, M., Colombo, R., Forni, F., Lanappe, G., Le Lijour, P., Manfredi, U., Montigny, F., and Sculati, M. Will Euro 6 reduce the NO_x emissions of new diesel cars? — Insights from on-road tests with Portable Emissions Measurement Systems (PEMS). *Atmospheric Environment* 62 (2012), pp. 657–665. DOI: https://doi.org/10.1016/j.atmosenv.2012.08.

056. URL: http://www.sciencedirect.com/science/article/pii/S1352231012008412.

[9] European Parliament and Council. *Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe*. http://data.europa.eu/eli/dir/2008/50/oj. 2008.

[10] European Environment Agency. *Air quality in Europe — 2016 report*. EEA Report. No 28/2016. 2016. URL: http://www.eea.europa.eu/publications/air-quality-in-europe-2016.

[11] Carslaw, D. C. and Carslaw, N. Detecting and characterising small changes in urban nitrogen dioxide concentrations. *Atmospheric Environment* 41.22 (2007), pp. 4723–4733. DOI: 10.1016/j.atmosenv.2007.03.034. URL: http://www.sciencedirect.com/science/article/pii/S1352231007002919.

[12] Alvarez, R., Weilenmann, M., and Favez, J.-Y. Evidence of increased mass fraction of NO_2 within real-world NO_x emissions of modern light vehicles — derived from a reliable online measuring method. *Atmospheric Environment* 42.19 (2008), pp. 4699–4707. URL: http://www.sciencedirect.com/science/article/pii/S1352231008000964.

[13] Keuken, M., Roemer, M., and van den Elshout, S. Trend analysis of urban NO_2 concentrations and the importance of direct NO_2 emissions versus ozone/NO_x equilibrium. *Atmospheric Environment* 43.31 (2009), pp. 4780–4783. URL: http://www.sciencedirect.com/science/article/pii/S1352231008006298.

[14] Williams, M. L. and Carslaw, D. C. New Directions: Science and policy — Out of step on NO_x and NO_2? *Atmospheric Environment* 45.23 (2011), pp. 3911–3912. DOI: https://doi.org/10.1016/j.atmosenv.2011.04.067. URL: http://www.sciencedirect.com/science/article/pii/S1352231011004560.

[15] Kiesewetter, G., Borken-Kleefeld, J., Schöpp, W., Heyes, C., Thunis, P., Bessagnet, B., Terrenoire, E., Gsella, A., and Amann, M. Modelling NO_2 concentrations at the street level in the GAINS integrated assessment model: projections under current legislation. *Atmospheric Chemistry and Physics* 14.2

(2014), pp. 813–829. DOI: https://doi.org/10.5194/acp-14-813-2014. URL: http://www.atmos-chem-phys.net/14/813/2014/.

[16] Carslaw, D. C., Murrells, T. P., Andersson, J., and Keenan, M. Have vehicle emissions of primary NO_2 peaked? *Faraday Discussions* 189.0 (2016), pp. 439–454. DOI: 10.1039/C5FD00162E. URL: http://dx.doi.org/10.1039/C5FD00162E.

[17] Carslaw, D. C. Evidence of an increasing NO_2/NO_x emissions ratio from road traffic emissions. *Atmospheric Environment* 39.26 (2005), pp. 4793–4802. DOI: https://doi.org/10.1016/j.atmosenv.2005.06.023. URL: http://www.sciencedirect.com/science/article/pii/S1352231005005443.

[18] Ligterink, N. E., Kadijk, G., and van Mensch, P. *Determination of Dutch NO_x emission factors for Euro-5 diesel passenger cars*. TNO 2012 R11099. 2012.

[19] Carslaw, D. C., Beevers, S. D., Tate, J. E., Westmoreland, E. J., and Williams, M. L. Recent evidence concerning higher NO_x emissions from passenger cars and light duty vehicles. *Atmospheric Environment* 45.39 (2011), pp. 7053–7063. DOI: 10.1016/j.atmosenv.2011.09.063. URL: http://www.sciencedirect.com/science/article/pii/S1352231011010260.

[20] World Health Organization. Chapter 7.1—Nitrogen dioxide. *WHO air quality guidelines for Europe, 2nd edition, 2000*. 2000. URL: http://www.euro.who.int/en/health-topics/environment-and-health/air-quality/publications/pre2009/who-air-quality-guidelines-for-europe,-2nd-edition,-2000-cd-rom-version.

[21] European Environment Agency. Premature deaths attributable to air pollution. 2016. URL: https://www.eea.europa.eu/media/newsreleases/many-europeans-still-exposed-to-air-pollution-2015/premature-deaths-attributable-to-air-pollution.

[22] European Environment Agency. Exceedances of air quality objectives due to traffic. 2016. URL: http://www.eea.europa.eu/data-and-maps/indicators/exceedances-of-air-quality-objectives/exceedances-of-air-quality-objectives-9.

[23] Grice, S., Stedman, J., Kent, A., Hobson, M., Norris, J., Abbott, J., and Cooke, S. Recent trends and projections of primary NO_2 emissions in Europe. *Atmospheric Environment* 43.13 (2009), pp. 2154–2167. DOI: https://doi.org/10.1016/j.atmosenv.2009.01.019. URL: http://www.sciencedirect.com/science/article/pii/S1352231009000508.

[24] Brand, C. Beyond 'Dieselgate': Implications of unaccounted and future air pollutant emissions and energy use for cars in the United Kingdom. *Energy Policy* 97 (2016), pp. 1–12. URL: http://www.sciencedirect.com/science/article/pii/S030142151630341X.

[25] Ntziachristos, L., Papadimitriou, G., Ligterink, N., and Hausberger, S. Implications of diesel emissions control failures to emission factors and road transport NO_x evolution. *Atmospheric Environment* 141 (2016), pp. 542–551. URL: http://www.sciencedirect.com/science/article/pii/S1352231016305568.

[26] Gkatzoflias, D., Kouridis, C., Ntziachristos, L., and Samaras, Z. *COPERT 4. Computer programme to calculate emissions from road transport*. User manual (version 9.0). 2012. URL: http://emisia.com/sites/default/files/COPERT4v9_manual.pdf.

[27] INFRAS. *Handbook emission factors for road transport (HBEFA)*. 2015. URL: http://www.hbefa.net/e/index.html.

[28] Department for Transport. Vehicle Emissions Testing Programme. 2016. URL: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/548148/vehicle-emissions-testing-programme-web.pdf.

[29] Johnson, T. V. Review of Diesel Emissions and Control. *SAE International Journal of Fuels and Lubricants* 3.1 (2010), pp. 16–29. URL: http://dx.doi.org/10.4271/2010-01-0301.

[30] Wild, R. J., Dubé, W. P., Aikin, K. C., Eilerman, S. J., Neuman, J. A., Peischl, J., Ryerson, T. B., and Brown, S. S. On-road measurements of vehicle NO_2/NO_x emission ratios in Denver, Colorado, USA. *Atmospheric Environment* 148 (2017), pp. 182–189. URL: http://www.sciencedirect.com/science/article/pii/S1352231016308469.

[31] European Commission. Transport Emissions — Air pollutants from road transport. 2016. URL: http://ec.europa.eu/environment/air/transport/road.htm.

[32] European Environment Agency. *AirBase – The European air quality database (Version 8)*. 2014. URL: https://www.eea.europa.eu/data-and-maps/data/airbase-the-european-air-quality-database-8.

[33] European Environment Agency. *Eionet Central Data Repository*. 2017. URL: http://cdr.eionet.europa.eu/.

[34] Grange, S. K. ***smonitor**: A framework and a collection of functions to allow for maintenance of air quality monitoring data*. 2018. URL: https://github.com/skgrange/smonitor.

[35] Carslaw, D. C. and Ropkins, K. *openair* — An R package for air quality data analysis. *Environmental Modelling & Software* 27–28.0 (2012), pp. 52–61. URL: http://www.sciencedirect.com/science/article/pii/S1364815211002064.

[36] Carslaw, D. and Ropkins, K. *openair: Open-source tools for the analysis of air pollution data*. 2015.

[37] Airparif. *Association de surveillance de la qualité de l'air en Île-de-France*. http://www.airparif.asso.fr/.

[38] Grange, S. K. *Technical note: **smonitor** Europe*. Tech. rep. Wolfson Atmospheric Chemistry Laboratories, University of York, 2017. DOI: https://doi.org/10.13140/RG.2.2.20555.49448/1. URL: https://doi.org/10.13140/RG.2.2.20555.49448/1.

[39] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2018. URL: https://www.R-project.org/.

[40] PostgreSQL Global Development Group. PostgreSQL. Version 9.5. URL: https://www.postgresql.org/.

[41] Parrish, D. D., Trainer, M., Hereid, D., Williams, E. J., Olszyna, K. J., Harley, R. A., Meagher, J. F., and Fehsenfeld, F. C. Decadal change in carbon monoxide to nitrogen oxide ratio in U.S. vehicular emissions. *Journal of Geophysical Research* 107.D12 (2002), ACH 5-1–ACH 5-9. URL: http://dx.doi.org/10.1029/2001JD000720.

[42] Hassler, B., McDonald, B. C., Frost, G. J., Borbon, A., Carslaw, D. C., Civerolo, K., Granier, C., Monks, P. S., Monks, S., Parrish, D. D., Pollack, I. B., Rosenlof, K. H., Ryerson, T. B., Schneidemesser, E. von, and Trainer, M. Analysis of long-term observations of NO_x and CO in megacities and application to constraining emissions inventories. *Geophysical Research Letters* 43 (2016). URL: http://dx.doi.org/10.1002/2016GL069894.

[43] European Commission. *European cities – the EU-OECD functional urban area definition*. 2015. URL: http://ec.europa.eu/eurostat/statistics-explained/index.php/European_cities_%E2%80%93_the_EU-OECD_functional_urban_area_definition.

[44] OpenStreetMap Foundation and OpenStreetMap contributors. *OpenStreetMap*. http://www.openstreetmap.org. 2016.

[45] Haklay, M. and Weber, P. OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing* 7.4 (2008), pp. 12–18.

[46] Steinbacher, M., Zellweger, C., Schwarzenbach, B., Bugmann, S., Buchmann, B., Ordóñez, C., Prevot, A. S. H., and Hueglin, C. Nitrogen oxide measurements at rural sites in Switzerland: Bias of conventional measurement techniques. *Journal of Geophysical Research* 112.D11 (2007), 13pp. DOI: 10.1029/2006JD007971. URL: http://dx.doi.org/10.1029/2006JD007971.

[47] Kirchstetter, T. W., Harley, R. A., and Littlejohn, D. Measurement of Nitrous Acid in Motor Vehicle Exhaust. *Environmental Science & Technology* 30.9 (1996), pp. 2843–2849. DOI: 10.1021/es960135y. URL: http://dx.doi.org/10.1021/es960135y.

[48] Venables, W. N. and Ripley, B. D. *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer, 2002. URL: http://www.stats.ox.ac.uk/pub/MASS4.

[49] Muggeo, V. M. Estimating regression models with unknown break-points. *Statistics in Medicine* 22 (2003), pp. 3055–3071.

[50] Muggeo, V. M. Segmented: an R package to fit regression models with broken-line relationships. *R news* 8.1 (2008), pp. 20–25.

[51] Jenkin, M. E. Analysis of sources and partitioning of oxidant in the UK—Part 2: contributions of nitrogen dioxide emissions and background ozone at a kerbside location in London. *Atmospheric Environment* 38.30 (2004), pp. 5131–5138. URL: http://www.sciencedirect.com/science/article/pii/S1352231004006193.

[52] Carslaw, D. C. and Rhys-Tyler, G. New insights from comprehensive on-road measurements of NO_x, NO_2 and NH_3 from vehicle emission remote sensing in London, UK. *Atmospheric Environment* 81.0 (2013), pp. 339–347. URL: http://www.sciencedirect.com/science/article/pii/S1352231013007140.

[53] Carslaw, D. C., Williams, M. L., Tate, J. E., and Beevers, S. D. The importance of high vehicle power for passenger car emissions. *Atmospheric Environment* 68.0 (2013), pp. 8–16. DOI: 10.1016/j.atmosenv.2012.11.033. URL: http://www.sciencedirect.com/science/article/pii/S1352231012010953.

[54] UK National Atmospheric Emission Inventory. *Primary NO_2 Emission Factors for Road Vehicles*. August 2014 update. 2014.

Chapter 5

Meteorological normalisation of Swiss PM_{10}

This work was originally was published in *Atmospheric Chemistry and Physics* on May 3, 2018.[†]

5.1 Abstract

Meteorological normalisation is a technique which accounts for changes in meteorology over time in an air quality time series. Controlling for such changes helps support robust trend analysis because there is more certainty that the observed trends are due to changes in emissions or chemistry, not changes in meteorology. Predictive random forest models (RF; a decision tree machine learning technique) were grown for 31 air quality monitoring sites in Switzerland using surface meteorological, synoptic scale, boundary layer height, and time variables to explain daily PM_{10} concentrations. The RF models were used to calculate meteorologically normalised trends which were formally tested and evaluated using the Theil-Sen estimator. Between 1997 and 2016, significantly decreasing normalised PM_{10} trends ranged between -0.09 and -1.16 $\mu g\,m^{-3}$ $year^{-1}$ with urban traffic sites experiencing the greatest mean decrease in PM_{10} concentrations at -0.77 $\mu g\,m^{-3}$ $year^{-1}$.

[†]https://doi.org/10.5194/acp-18-6223-2018

Similar magnitudes have been reported for normalised PM_{10} trends for earlier time periods in Switzerland which indicates PM_{10} concentrations are continuing to decrease at similar rates as in the past. The ability for RF models to be interpreted was leveraged using partial dependence plots to explain the observed trends and relevant physical and chemical processes influencing PM_{10} concentrations. Notably, two regimes were suggested by the models which cause elevated PM_{10} concentrations in Switzerland: one related to poor dispersion conditions and a second resulting from high rates of secondary PM generation in deep, photochemically active boundary layers. The RF meteorological normalisation process was found to be robust, user friendly and simple to implement, and readily interpretable which suggests the technique could be useful in many air quality exploratory data analysis situations.

5.2 Introduction

5.2.1 Air quality trend analysis

Trend analysis of ambient air quality data is a common and important procedure. The goal of such trend analysis usually involves the confirmation, or lack of confirmation of a statistically significant change in pollutant concentrations over time. If pollutant concentrations are significantly increasing or decreasing, there is evidence that air quality is better or worse than in the past and conclusions such as these are useful for scientists, policy makers, and the public.[1] However, air quality trend analysis is complicated because it is usually unknown if the behaviour of the trend is driven by changes in meteorology or changes in emissions or atmospheric chemistry.[2–6] The former is usually is of greatest importance for policy makers because investigation in changes in emissions, and in turn, the perturbations on ambient pollutant concentrations is how efficacy of intervention activities are judged.[7,8] Despite the uncertainty surrounding the drivers of

air pollutant trends, this issue is often acknowledged but rarely robustly compensated for.

The issue surrounding meteorology and air quality trend analysis arises because air quality and pollutant concentrations are highly dependent on meteorological conditions across all spatial scales.[9] Wind speed, wind direction, atmospheric temperature and stability can be expected to have large influences on pollutant concentrations at most locations. The influence of such meteorological variables can be much greater than an intervention activity which results in meteorological conditions often obscuring or exacerbating trends.[10] In situations where these processes are not accounted for, a calculated trend is less likely to represent changes in pollutant emissions due to air quality management efforts and therefore erroneous conclusions can be made on what is causing the observed trend.

The methods used for trend analysis are diverse and range from simple least squares linear regression analysis to numerically complex methods often requiring multiple pre-processing or work-up steps before the final trend test is conducted.[1,11,12] When trends are found to be monotonic, *i.e.* constantly changing with time, the robust non-parametric linear regression Mann-Kendal test is often used.[13] The Mann-Kendal test can be supplemented by using the Theil-Sen estimator and bootstrapping techniques which increase the test's robustness and can account for autocorrelation in the time series.[14–16] Although methods for the testing of monotonic trends are mature and are in common usage in air quality and other environmental applications[17], much of the effort of trend analysis is put into the preprocessing steps which generally involves deciding what aggregation period and function to use as well as handling the removal of the seasonal component if necessary (an annual cyclical pattern). Common techniques to allow for removal of the seasonal component of a time series is classical decomposition using loess (often called seasonal and trend decomposition using loess; STL) and Kolmogorov-Zurbenko filters.[6,18,19] Although these decomposition methods help treat the time series for further trend analysis,

they alone do not address changes of meteorology over time.

5.2.2 Meteorological normalisation

A method to control or take into account meteorology effects on pollutant concentrations involves the development and use of predictive statistical models.[8,11,20–22] Such models attempt to use a number of explanatory variables such as surface measurements of wind behaviour, atmospheric temperature, and pressure to explain the variability of pollutant concentrations. Time variables such as Julian day (day of the year), weekday, and hour of the day can also be used as predictors. These time variables act as proxies for emission strength because pollutant emissions or generation processes vary by the time of day, day of the week, and season.[23] If the predictive models are found to explain an adequate amount of the variation in pollutant concentration, the model can be used to account for the influence of meteorological variables on the pollutant concentration. The explanation of some of the variation in a time series also has the side effect of allowing significant trends to be detected earlier because of the reduction of estimate uncertainty. This technique is known by a few different names but here, we refer to the technique as *meteorological normalisation*.

The application of meteorological normalisation approaches are however complicated due to how pollutant concentrations vary based on meteorological variables. For example, for a traffic sourced pollutant such as nitrogen dioxide (NO_2), it would be expected that concentrations would decrease with increasing wind speed due to atmospheric dilution and dispersion processes.[24] However, this process is highly unlikely to be linear and when a monitoring site is located adjacent to a kerb, the effect of dilution based on the wind speed would also be highly dependent on wind direction. There would be further complication if the monitoring site was located within a street canyon. When variables depend on one another (or among more than two variables) in such a way, this is termed interaction.[25] Interaction effects generally require special treatment in most statistical models.

Additionally normality, homoscedasticity, multicollinearity, and independence should also be addressed before and during statistical modelling. All of these features are commonly encountered in air quality time series which can make the use statistical techniques highly burdensome in this domain.

5.2.3 Machine learning

In the past three decades, there has been large development in the field of what is now known as machine learning (ML). ML is a fusion of statistics, data science, and computing which experiences use across a very wide range of applications.[26,27] ML is a diverse topic but it has seen the development of many predictive models which offer alternatives to "classical" statistical models for exploratory data analysis. Some of the more popular ML predictive models include decision tree methods such as boosted regression trees and random forest, the kernel methods which include support vector machines, and finally artificial neural networks.[28] These ML methods, when used in regression mode, can be used in similar applications as multiple regression models such as general additive models (GAMs). These ML techniques are non-parametric and have the critical advantage of not needing to address many of the assumptions needed for statistical models such as sample normality, homoscedasticity, independence, adherence to other strict parametric assumptions, and the careful handling of interaction effects.[29] ML predictive models have the potential to supplement more classical statistical techniques which may result in improved air quality trend analysis.

5.2.3.1 Decision trees and random forest

Random forest (RF) (also known as decision forests) which is utilised in this study is an ensemble decision tree ML method.[30,31] Decision trees use a binary recursive classifying algorithm which creates "pure" nodes by splitting observations into two homologous groups. The recursive nature of the al-

gorithm means splitting is repeated until node purity is achieved. Together the entire series of splits, individually called nodes or branches, is referred to as a tree. The recursive algorithm will always correctly classify input data if the trees are allowed to grow to their maximum depth. Algorithms of this sort are called greedy.[32] This greedy behaviour can result in very deep trees (especially with continuous numeric variables) where the final split is only evaluating two observations *i.e.*, a singleton node. Models such these will very rarely generalise to new data which was not used to train the model. Therefore, decision trees are prone to overfitting.[33] RF controls for this disadvantage by growing many individual decision trees from a training set using a process called bagging (bootstrap aggregation). RF is an ensemble method because the model consists of many individual trees/-models/learners grown from bagged data but when used for prediction, all the trees' outputs are used together (Figure 5.1).

Bagging refers to randomly sampling observations with replacement from the training set along with sampling of explanatory variables.[35] A set which results from bagging is called out-of-bag data (OOB) and OOB data will always be lacking some of the input data. When a single tree is grown from OOB data, it is unlikely to contain the same observations and variables used by other trees if the process is repeated. RF models usually contain a few hundred trees using OOB data and this creates a forest which consists of many decorrelated trees which have been trained on different subsets of the training set (Figure 5.1). Every tree can then be used to predict and the predictions are aggregated to form a single prediction. In regression applications, the mean of predictions is used. Somewhat counter intuitively, the bagging process and ensemble predictions addresses decision trees' tendency to overfit training sets.[36] This allows RF to produce predictive models which generalise well and predictive performance is generally considered among the best of any ML technique.[37]

RF also has the advantage of not being a "black-box" method.[38] Decision trees are one of the few ML techniques where the learning process can

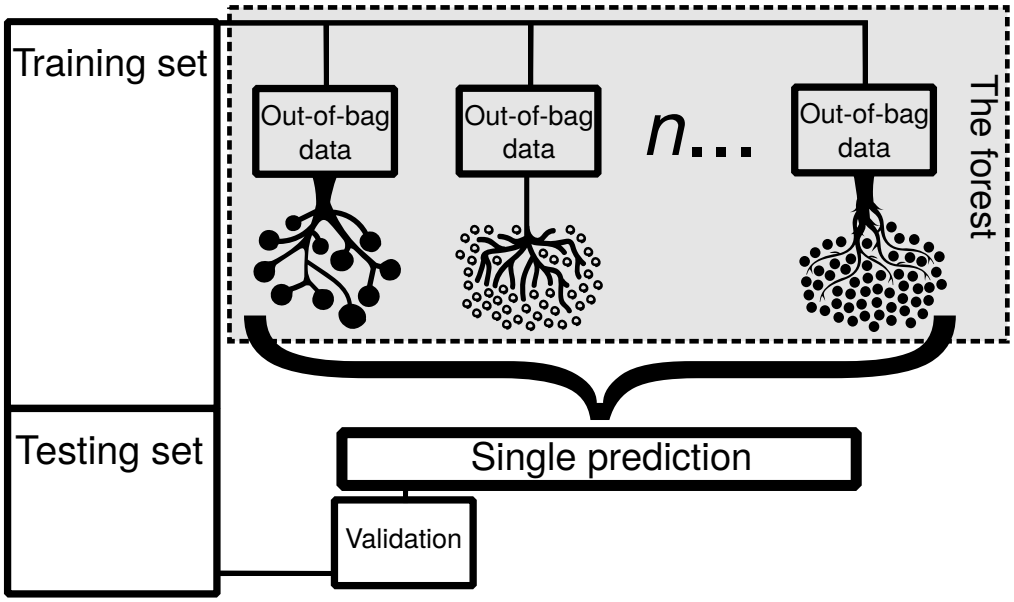

Figure 5.1: Conceptual diagram of a random forest model. Many out-of-bag samples are taken from the training set and different decision trees are grown. After many decision trees are grown, termed the forest, all trees are used to form a single prediction. The predictions can then be validated using the test set which is withheld from the training process. Tree icons are from freepik.com [34].

be explained, investigated, and interpreted. In the case of artificial neural networks or kernel based learning methods, this is much more difficult to do.[31,33] RF models can be investigated with partial dependence plots which demonstrate the relationships among variables and a variable's importance as a predictor can be determined. RF can be used in unsupervised, regression, or classification modes, accepts numeric and categorical variables, and is known to be simpler to tune when compared to other decision tree methods which usually require pruning; a process which removes some of the grown branches from the forest. The combination of these attributes has made RF a popular ML technique.[29,36]

5.2.4 Objectives

Improvements in the pre-processing steps for air quality trend analysis need to be made which control, or account for meteorology and allow for more robust trend and intervention exploration. This paper has the overall objective to present a meteorological normalisation technique which uses RF predictive models to prepare ambient atmospheric pollutant concentration data for trend analysis. Specifically, this paper will (*i*) present a meteorological normalisation technique using RF predictive models using routine data which will be accessible to most data users, (*ii*) present a trend analysis of the meteorologically normalised time series, and (*iii*) use RF's advantage of being able to interpret the learning processes to explain the trends which are observed. Daily PM_{10} observations from across Switzerland will be used for the analysis. The use of daily Swiss PM_{10} data was chosen because the data record and capture rates are excellent, and a previous study[39] conducted a PM_{10} trend analysis using a different method for observations between 1991 and 2008. Therefore, this work also updates and extends previous work.

5.3 Methods

5.3.1 Data

Routine air quality observations from Switzerland were used in this study and these data were accessed from the European Environment Agency (EEA) AirBase and Air Quality e-Reporting (AQER) data repositories.[40,41] The AirBase repository includes data between 1969 and 2012 (inclusive) while the AQER repository contains data from 2013 onwards. These two repositories contain monitoring sites which are within Switzerland's National Air Pollution Monitoring Network (NABEL) and sites which are managed by the Swiss Cantons (states).[42,43] Data from the two repositories have different data models and file formats which required transformation and processing into a standardised relational data model called **smonitor**.[44,45] The Härkingen-A1 and Sion-Aéroport sites' data are not submitted to the EEA, therefore these data were requested and delivered directly from the Swiss Federal Office for the Environment (FOEN).

Daily PM_{10} observations were used as the pollutant of interest and in the models as the dependent variable. Observations between 1997 and 2016 were used and the observations were collected with the use of commercially available gravimetric instrumentation and are subjected to quality assurance and control procedures.[43] A total of 186 400 PM_{10} observations from 31 sites were used. The sites were classified into six site types: rural, rural mountain, urban background, suburban, rural motorway, or urban traffic based on classifications in the AQER reporting system. For site locations and details see Table 5.1 and Figure 5.2.

The 31 PM_{10} monitoring sites where chosen for their suitably for use in trend analysis. The main condition was that PM_{10} observations needed to be unbroken for at least five years. One exception was made for Zürich-Schimmelstrasse. Zürich-Schimmelstrasse has a broken PM_{10} time series due to PM_{10} monitoring occurring every second year between 2002 and 2010, however, these data were still considered valuable to include in the

analysis. All other sites had very high data capture rates (median of 99 %) for the duration they were operational. Five monitoring sites were closed before, or did not have PM_{10} data to the end of the analysed time period (the end of 2016) but until their date of closure, had uninterrupted PM_{10} time series.

Table 5.1: Information for the Swiss PM_{10} and meteorological monitoring sites used in this study.

Site name	Latitude	Longitude	Elevation (m)	Site type	Site name ISD (met.)
Avully-Passeiry	46.163	6.005	427	Rural	Geneva Cointrin
Magadino-Cadenazzo	46.160	8.934	203	Rural	Locarno - Magadino
Payerne	46.813	6.944	489	Rural	Payerne
Saxon	46.139	7.148	460	Rural	Sion
Tänikon	47.480	8.905	538	Rural	Aadorf-Taenikon
Härkingen-A1	47.312	7.821	431	Rural motorway	Wynau
Sion-Aéroport-A9	46.220	7.342	483	Rural motorway	Sion
Chaumont	47.050	6.979	1136	Rural mountain	Chasseral
Rigi-Seebodenalp	47.067	8.463	1031	Rural mountain	Luzern
Basel-Binningen	47.541	7.583	316	Suburban	Bale Mulhouse
Dübendorf-EMPA	47.403	8.613	432	Suburban	Zuerich-Fluntern
Ebikon-Sedel	47.068	8.301	482	Suburban	Luzern
Ittigen	46.976	7.479	460	Suburban	Bern-Zollikofen
Lugano-Pregassona	46.026	8.968	305	Suburban	Lugano
Meyrin-Vaudagne	46.231	6.074	439	Suburban	Geneva Cointrin
Opfikon-Balsberg	47.439	8.570	430	Suburban	Zuerich-Fluntern
Thônex-Foron	46.196	6.211	422	Suburban	Geneva Cointrin
Basel-St-Johann	47.566	7.582	260	Urban background	Bale Mulhouse
Lugano-Università	46.011	8.957	280	Urban background	Lugano
Luzern-Museggstrasse	47.056	8.310	460	Urban background	Luzern
Winterthur-Obertor	47.500	8.732	448	Urban background	Zuerich-Fluntern
Zürich-Kaserne	47.378	8.530	409	Urban background	Zuerich-Fluntern
Basel-Feldbergstrasse	47.567	7.595	255	Urban traffic	Bale Mulhouse
Bern-Bollwerk	46.951	7.441	536	Urban traffic	Bern Belp
Bern-Brunngasshalde	46.949	7.450	533	Urban traffic	Bern-Zollikofen
Genève-Ile	46.206	6.143	375	Urban traffic	Geneva Cointrin
Genève-Wilson	46.216	6.151	376	Urban traffic	Geneva Cointrin
Lausanne-César-Roux	46.522	6.640	530	Urban traffic	Geneva Cointrin
St-Gallen-Rorschacherstrasse	47.429	9.387	660	Urban traffic	St. Gallen
Zürich-Schimmelstrasse	47.371	8.524	415	Urban traffic	Zuerich-Fluntern
Zürich-Stampfenbachstrasse	47.387	8.540	445	Urban traffic	Zuerich-Fluntern

Surface meteorological variables to be included in the modelling process such as wind speed, wind direction, and atmospheric temperature were ac-

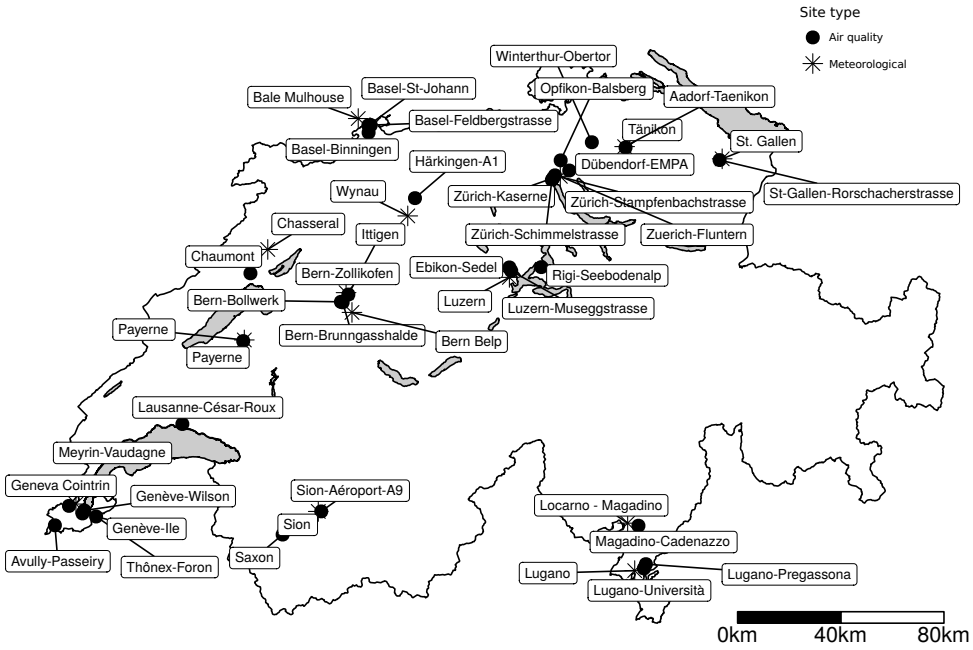

Figure 5.2: Locations of the air quality and meteorological sites included in the analysis. The map outline is the extent of Switzerland.

cessed from the Integrated Surface Database (ISD) with the **worldmet** R package.[46,47] These observations are generally available as hourly means and were therefore aggregated to daily averages. The wind speed aggregation used was the scalar averages which represents average atmospheric motion well at this aggregation period.[48] Generally, the closest ISD site with a complete time series was matched to an air quality monitoring site, but there were cases where the data record was poor for the closest site, or it was unrepresentative (usually due to large differences in elevation) so another ISD site was used instead. Some air quality monitoring sites monitor meteorological variables, but often the time series were not complete in the ISD database and another site was therefore supplemented. Fourteen unique ISD sites were used and Table 5.1 shows which ISD site was used for each of the 31 air quality monitoring sites.

Synoptic scale weather patterns were included into the models by using the Swiss Weather Type Classifications (WTC). The WTC is an objec-

tive and automatic classification scheme which is used to describe broad synoptic scale circulation patterns in Switzerland. There are ten different WTCs types but only the CAP9 classification was used which defines nine distinct clusters of synoptic weather patterns calculated by principal component analysis.[49] Descriptions of what these nine classes represent are displayed in Table 5.2.

Table 5.2: The nine synoptic scale weather type classifications (WTC) used in this study.[49]

CAP9 class	CAP9 description
1	North-East, indifferent
2	West-South-West, cyclonic, flat pressure
3	Westerly flow over Northern Europe
4	East, indifferent
5	High Pressure over the Alps
6	North, cyclonic
7	West-South-West, cyclonic
8	High Pressure over Central Europe
9	Westerly flow over Southern Europe, cyclonic

Modelled daily boundary layer heights between 1997 and 2016 were sourced from the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA-Interim data portal.[50] The highest spatial resolution outputs were used which were at 0.125×0.125 decimal degrees. The NetCDF ECMWF model outputs were promoted to a raster stack and the midday boundary layer heights were extracted for each of the 31 monitoring sites.[51,52] Many of the Swiss urban monitoring sites are within close proximity and therefore only 23 unique raster cells were needed to represent the 31 sites. After the raster extraction, daily time series of boundary layer heights for each site were generated. The modelled ECMWF outputs were tested against radio sounding observations at Payerne before 2010 when such data exists. Although the two data sets did not agree well, a positive correlation

was present and inclusion of boundary layer variable was done to allow the models to have a predictor which represented approximate atmospheric stability and the modelled data was judged to be suitable for this purpose.

For each of the 23 raster cells, daily back trajectories were calculated using the HYSPLIT model for the monitored period of PM_{10} (1997–2016).[53] The back trajectories were calculated backwards in time for 120 hours and used half the mean monthly boundary layer height as their starting height. This start height ensured that the back trajectory receptor was aloft, but remained within the boundary layer throughout the year. The back trajectories were then clustered into six clusters using the Euclidian distance and these clusters were used to represent the common air masses the PM_{10} monitoring sites were exposed to. The use of six clusters was a heuristic, but the six clusters represented distinct air masses and they were very stable across the 23 receptor locations. The HYSPLIT clustering function in **openair** was used to determine these clusters.[54]

5.3.2 Modelling

RF models which used PM_{10} as the dependent variable for each of the 31 air quality monitoring sites were grown. All RF models used the same explanatory variables to predict daily PM_{10} concentrations. The explanatory variables were: wind speed, wind direction, atmospheric temperature, synoptic weather pattern, boundary layer height, air mass cluster based on the HYSPLIT back trajectories, a linear trend term which was the Unix time of the observation (number of seconds since 1 January, 1970), Julian day (day of the year) as the seasonal term, and day of the week. The air mass cluster, the synoptic weather pattern, and day of the week variables were categorical variables while all others were numeric. All variables were used within their response scale with no transformations being applied. The **randomForest** R package was used as the interface to the RF functions reported by Breiman [30].[55] A daily PM_{10} concentration was only modelled if valid wind speed data was available for that day. For all other input variables,

missing data was imputed with the median of numeric variables and the mode for categorical variables. Training of the models was conducted on 80 % of the input data and the other 20 % was withheld from the training and used to validate the models once they had been grown.

RF only requires a handful of tuning parameters (also called hyper parameters) to be specified by the user.[29,55] To determine the optimal values, many models were run with different combinations of tuning parameters. The model performance statistics using the testing set (data withheld from the training step) and run times were evaluated to judge what hyper parameters grew the best performing models. For this application, the models were found to be rather insensitive to tuning parameters. However, the number of variables used to grow a tree was set to three, the minimum node-size or depth was five, and the number of trees within a forest was set at 300 for all models.

5.3.2.1 Meteorological normalisation

The meteorological normalisation of the daily PM_{10} time series was achieved by repeatedly sampling and predicting using individual site RF models, rather than attempting to solve-for, and then remove the short term variation in a time series. The RF predictive model for a site was used to predict every PM_{10} concentration 1000 times. For every prediction, the explanatory variables, with the exception of the trend term, were sampled without replacement and randomly allocated to a dependent variable observation (a PM_{10} concentration). The 1000 predictions were then aggregated using the arithmetic mean and this represented "average" meteorological conditions and hence, this was the meteorologically normalised trend. If more than a thousand predictions were made, only a very minor reduction of noise was achieved. The functions used to grow the RF models and apply the meteorological normalisation procedure reported here are available in the **normalweatherr** R package.[56]

5.3.3 Trend tests

After the normalised time series for a site had been calculated, formal trend tests were preformed. The Theil-Sen estimator accounting for autocorrelation was used at the 95 % confidence level ($\alpha = 0.05$) to indicate a significant trend. The autocorrelation consideration process results in more conservative confidence intervals for the trend estimates. These functions were also provided by the **openair** R package.[54]

5.4 Results and discussion

5.4.1 Random forest model evaluation

The predictive random forest (RF) models performed well for most PM_{10} monitoring sites. All mean squared errors (MSE) and R^2 values are displayed in tabular form in Table 5.3. R^2 values ranged from 54 to 71 % (Figure 5.3). This indicates for some sites in Switzerland PM_{10} concentrations could be well explained by a combination of surface meteorological conditions, boundary layer height, synoptic scale conditions, back trajectory receptor air mass clusters, and time variables which acted as proxies for emission strength. There were only two obvious patterns observed between site type and predictive model performance: the rural motorway sites performed in a similar way and the rural mountain sites' models generally performed worse than other site types when using the R^2 metric. However, there were only two of each of these site types analysed in this study, and the other four site types did not demonstrate any conclusive grouping with model performance measures (Figure 5.3).

The most important explanatory variable for PM_{10} concentrations depended on which site was being investigated. However, generally, wind speed was the variable with the greatest importance for prediction (Figure 5.4). Variable importance was defined as the difference between the MSE of each tree in the forest, minus the MSE for each independent variable,

Table 5.3: Random forest model performance statistics for 31 PM_{10} air quality monitoring sites in Switzerland.

ID	Site	Site type	MSE	R^2 (%)
1	Avully-Passeiry	Rural	54.824	59.980
2	Magadino-Cadenazzo	Rural	129.356	56.898
3	Payerne	Rural	60.854	62.431
4	Saxon	Rural	64.023	62.097
5	Tänikon	Rural	51.140	67.523
6	Härkingen-A1	Rural motorway	84.145	65.531
7	Sion-Aéroport-A9	Rural motorway	53.355	64.646
8	Chaumont	Rural mountain	26.095	61.019
9	Rigi-Seebodenalp	Rural mountain	32.276	53.513
10	Basel-Binningen	Suburban	65.807	64.247
11	Dübendorf-EMPA	Suburban	64.563	63.084
12	Ebikon-Sedel	Suburban	68.702	54.373
13	Ittigen	Suburban	68.965	64.415
14	Lugano-Pregassona	Suburban	84.349	55.492
15	Meyrin-Vaudagne	Suburban	52.188	59.037
16	Opfikon-Balsberg	Suburban	57.011	62.900
17	Thônex-Foron	Suburban	61.899	66.192
18	Basel-St-Johann	Urban background	63.320	66.413
19	Lugano-Università	Urban background	173.909	55.792
20	Luzern-Museggstrasse	Urban background	89.484	62.690
21	Winterthur-Obertor	Urban background	68.498	57.971
22	Zürich-Kaserne	Urban background	73.583	61.867
23	Basel-Feldbergstrasse	Urban traffic	62.058	63.296
24	Bern-Bollwerk	Urban traffic	94.146	67.708
25	Bern-Brunngasshalde	Urban traffic	66.208	57.540
26	Genève-Ile	Urban traffic	66.777	59.299
27	Genève-Wilson	Urban traffic	80.017	62.025
28	Lausanne-César-Roux	Urban traffic	80.206	61.248
29	St-Gallen-Rorschacherstrasse	Urban traffic	55.139	60.131
30	Zürich-Schimmelstrasse	Urban traffic	91.317	70.609
31	Zürich-Stampfenbachstrasse	Urban traffic	75.976	61.974

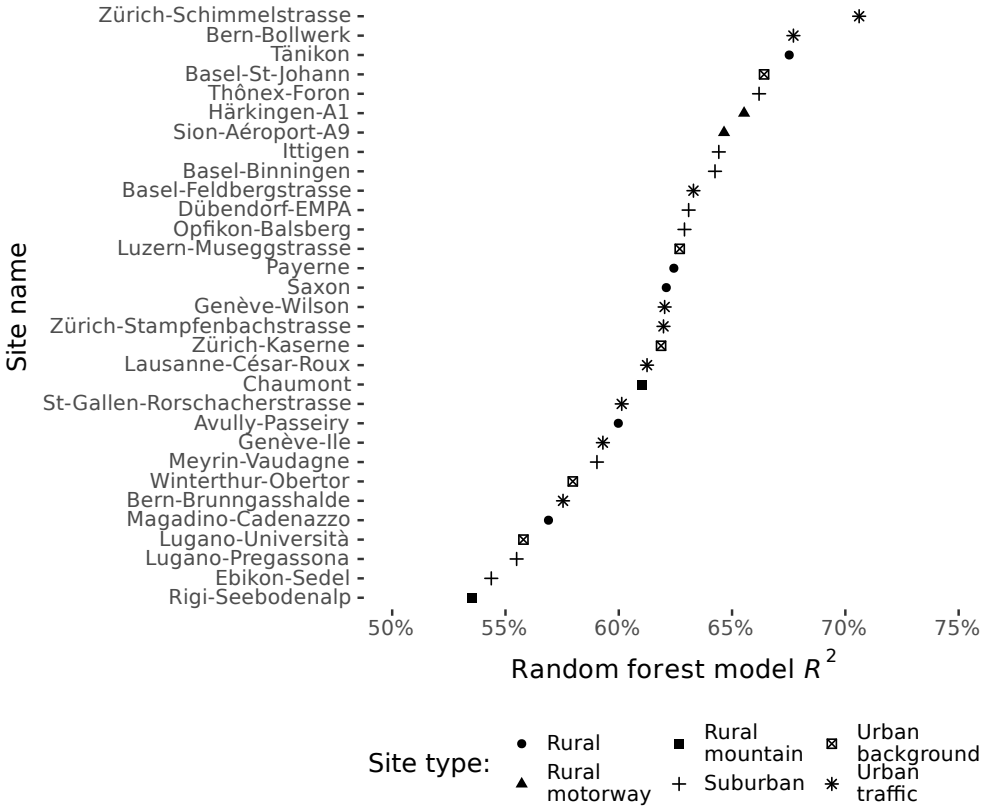

Figure 5.3: The R^2 values for the 31 random forest models grown for the Swiss PM_{10} monitoring sites.

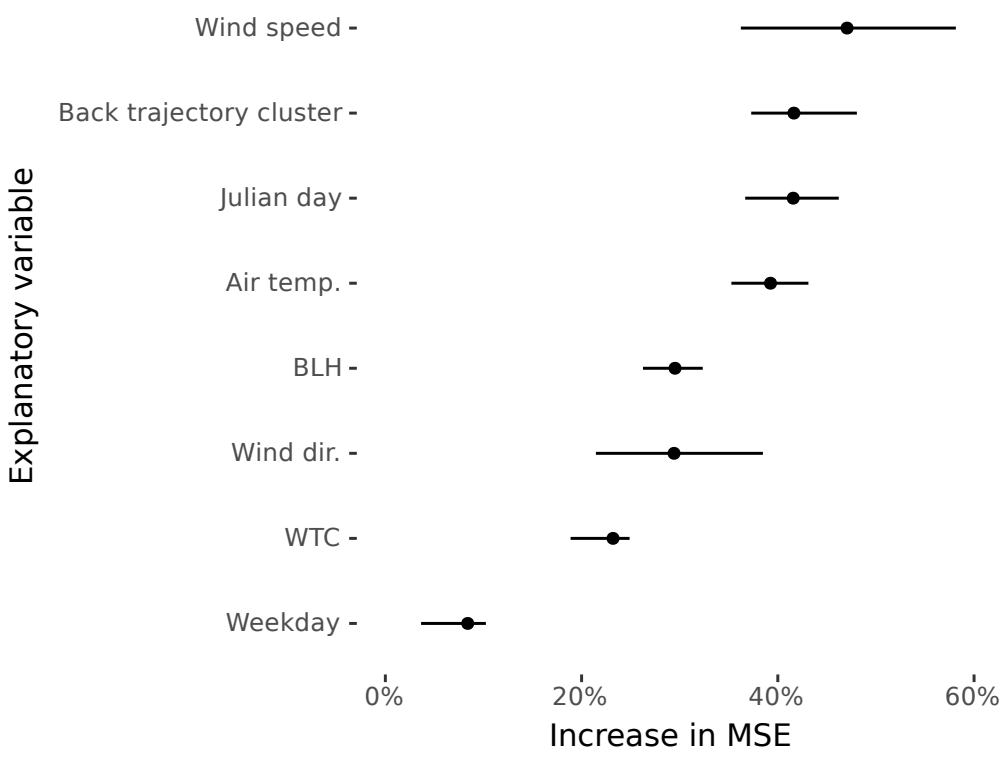

Figure 5.4: Variable importance for the 31 Swiss PM_{10} monitoring sites' random forest models. Dots represent the mean increase in mean square error (MSE) and the lines represent the interquartile range for each variable.

scaled by the standard deviation of the differences.[55] Other sites demonstrated that the seasonal term (Julian day), or trajectory cluster were the most important variables to explain variability in PM_{10} concentrations (Figure 5.4). This indicates that both local and regional scale processes were important when explaining PM_{10} concentrations in Switzerland. Day of the week and the synoptic-scale classification (WTC) were generally the least important variables in the RF models, but both variables always contributed to the models' predictive ability (Figure 5.4). Including variables with little predictive power does not negatively effect the performance of RF models and therefore there was no attempt to remove such variables from the models. Interestingly, wind direction was often a relatively unimportant variable (Figure 5.4). This may be due to daily wind direction averages not contributing much information gain in the model because the aggregation period results in the metric representing atmospheric motion rather poorly. For all of the 31 sites, the normalised PM_{10} was approximately monotonic and no seasonal component was apparent which made formal trend tests suitable.

5.4.2 PM_{10} trend analysis

In all but two PM_{10} Swiss monitoring sites, normalised PM_{10} concentrations were found to be significantly decreasing at the $\alpha = 0.05$ level between 1997 and 2016. Significantly decreasing normalised PM_{10} trends at individual sites ranged from -0.09 to -1.16 $\mu g\,m^{-3}$ $year^{-1}$ (Figure 5.5). These values were similar to the normalised trends reported by Barmpadimos et al. [39] of -0.15 to -1.2 $\mu g\,m^{-3}$ $year^{-1}$ which analysed Swiss PM_{10} trends between 1991 and 2008 with a different method (general additive models; GAMs). The similarities between the two studies suggests that PM_{10} concentrations have continued to reduce at the same rate as reported in the past, which also validates the performance of emission control measures relating to vehicular and heating PM emissions and confirms the trends that were modelled based on emission inventories and their projections.[57] Luzern-

Museggstrasse was the only monitoring site which demonstrated a significantly increasing normalised PM_{10} trend of 0.14 $\mu g\,m^{-3}$ $year^{-1}$. However, this facility stopped monitoring PM_{10} in 2009 and therefore it is unknown if this trend continued to more recent times. The two monitoring sites in Geneva also did not have PM_{10} observations to the end of the analysis period. PM_{10} at Genève-Wilson demonstrated no significant normalised trend and Genève-Ile had the least significant normalised PM_{10} trend across the 31 sites analysed (Figure 5.5). This may suggest that Geneva's PM_{10} trends are different from the rest of Switzerland, but with the lack of more recent observations, this is uncertain.

Sites classified as "urban traffic" had a greater decreasing trend when compared to other site types (Figure 5.6). When the six site type trends were aggregated together, the stronger decreasing trend for traffic sites was clear with an average trend of -0.77 $\mu g\,m^{-3}$ $year^{-1}$, compared to the other site types which ranged between -0.39 and -0.63 $\mu g\,m^{-3}$ $year^{-1}$ (Figure 5.6). Barmpadimos et al. [39] also reported trends based on site type but their site type definitions were not the same as used in this study so they should not be directly compared. The higher first four points in the rural panel of Figure 5.6 was caused by the aggregated time series only containing the Magadino-Cadenazzo monitoring site at the very beginning of the analysis period. Magadino-Cadenazzo is located south of the Alps and experiences higher average concentrations of PM_{10} compared to the other rural sites. Without the observations from the other rural sites, these higher concentrations leveraged the mean seen in Figure 5.6. These observations were still included in the analysis and the Theil-Sen estimator used is hardened against outliers so this will have minimal influence on the trend estimate.

Difference in annual mean PM_{10} concentrations between the rural and urban traffic site types for 2016, the final year of analysis, was 4.7 $\mu g\,m^{-3}$ compared to 9.8 $\mu g\,m^{-3}$ in 1997. The deltas between rural and other site types (excluding the mountainous sites) also decreased during the analysis period. This suggests the locations which are influenced by immediate PM_{10}

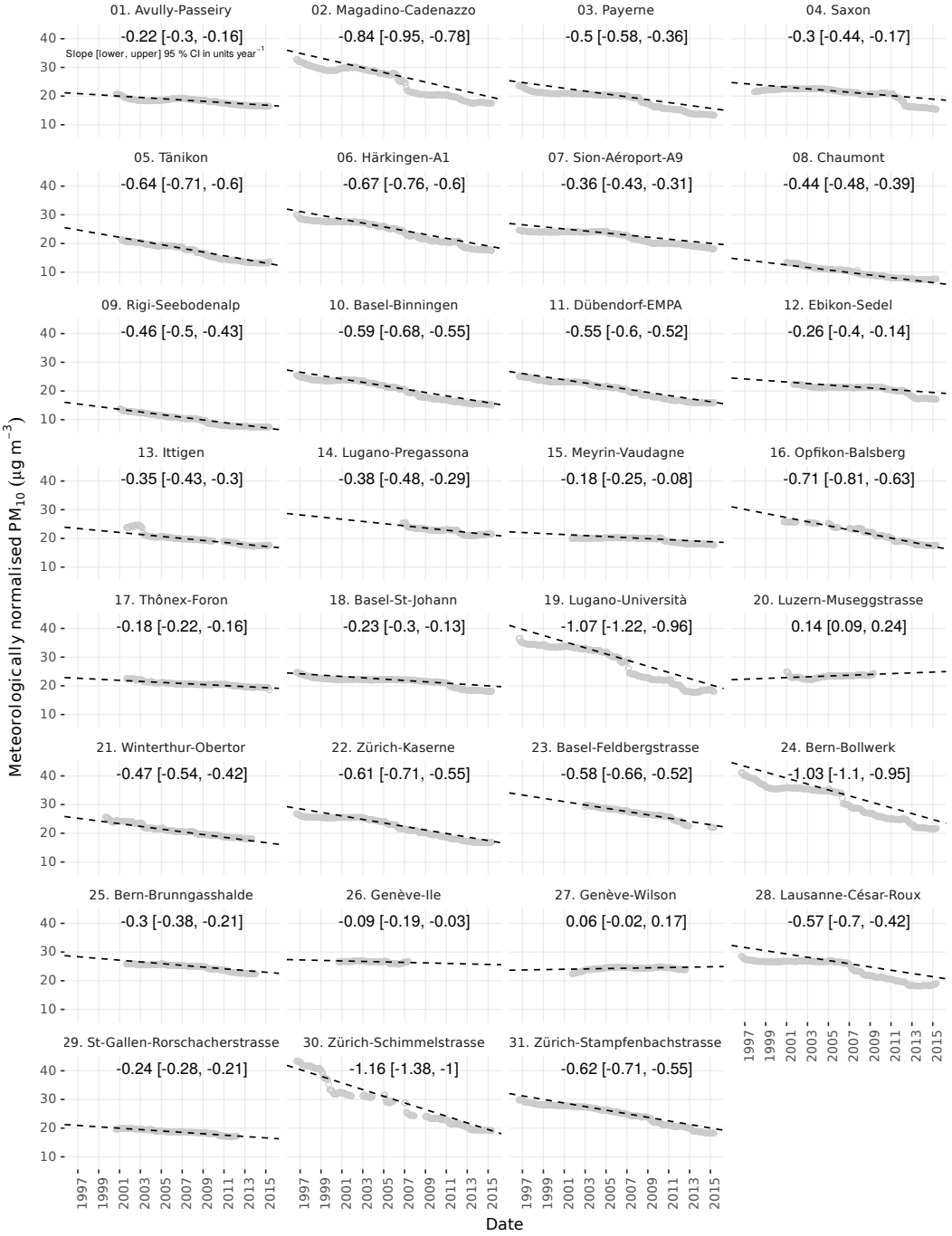

Figure 5.5: Meteorologically normalised PM_{10} trends for the 31 sites analysed in Switzerland between 1997 and 2016. The annotation format is: Slope [Lower 95% CI, Upper 95% CI] in $\mu g\,m^{-3}$ year $^{-1}$.

sources are becoming less polluted by local emissions and are increasingly heading towards rural background levels. The rural and urban background sites' trend metrics are very similar indicating that these two site types are behaving in a very similar way in respect to changes to PM_{10} concentrations over time.

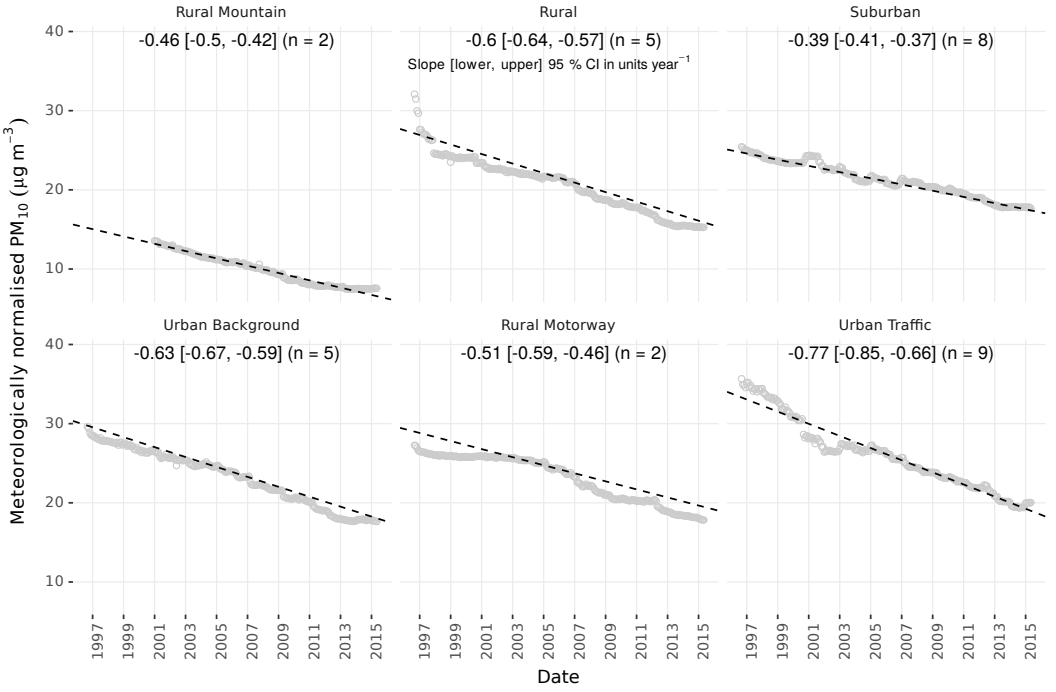

Figure 5.6: Aggregated meteorologically normalised PM_{10} trends for the six site types in Switzerland between 1997 and 2016. Points represent the aggregated meteorologically normalised monthly means and lines represent the trend estimate. The annotation format is: Slope [Lower 95% CI, Upper 95% CI] in $\mu g\,m^{-3}$ year $^{-1}$, and n represents the number of sites in the group.

The site type classifications used in this study can be sorted by their increasing anthropogenic PM_{10} load in this order: rural mountain, rural, suburban, urban background, and urban traffic. Site types which experience more anthropogenic PM_{10} emissions could be expected to demonstrate greater reductions in PM_{10} concentrations when emission inventions or controls are applied. This continuum is only partially shown in the trend mag-

nitudes however with suburban and rural motorway sites not conforming to this expected pattern (Figure 5.6). In fact, the suburban sites demonstrate the smallest decrease in PM_{10} concentrations.

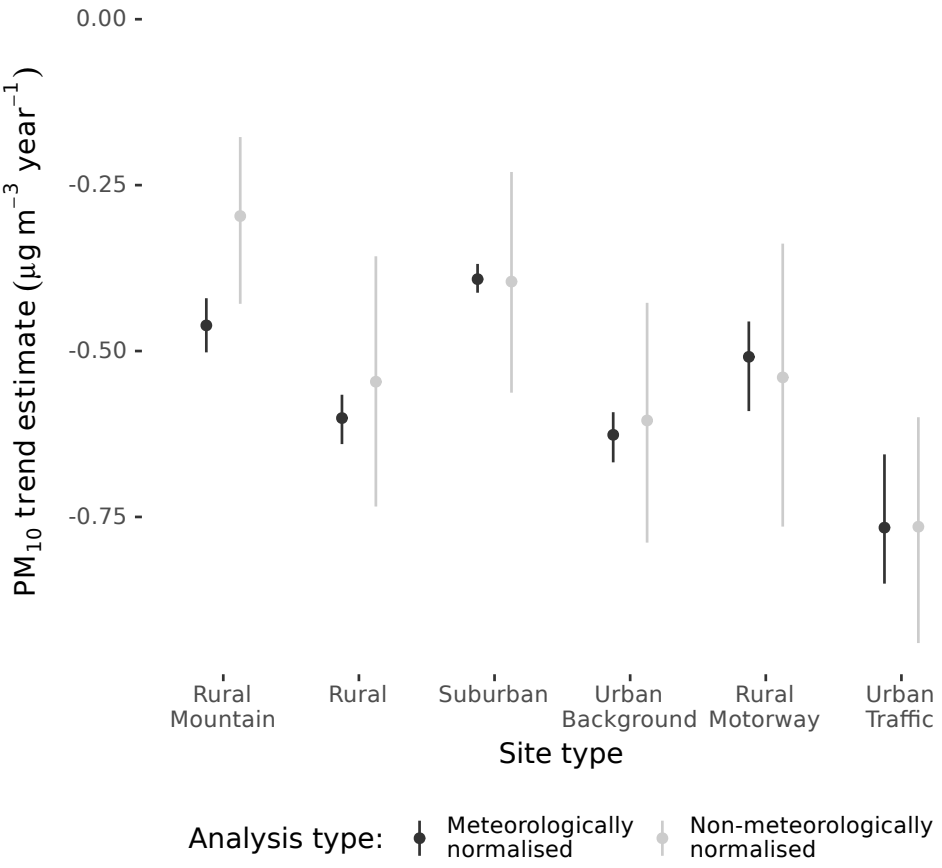

Figure 5.7: PM_{10} trend slope estimates of meteorological normalised and non-meteorological normalised observations for five site types in Switzerland between 1997 and 2016. The line ranges represent the 95 % confidence intervals of the slope estimates.

The rural motorway trends can be explained because although PM (tailpipe) emissions for road traffic have decreased in Switzerland between 1997 and 2016, the volume of traffic using the adjacent roads has increased.[58] This increase in traffic would have offset the lower emissions during the time period and thus PM concentrations would not have decreased as much as could be expected based on vehicular emissions alone. The suburban sites' lack of decrease is more difficult to explain. There are many processes which

could explain this feature, but we attribute this result due to changes in the surrounding environment of the suburban sites. Many of the monitoring sites in Switzerland which are classed as suburban have become increasingly urban during the period of analysis (1997 and 2016). Therefore, some of these suburban monitoring sites are being influenced by more urban-like processes and emissions due to the development in their vicinity.

Woodburning is a source of PM_{10} in the alpine, suburban, and urban areas in Switzerland. The number of woodburning appliances and heating demand is decreasing over time and this change will contribute to the trends observed in Figure 5.6.[59] However, the quantification of the reduction in woodburning activity on PM_{10} concentrations among the different site types is cannot be conducted with the current data concerning woodburner usage.

The comparison of the RF meteorological normalisation models with other techniques was not a primary objective of this work. However, it is important to consider what effect meteorological normalisation had on the trend estimates. To investigate this, the PM_{10} observations which were subjected to the meteorological normalisation process were aggregated to monthly means and their trends tested with the Theil-Sen test with identical parameters as used on the normalised time series. This could be considered a "standard" and routine procedure for air quality data analysis. With the exception of the rural motorway sites, the normalised trend estimate was found to be greater (more negative), than the non-normalised trend estimates (Figure 5.7). This indicates that meteorology in Switzerland between 1997 and 2016 has masked or obscured changes in PM_{10} emissions during the same period in the observational record. Because the meteorological normalisation technique helps to explain variation in PM_{10} concentrations, the normalised trend estimates had a much lower range of uncertainty when compared to the aggregated observations in all cases (Figure 5.7). Therefore, not only did the meteorological normalisation technique generally estimate more negative trends compared to standard methods, the trends calculated

were more robust and less uncertain when compared to a routine analysis method which would lead to quicker identification of significant trends.

5.4.2.1 Explaining the observed trends

One of the primary advantages of decision tree methods like RF over other machine learning techniques is the ability to interpret and explain the models and discussion of this is presented in Section 5.2.3.1. Here, this advantage will be leveraged to help explain some of the features in the PM_{10} trends in Switzerland between 1997 and 2016.

Partial dependence plots allow RF models to be evaluated and to confirm how the explanatory variables are being used in the models for prediction.[38] For the application presented here, there are general physical and chemical processes which should be confirmed in the RF models. For example, it can be expected that PM_{10} concentrations will be inversely related to wind speed due to increased atmospheric dispersion, and that wintertime concentrations will be higher than other seasons resulting from a combination of greater emissions and atmospheric stability. These general predictions and processes were confirmed by the RF models' partial dependence plots (one site shown as an example in Figure 5.8).

The partial dependence plots of the Zürich-Stampfenbachstrasse RF model (Figure 5.8) showed some interesting features and were typical for Switzerland's traffic influenced sites. The y (vertical) axis for each plot represents the dependence of PM_{10} concentration on one variable if all other variables are fixed at their average level. The most important variable at this location was wind speed and the non-linear relationship is present in Figure 5.8. When wind speeds were very low, the PM_{10} concentrations were on average over 38 $\mu g\,m^{-3}$ day^{-1} but the influence on PM_{10} concentrations was strong and therefore at wind speeds greater than 3 $m\,s^{-1}$, average concentrations decreased to under 22 $\mu g\,m^{-3}$ day^{-1} (Figure 5.8). There was minimal evidence of increasing PM_{10} concentrations at high wind speeds due to resuspension of wind blown PM at any monitoring site in the RF models.

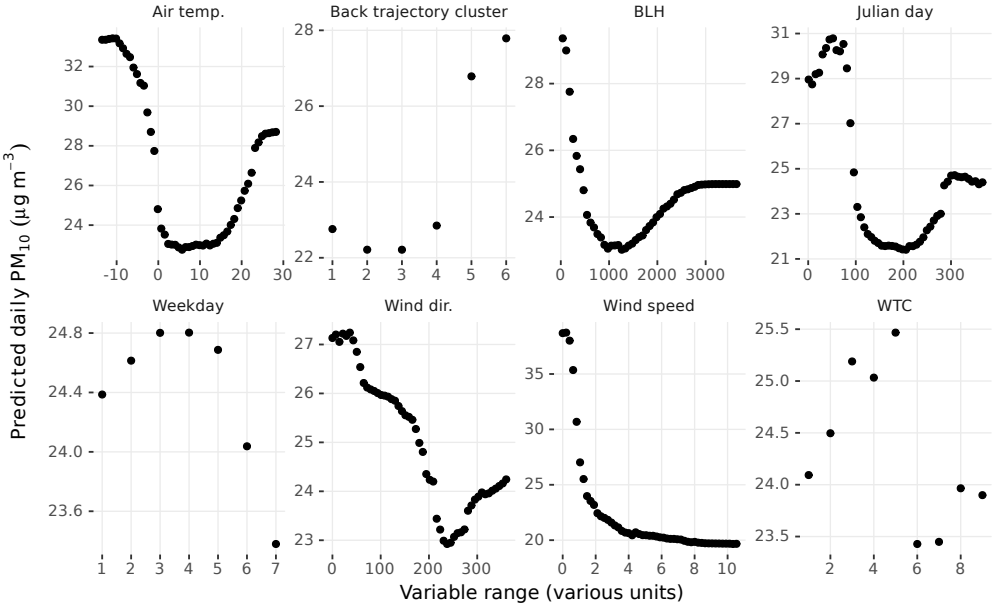

Figure 5.8: Partial dependence plots of the explanatory variables used in the Zürich-Stampfenbachstrasse PM_{10} random forest model.

Weekday was the variable of least importance for the Zürich- Stampfenbachstrasse RF model but the partial dependence plot still demonstrates what would be expected. Weekdays (days 1–5; Monday–Friday) were more polluted than the weekend days due to higher traffic sourced emissions, but the variability of PM_{10} concentrations among the weekdays was less than 2 $\mu g\,m^{-3}$ day^{-1}, *i.e.*, the response scale was small (Figure 5.8). There was evidence of a sequential loading process over the weekdays which peaked on Thursdays (day 4) and also lower concentrations during the early working week (Monday and Tuesday; days 1 and 2) which resulted from reduced precursor PM emissions during the weekend, especially Sunday.

The seasonal component represented by Julian day showed a similar pattern to air temperature (Figure 5.8). Despite the similar shapes of dependencies on PM_{10} for these variables, they represent rather different processes. The Julian day dependence represents the changes in local and regional emissions which influence PM_{10} concentrations over the course of the year. In the case of Zürich-Stampfenbachstrasse, this will be dominated by changes in regional background concentrations with the addition of lo-

cal traffic emissions. The seasonal variation of emissions which effect PM_{10} concentrations at Zürich-Stampfenbachstrasse spans 10 $\mu g\,m^{-3}$, and this indicates that the seasonal effect is important to consider. When Julian day was removed from the RF models, the dependence on air temperature and boundary layer height did not change and this shows that the models were able to differentiate the different processes correctly despite their collinearity.

The back trajectory cluster variable was important for many PM_{10} monitoring sites including Zürich-Stampfenbachstrasse (Figure 5.4 and 5.8). The decoded clusters' descriptions displayed in Figure 5.8 can be found in Table 5.4 but the two most polluted air masses, 5 and 6 represented a local flow from south west Switzerland and a strong north east flow from Poland and southern Germany respectively (Figure 5.9). This indicates that air masses from surrounding European states can cause polluted PM_{10} conditions in Zürich, as can periods of calm and localised flows.

Table 5.4: The six decoded HYSPLIT back trajectory clusters. The integer cluster key was used in the random forest models and the decoded cluster was determined after the cluster analysis.

Cluster	Decoded cluster
1	Strong northerly flow from north sea
2	Very strong north west flow from Atlantic Ocean
3	Westerly flow from Atlantic Ocean
4	South west flow from France and western Switzerland
5	Local flow from south west Switzerland
6	Strong north east flow from Poland and southern Germany

The partial dependence plots indicate that most monitoring sites experience their minimum PM_{10} concentrations when the boundary layer is $\approx$ 1000 metres high, but concentrations increase again once the boundary layer increases over 2000 metres (Figure 5.8). This is an interesting phenomenon and it suggests that there are two regimes in Switzerland which

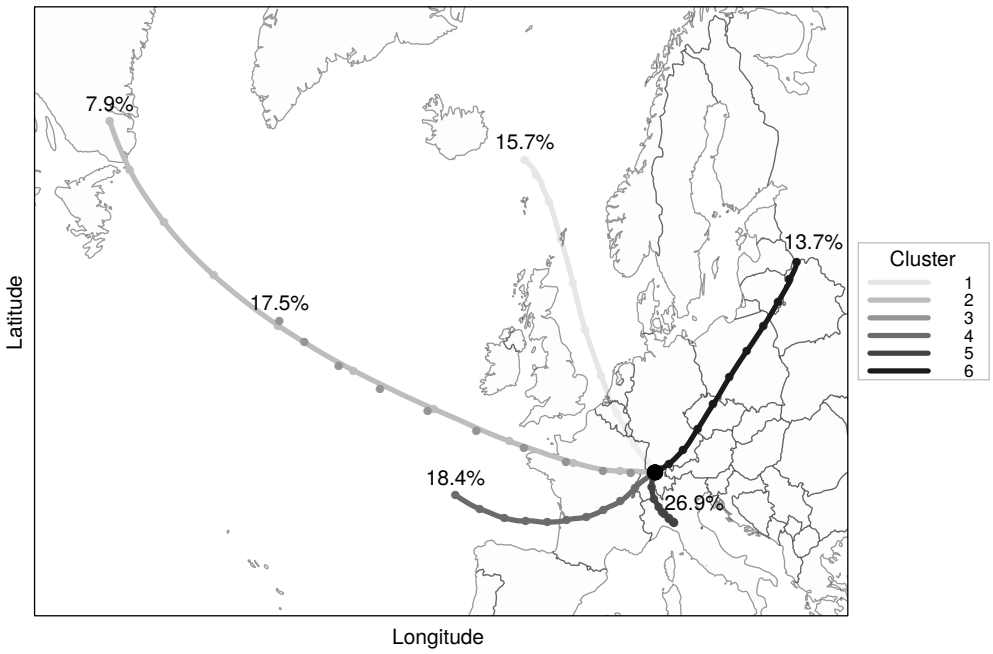

Figure 5.9: The six back trajectory clusters for the Zürich receptor location between 1997 and 2016 which were used by the random forest PM_{10} models. The clusters are decoded in Table 5.4 and the percentages indicate the frequency of occurrence.

drive elevated PM_{10} concentrations. The first is the obvious (and expected) combination of low temperatures, low boundary heights, and high rates of surface-based emissions during wintertime. These factors combine to create a poor dispersive environment which leads to high pollutant concentrations. The second regime which causes elevated PM_{10} concentrations is active when temperatures are above 20 °C and the boundary layer is above 2000 metres (Figure 5.8). These conditions occur with every air mass cluster and under all synoptic weather patterns which are experienced at these higher temperatures. Therefore, this regime is associated with warm, dry, dusty, and deep convective boundary layer conditions which favour transportation of PM_{10} from other locations and the generation of secondary aerosol and other processes driven by photochemistry. Daily sulfur (in PM_{10}) observations are available at the Payerne monitoring site and SO_4^{2-} concentrations do indeed increase at higher boundary layer heights while primary pollutants such as NO_x do not (Figure 5.10). These results are consistent with enhanced sulphate formation in summertime when the formation of sulphate through photochemistry is most important. By contrast, the concentration of primary pollutants such as NO_x tend to decrease with increasing boundary layer height due to increased mixing.

The partial dependence plots of the seasonal and trend components also demonstrate that while the trend component decreased between 1997 and 2016, the seasonal component also decreased at some of the Swiss PM_{10} monitoring sites. The best example of this was demonstrated at Magadino-Cadenazzo, a rural site in Ticino in the south of Switzerland (Table 5.1 and Figure 5.2). The decrease in the seasonal component was especially true after 2006 and during early winter at Magadino-Cadenazzo (December; Figure 5.11(a)). As discussed above, this further validates air pollutant emission controls and interventions because both the background concentration and the local loading of PM_{10} during winter is decreasing simultaneously. There is evidence however that the wintertime loading has plateaued since approximately 2014 at this monitoring site (Figure 5.11(b)).

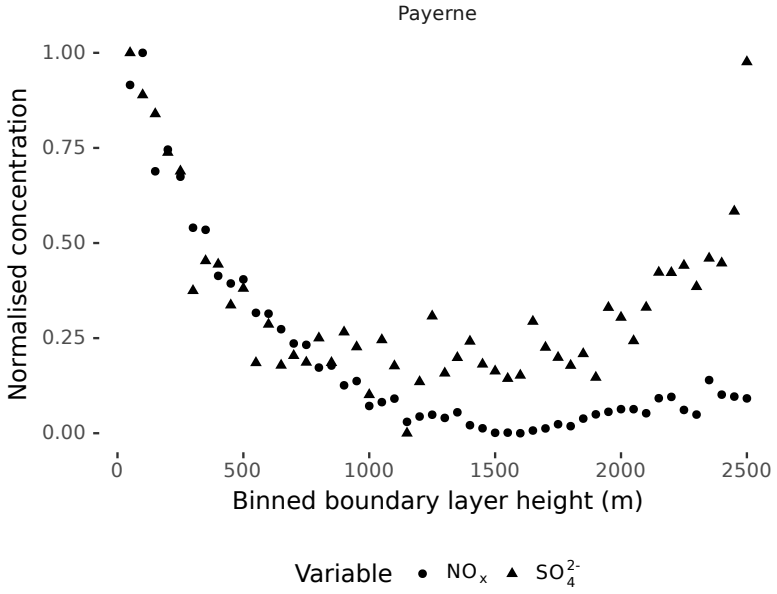

Figure 5.10: Mean normalised concentrations of SO_4^{2-}, a secondary PM species and NO_x for binned boundary layer heights (bin was set at 50 metres) at Payerne between 1997 and 2016.

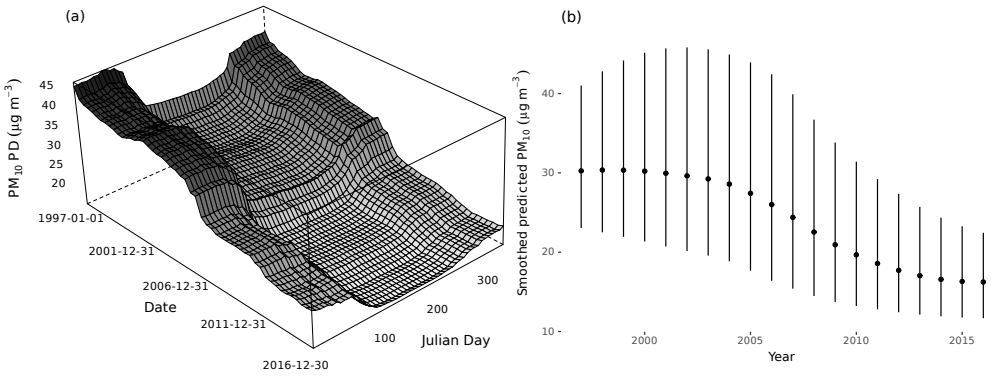

Figure 5.11: (a) PM_{10} partial dependence on trend and seasonal components (Date and Julian day respectively) and (b) annual predicted seasonal component at Magadino-Cadenazzo where dots represent the mean and lines indicate the amplitude of the seasonal component.

The rural mountain Chaumont and Rigi-Seebodenalp monitoring sites have low PM_{10} concentrations when compared to the other Swiss sites and site types (Figure 5.5 and Figure 5.6). Both of these locations are isolated and are located above 1000 metres of elevation (Table 5.1 and Figure 5.2). Therefore, these two monitoring sites represent pristine locations. The PM_{10} concentrations at both locations decreased at $\approx$ -0.45 $\mu g\,m^{-3}$ $year^{-1}$ between 1997 and 2016 indicating a wider-scale European reduction in PM_{10} and its precursors.[13] Interestingly, the normalised trend at Rigi-Seebodenalp showed an additional PM_{10} loading between April 8 and 26, 2010 due to the Eyjafjallajökull Icelandic volcanic eruption[60,61] but at Chaumont, this was not discernible (not shown). This demonstrates that the two sites do behave differently and are exposed to different processes at times. The differences between the two sites are not clear in the concentration data alone and demonstrates a potentially useful side effect of the technique where it can be used to investigate abnormal events.

The RF models for these two rural and mountainous locations also demonstrated different processes compared to other site types. The most interesting feature was that the relationship between air temperature and boundary layer height with PM_{10} concentrations differed from the other Swiss monitoring sites. The two mountainous sites experienced their highest PM_{10} concentrations at high temperatures (Chaumont shown in Figure 5.12(a)). This difference in dependence was due to these monitoring locations being intermittently above the boundary layer, which was also confirmed with the boundary layer height partial dependence plots (Figure 5.12(b)). When these elevated sites were within the boundary layer during warmer periods, the relatively well mixed PM_{10} influenced the monitoring locations, but during cooler times, the sites were located in the free troposphere decoupled from surface based emissions. This generally resulted in the elevated monitoring sites experiencing lower concentrations of PM_{10} during cooler periods which was not the case for monitoring sites located at lower elevations, for example, Basel-St-Johann, an urban background site located

at 260 metres of elevation (Figure 5.12).

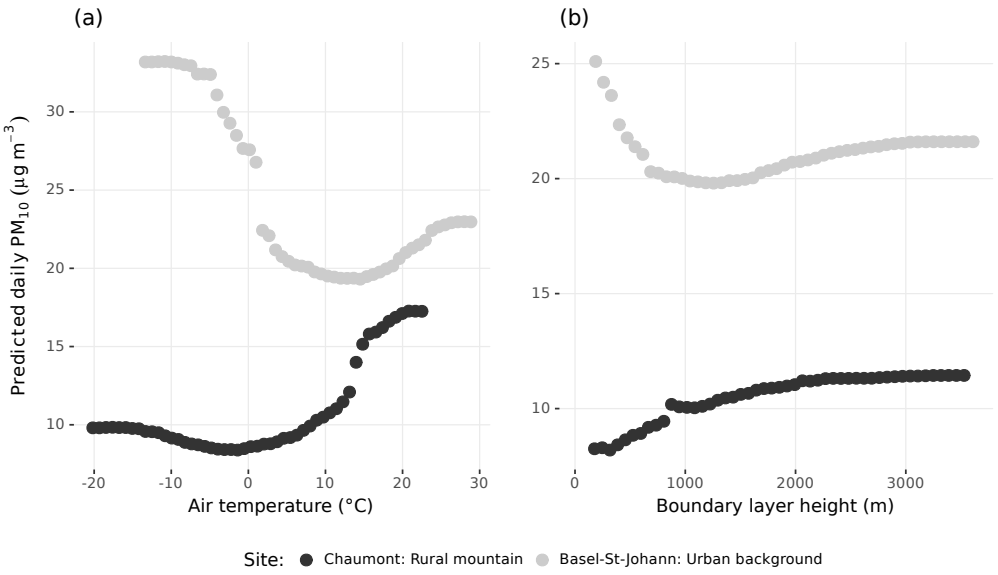

Figure 5.12: Partial dependence of PM_{10} concentrations on (a) air temperature and (b) boundary layer height at two monitoring sites with different site type classifications.

5.5 Conclusions

This paper presented a meteorological normalised PM_{10} trend analysis using daily data from Switzerland. Random forest (RF) predictive models which were used to explain variation of PM_{10} concentrations using surface meteorology, synoptic scale weather patterns, boundary layer height, back trajectory clusters, and time variables. The models were then used to prepare the PM_{10} time series to create a meteorological normalised trend which was suitable for formal trend analysis.

The RF performed well for the 31 monitoring sites with R^2 values up to 71 %. Wind speed, Julian day (the seasonal component), and back trajectory cluster were generally the most important predictors for PM_{10} concentration. For 29 of the 31 monitoring sites analysed, PM_{10} concentrations were found to be significantly decreasing at rates between -0.09 and

-1.16 $\mu g\,m^{-3}$ $year^{-1}$ and on average, urban traffic sites demonstrated the greatest decrease of -0.77 $\mu g\,m^{-3}$ $year^{-1}$. The RF models' learning process was interpreted with partial dependence plots to explain the trends observed. There was evidence of a decrease in the seasonal component at some sites, *i.e.*, the wintertime loading has decreased, and the monitoring sites above 1000 metres of elevation showed interesting dependences on air temperature which were not demonstrated at other sites because they are intermittently located above the boundary layer. The models also indicated that across Switzerland, elevated PM_{10} concentrations occur in poor dispersion conditions as well as at high temperatures with a deep boundary layers due to high rates secondary PM generation resulting from photochemical processes.

The meteorological normalisation technique using RF was found to be helpful in the PM_{10} trend analysis conducted and resulted in more negative and less uncertain trend estimates compared to another standard analysis method. The predictive modelling framework and technique was found to be easy to implement and user friendly because RF does not need to conform to strict parametric assumptions. The technique described could be used in many air quality exploratory data analysis applications.

5.6 References

[1] Porter, P. S., Rao, S. T., Zurbenko, I. G., Dunker, A. M., and Wolff, G. T. Ozone Air Quality over North America: Part II: An Analysis of Trend Detection and Attribution Techniques. *Journal of the Air & Waste Management Association* 51.2 (2001). PMID: 28060596, pp. 283–306. DOI: 10.1080/10473289.2001.10464261. URL: http://dx.doi.org/10.1080/10473289.2001.10464261.

[2] Rao, S. T. and Zurbenko, I. G. Detecting and Tracking Changes in Ozone Air Quality. *Journal of the Air & Waste Management Association* 44.9 (1994), pp. 1089–1092. DOI: 10.1080/10473289.1994.10467303. URL: http://dx.doi.org/10.1080/10473289.1994.10467303.

[3] Pryor, S., McKendry, I., and Steyn, D. Synoptic-scale meteorological variability and surface ozone concentrations in Vancouver, British Columbia. *Journal of Applied Meteorology* 34.8 (1995), pp. 1824–1833. DOI: 10.1175/1520-0450(1995)034<1824:SSMVAS>2.0.CO;2. URL: https://doi.org/10.1175/1520-0450(1995)034%3C1824:SSMVAS%3E2.0.CO;2.

[4] Libiseller, C. and Grimvall, A. Model selection for local and regional meteorological normalisation of background concentrations of tropospheric ozone. *Atmospheric Environment* 37.28 (2003), pp. 3923–3931. URL: http://www.sciencedirect.com/science/article/pii/S1352231003005028.

[5] Libiseller, C., Grimvall, A., Waldén, J., and Saari, H. Meteorological normalisation and non-parametric smoothing for quality assessment and trend analysis of tropospheric ozone data. *Environmental Monitoring and Assessment* 100.1 (2005), pp. 33–52. DOI: 10.1007/s10661-005-7059-2. URL: http://dx.doi.org/10.1007/s10661-005-7059-2.

[6] Wise, E. K. and Comrie, A. C. Extending the Kolmogorov–Zurbenko Filter: Application to Ozone, Particulate Matter, and Meteorological Trends. *Journal of the Air & Waste Management Association* 55.8 (2005), pp. 1208–1216. DOI: 10.1080/10473289.2005.10464718. URL: http://dx.doi.org/10.1080/10473289.2005.10464718.

[7] Zeldin, M. D. and Meisel, W. S. *Use of Meteorological Data in Air Quality Trend Analysis*. Tech. rep. Publication number: EPA-450/3-78-024. United States Environmental Protection Agency, 1978.

[8] Carslaw, D. C., Ropkins, K., and Bell, M. C. Change-Point Detection of Gaseous and Particulate Traffic-Related Pollutants at a Roadside Location. *Environmental Science & Technology* 40.22 (2006), pp. 6912–6918. DOI: 10.1021/es060543u. URL: http://dx.doi.org/10.1021/es060543u.

[9] Stull, R. B. *An Introduction to Boundary Layer Meteorology*. London: Kluwer Academic Publishers, 1988, 666pp.

[10] Anh, V., Duc, H., and Azzi, M. Modeling anthropogenic trends in air quality data. *Journal of the Air & Waste Management Association* 47.1 (1997), pp. 66–71. DOI: doi:10.1080/10473289.1997.10464406. URL: https://doi.org/10.1080/10473289.1997.10464406.

[11] Lou Thompson, M., Reynolds, J., Cox, L. H., Guttorp, P., and Sampson, P. D. A review of statistical methods for the meteorological adjustment of tropospheric ozone. *Atmospheric Environment* 35.3 (2001), pp. 617–630. URL: http://www.sciencedirect.com/science/article/pii/S1352231000002612.

[12] Marchetto, A., Rogora, M., and Arisci, S. Trend analysis of atmospheric deposition data: A comparison of statistical approaches. *Atmospheric Environment* 64.0 (2013), pp. 95–102. DOI: 10.1016/j.atmosenv.2012.08.020. URL: http://www.sciencedirect.com/science/article/pii/S1352310120079 23.

[13] Guerreiro, C. B., Foltescu, V., and Leeuw, F. de. Air quality status and trends in Europe. *Atmospheric Environment* 98.0 (2014), pp. 376–384. URL: http://www.sciencedirect.com/science/article/pii/S1352310140071 09.

[14] Siegel, A. F. Robust Regression Using Repeated Medians. *Biometrika* 69.1 (1982), pp. 242–244. URL: http://www.jstor.org/stable/2335877.

[15] Hamed, K. H. and Ramachandra Rao, A. A modified Mann-Kendall trend test for autocorrelated data. *Journal of Hydrology* 204.1–4 (1998), pp. 182–196. URL: http://www.sciencedirect.com/science/article/pii/S002216949700125X.

[16] Salmi, T., Määttä, A., Anttila, P., Ruoho-Airola, T., and Amnell, T. Detecting trends of annual values of atmospheric pollutants by the Mann-Kendall test and Sen's slope estimates - the Excel template application MAKESENS. Finnish Meteorological Institute. Publications on air quality. No. 31. 2002.

[17] Meals, D. W., Spooner, J., Dressing, S. A., and Harcum, J. B. Statistical analysis for monotonic trends. Tech Notes 6. Developed for U.S. Environmental Protection Agency by Tetra Tech, Inc., Fairfax, VA. 2011. URL: https://www.epa.gov/sites/production/files/2016-05/documents/tech_notes_6_dec2013_trend.pdf.

[18] Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I. STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Official Statistics* 6.1 (1990). URL: http://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/stl-a-seasonal-trend-decomposition-procedure-based-on-loess.pdf.

[19] Yang, W. and Zurbenko, I. Kolmogorov–Zurbenko filters. *Wiley Interdisciplinary Reviews: Computational Statistics* 2.3 (2010), pp. 340–351. URL: http://dx.doi.org/10.1002/wics.71.

[20] Beevers, S., Carslaw, D., Westmoreland, E., and Mittal, H. *Air pollution and emissions trends in London*. Tech. rep. King's College London, Environmental Research Group Leeds University, Institute for Transport studies, 2009. URL: http://naei.defra.gov.uk/reports/reports?report_id=589.

[21] Carslaw, D. and Priestman, M. *Analysis of the 2013 vehicle emission remote sensing campaigns data*. Tech. rep. King's College London, 2015. URL: https://uk-air.defra.gov.uk/library/reports?report_id=831.

[22] Fuller, G. and Carslaw, D. *Putney High Street air quality. Part 2: Bridge closure, loading and parking changes*. Tech. rep. King's College London & the University of York, 2017.

[23] Derwent, R., Middleton, D., Field, R., Goldstone, M., Lester, J., and Perry, R. Analysis and interpretation of air quality data from an urban roadside location in Central London over the period from July 1991 to July 1992. *Atmospheric Environment* 29.8 (1995), pp. 923–946. DOI: 10.1016/1352-

2310(94)00219-B. URL: http://www.sciencedirect.com/science/article/pii/135223109400219B.

[24] Hitchins, J., Morawska, L., Wolff, R., and Gilbert, D. Concentrations of submicrometre particles from vehicle emissions near a major road. *Atmospheric Environment* 34.1 (2000), pp. 51–59. URL: http://www.sciencedirect.com/science/article/pii/S1352231099003040.

[25] Cox, D. R. Interaction. *International Statistical Review/Revue Internationale de Statistique* (1984), pp. 1–24.

[26] Smola, A. and Vishwanathan, S. V. N. *Introduction to Machine Learning*. Cambridge University Press, Cambridge, United Kingdom, 2008.

[27] Kuhn, M. Predictive Modeling with R and the **caret** Package. useR! conference July 10-12 2013 University of Castilla-La Mancha, Albacete, Spain. 2013. URL: https://www.r-project.org/conferences/useR-2013/Tutorials/Kuhn.html.

[28] Friedman, J. H. Recent Advances in Predictive (Machine) Learning. *Journal of Classification* 23.2 (2006), pp. 175–197. URL: http://dx.doi.org/10.1007/s00357-006-0012-4.

[29] Immitzer, M., Atzberger, C., and Koukal, T. Tree Species Classification with Random Forest Using Very High Spatial Resolution 8-Band WorldView-2 Satellite Data. *Remote Sensing* 4.9 (2012), p. 2661. URL: http://www.mdpi.com/2072-4292/4/9/2661.

[30] Breiman, L. Random Forests. *Machine Learning* 45.1 (2001), pp. 5–32. DOI: 10.1023/A:1010933404324. URL: http://dx.doi.org/10.1023/A:1010933404324.

[31] Tong, W., Hong, H., Fang, H., Xie, Q., and Perkins, R. Decision Forest: Combining the Predictions of Multiple Independent Decision Tree Models. *Journal of Chemical Information and Computer Sciences* 43.2 (2003), pp. 525–531. DOI: 10.1021/ci020058s. URL: http://dx.doi.org/10.1021/ci020058s.

[32] Biau, G., Devroye, L., and Lugosi, G. Consistency of Random Forests and Other Averaging Classifiers. *Journal of Machine Learning Research* 9 (2008), pp. 2015–2033.

[33] Kotsiantis, S. B. Decision trees: a recent overview. *Artificial Intelligence Review* 39.4 (2013), pp. 261–283. URL: http://dx.doi.org/10.1007/s10462-011-9272-4.

[34] freepik.com. *FlatIcon*. 2017. URL: www.freepik.com.

[35] Breiman, L. Bagging predictors. *Machine Learning* 24.2 (1996), pp. 123–140. DOI: 10.1007/BF00058655. URL: http://dx.doi.org/10.1007/BF00058655.

[36] Friedman, J., Hastie, T., and Tibshirani, R. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Second. Vol. 1. Springer series in statistics Springer, Berlin, 2001.

[37] Caruana, R. and Niculescu-Mizil, A. An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 161–168.

[38] Jones, Z. and Linder, F. Exploratory Data Analysis using Random Forests. 73rd annual MPSA conference, April 16-19, 2015, Chicago, United States of America. 2015. URL: https://pdfs.semanticscholar.org/e7b7/3565b07a7f1369a20b1055f222423f0feb34.pdf.

[39] Barmpadimos, I., Hueglin, C., Keller, J., Henne, S., and Prévôt, A. S. H. Influence of meteorology on PM_{10} trends and variability in Switzerland from 1991 to 2008. *Atmospheric Chemistry and Physics* 11.4 (2011), pp. 1813–1835. URL: http://www.atmos-chem-phys.net/11/1813/2011/.

[40] European Environment Agency. *AirBase – The European air quality database (Version 8)*. 2014. URL: https://www.eea.europa.eu/data-and-maps/data/airbase-the-european-air-quality-database-8.

[41] European Environment Agency. *Eionet Central Data Repository*. 2017. URL: http://cdr.eionet.europa.eu/.

[42] Federal Office for the Environment. Messstationen des NABEL — Stations de mesure NABEL. Technischer Bericht NABEL 2013. 2014. URL: https://www.bafu.admin.ch/dam/bafu/en/dokumente/luft/fachinfo-daten/nabel-messstationen.pdf.download.pdf/nabel-messstationen.pdf.

[43] Federal Office for the Environment. National Air Pollution Monitoring Network (NABEL). 2017. URL: https://www.bafu.admin.ch/bafu/en/home/topics/air/state/data/national-air-pollution-monitoring-network--nabel-.html.

[44] Grange, S. K. ***smonitor**: A framework and a collection of functions to allow for maintenance of air quality monitoring data*. 2018. URL: https://github.com/skgrange/smonitor.

[45] Grange, S. K. *Technical note: **smonitor** Europe*. Tech. rep. Wolfson Atmospheric Chemistry Laboratories, University of York, 2017. DOI: https://doi.org/10.13140/RG.2.2.20555.49448/1. URL: https://doi.org/10.13140/RG.2.2.20555.49448/1.

[46] NOAA. Integrated Surface Database (ISD). 2016. URL: https://www.ncdc.noaa.gov/isd.

[47] Carslaw, D. ***worldmet**: Import Surface Meteorological Data from NOAA Integrated Surface Database (ISD)*. R package version 0.7.5. 2017. URL: http://github.com/davidcarslaw/worldmet.

[48] Grange, S. K. Technical note: Averaging wind speeds and directions. 2014. URL: https://drive.google.com/open?id=0B7lRD1eNlhErbWNQeTVPTy1pRWc.

[49] Weusthoff, T. Weather Type Classification at MeteoSwiss—Introduction of new automatic classifications schemes. *Arbeitsberichte der MeteoSchweiz* 235 (2011), 46 pp.

[50] Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., Rosnay, P. de, Tavolato, C., Thépaut, J.-N., and Vitart, F. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society* 137.656 (2011), pp. 553–597. URL: http://dx.doi.org/10.1002/qj.828.

[51] Hijmans, R. J. ***raster**: Geographic Data Analysis and Modeling*. R package. URL: https://CRAN.R-project.org/package=raster.

[52] Pierce, D. ***ncdf4**: Interface to Unidata netCDF (Version 4 or Earlier) Format Data Files*. R package version 1.16. 2017. URL: https://CRAN.R-project.org/package=ncdf4.

[53] Stein, A. F., Draxler, R. R., Rolph, G. D., Stunder, B. J. B., Cohen, M. D., and Ngan, F. NOAA's HYSPLIT Atmospheric Transport and Dispersion Modeling System. *Bulletin of the American Meteorological Society* 96.12 (2015), pp. 2059–2077. DOI: 10.1175/BAMS-D-14-00110.1. URL: http://dx.doi.org/10.1175/BAMS-D-14-00110.1.

[54] Carslaw, D. C. and Ropkins, K. *openair — An R package for air quality data analysis*. *Environmental Modelling & Software* 27–28.0 (2012), pp. 52–61. URL: http://www.sciencedirect.com/science/article/pii/S1364815211002064.

[55] Liaw, A. and Wiener, M. Classification and Regression by **randomForest**. *R News* 2.3 (2002), pp. 18–22. URL: http://CRAN.R-project.org/doc/Rnews/.

[56] Grange, S. K. ***normalweatherr**: Package to conduct meteorological/weather normalisation on air quality data (deprecated)*. 2017. URL: https://github.com/skgrange/normalweatherr.

[57] Heldstab, J., Leippert, F., Wüthrich, P., Künzle, T., and Stampfli, M. PM_{10} and $PM_{2.5}$ ambient concentrations in Switzerland. Modelling results for 2005, 2010, 2020. Federal Office for the Environment. 2013. URL: https://www.bafu.admin.ch/dam/bafu/en/dokumente/luft/uw-umwelt-wissen/pm10_und_pm2_5_immissioneninderschweizzusammenfassung.pdf.download.pdf/pm10_and_pm2_5_ambientconcentrationsinswitzerland.pdf.

[58] Bundesamt für Strassen. Verkehrsentwicklung und Verfügbarkeit der Nationalstrassen — Jahresbericht 2016. Eidgenössisches Departement für Umwelt, Verkehr, Energie und Kommunikation UVEK. 2017. URL: https://www.astra.admin.ch/astra/de/home/themen/nationalstrassen/verkehrsfluss-stauaufkommen/verkehrsfluss-nationalstrassen.html.

[59] Stettler, Y. and Betbèze, F. Schweizerische Holzenergiestatistik. Erhebung für das Jahr 2016. 2017. URL: http://www.bfe.admin.ch/php/modules/publikationen/stream.php?extlang=de&name=de_57834276.pdf.

[60] Bukowiecki, N., Zieger, P., Weingartner, E., Jurányi, Z., Gysel, M., Neininger, B., Schneider, B., Hueglin, C., Ulrich, A., Wichser, A., Henne, S., Brunner, D., Kaegi, R., Schwikowski, M., Tobler, L., Wienhold, F. G., Engel, I., Buchmann, B., Peter, T., and Baltensperger, U. Ground-based and airborne in-situ measurements of the Eyjafjallajökull volcanic aerosol plume in Switzerland in spring 2010. *Atmospheric Chemistry and Physics* 11.19 (2011), pp. 10011–10030. URL: https://www.atmos-chem-phys.net/11/10011/2011/.

[61] Thorsteinsson, T., Jóhannsson, T., Stohl, A., and Kristiansen, N. I. High levels of particulate matter in Iceland due to direct ash emissions by the Eyjafjallajökull eruption and resuspension of deposited ash. *Journal of Geophysical Research* 117.B9 (2012). URL: http://dx.doi.org/10.1029/2011JB008756.

Chapter 6

Exploring air quality interventions

This work was originally published in *Science of the Total Environment* on October 25, 2018.[†]

6.1 Abstract

Interventions used to improve air quality are often difficult to detect in air quality time series due to the complex nature of the atmosphere. Meteorological normalisation is a technique which controls for meteorology/weather over time in an air quality time series so intervention exploration (and trend analysis) can be assessed in a robust way. A meteorological normalisation technique, based on the random forest machine learning algorithm was applied to routinely collected observations from two locations where known interventions were imposed on transportation activities which were expected to change ambient pollutant concentrations. The application of progressively stringent limits on the content of sulfur in marine fuels was very clearly represented in ambient sulfur dioxide (SO_2) monitoring data in Dover, a port city in the South East of England. When the technique was applied to the oxides of nitrogen (NO_x and NO_2) time series at London

[†]https://doi.org/10.1016/j.scitotenv.2018.10.344

Marylebone Road (a Central London monitoring site located in a complex urban environment), the normalised time series highlighted clear changes in NO_2 and NO_x which were linked to changes in primary (directly emitted) NO_2 emissions at the location. The clear features in the time series were illuminated by the meteorological normalisation procedure and were not observable in the raw concentration data alone. The lack of a need for specialised inputs, and the efficient handling of collinearity and interaction effects makes the technique flexible and suitable for a range of potential applications for air quality intervention exploration.

6.2 Graphical abstract

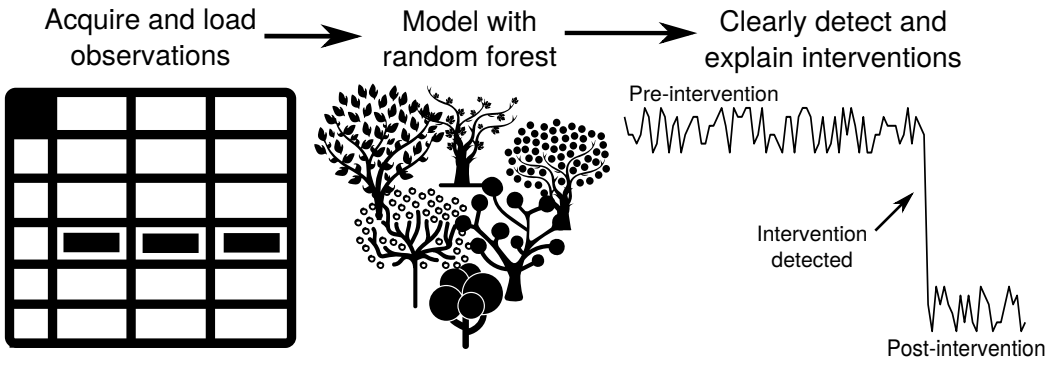

Figure 6.1: Graphical abstract. Table and tree icons are from freepik.com [1].

6.3 Introduction

Across all spatial and temporal scales, weather influences concentrations of atmospheric pollutants and in turn ambient air quality.[2,3] The effects of weather (or meteorology) on air quality are often much greater than intervention or management efforts to control air pollution and therefore intervention events can be very difficult to detect and quantify within an

observational record.[4] Similarly, when considering trends in ambient air pollution, it can be difficult to know whether a change in concentration is due to meteorology or a change in emission source strength. Meteorological variation can therefore frustrate the analysis of trends in different pollutant species. If meteorology is not controlled or accounted for, the changes in pollutant concentrations observed may be contaminated with meteorological variation rather than emission or chemically induced perturbations which can lead to erroneous conclusions concerning the efficacy of air quality management strategies.[5,6] This issue is often acknowledged, but infrequently addressed.

Meteorological normalisation is one technique which can be used to control for meteorology over time in air quality time series. The central philosophy of meteorological normalisation is to reduce variability in an air quality time series with statistical modelling. The reduction of variability is achieved by training a model which can explain some of the variation of pollutant concentrations through a number of independent variables. The independent variables used are typically surface-based meteorological observations and time variables which act as proxies for regular emission patterns such as hour of day and season.[7] However, in practice, any independent variable which could explain variations in pollutant concentrations could be used. Once the model has been trained and it is found that it can explain an adequate amount of the dependent variable's variation, the model can be used to remove the influence the independent variables have on the dependent variable by sampling and predicting. The time series which results can then be exposed to further exploratory data analysis (EDA) techniques such as formal trend analysis and/or intervention exploration.[8] The normalised time series is in the pollutant's original units and can be thought of as concentrations in "average" or invariant weather conditions.

There has been some air quality research conducted which uses the idea of change-point analysis to investigate changes in atmospheric pollutant

concentrations, for example Carslaw et al. [9] and Carslaw and Carslaw [10]. Methods such as these rely on regime changes where a time series abruptly shifts from one regime to another.[11] In the air quality domain, this rarely happens, since changes are usually nuanced and occur progressively with much variability which makes the generality of this approach for investigating intervention efforts poor. Meteorological normalisation is potentially a more general approach which enables its use in a greater range of applications.

Atmospheric processes are complex, non-linear, and observations commonly record collinearity with other observations. These attributes make the process of statistical modelling very challenging, especially so with parametric methods.[12] With the rise of machine learning algorithms, these attributes can be much more easily accommodated due to the non-parametric and robust nature of these techniques.[13] The meteorological normalisation technique used here uses random forest, an ensemble decision tree machine learning method as the modelling algorithm.

Random forest has been described very well and in depth elsewhere (see Grange et al. [8], Friedman et al. [13], Breiman [14], Tong et al. [15], Ziegler and König [16], and Jones and Linder [17]). However in brief, a single decision tree is formed from a series of binary splits which results in homologous or "pure" groups. The splitting process is recursive which means splitting occurs until purity is achieved if the tree is allowed to grow to its maximum depth. Decision trees make no assumptions on the input data structure (they are non-parametric), allow for interaction and collinearity among variables, and will ignore variables which are irrelevant to the dependant variable.[16] Decision trees are fast to train, fast to make predictions, and are conceptually simple to understand. However, they suffer heavily from overfitting, an issue where the model represents the training set well, but does not generalise to sets which were not used for training.[17] Using a model which predicts pollutant concentrations and suffers from overfitting would result in the model being contaminated with noise from the training set and

unreliable predictions would impede analyses.

Random forest is an algorithm which controls for the tendency of decision trees to overfit. The algorithm achieves this by sampling (with replacement) the training set with a process called bagging (bootstrap aggregation).[18] In modern usage, sampling of the independent variables is usually done during bagging too. Bagging results in a new, sampled set called out-of-bag (OOB) data. A decision tree is then grown on the OOB data. The bagging-then-tree growth is repeated, generally a few hundred times. Because OOB data is sampled, all the decision trees are grown on differing observations and independent variables which leads to a "forest" of decorrelated trees. After training, all the individual trees within the forest are used to predict, but their predictions are aggregated as a mean (or the mode for categorical dependent variables) and that forms the single ensemble prediction for the model.

The meteorological normalisation technique is pragmatic in respect to the input variables required for many common applications. Generally, routinely accessible surface meteorological variables are very effective for the process and specialised or obscure variables are generally not necessary for the technique to be applied. Although traffic counts, upper air data, and outputs from weather models will usually strengthen a model's explanatory power, the existence or access to such variables is not a prerequisite, an attribute which is very useful for most situations where such inputs are not available. For pollutants which are primarily controlled by regional scale processes, most notably particulate matter (PM) and ozone (O_3), additional variables such as boundary layer height, air mass cluster, or back trajectory information would however be beneficial to include if possible and examples can be found elsewhere, for example Grange et al. [8].

The temporal variables used as independent variables in the meteorological normalisation models: Julian day, weekday, and hour of day are included not for their direct influence on atmospheric concentrations, but because they act as proxies for cyclical emission patterns. Hour of day for ex-

ample offers a term to explain emissions with a diurnal cycle such as traffic-related rush hour emissions or domestic heating phases, while Julian day is a seasonal term which represents emissions or atmospheric chemistry which varies seasonally. These processes are generally strong drivers of concentrations of most atmospheric pollutants.[19] Random forest's ability to handle collinearity and interaction between these and the other independent variables used and the lack of need of specialised or exotic inputs results in a flexible tool kit for probing the influences of interventions on air quality time series.

6.4 Objectives

The primary objective of this paper is to apply a meteorological normalisation technique based on random forest, a machine learning algorithm to detect interventions in air quality monitoring data. This is done to gain understanding of what physical and chemical processes are driving ambient pollutant concentrations and highlight the suitability and potential of the technique to other applications.

Two case studies are presented using routine data sets in Dover, South East England where sulfur fuel limits of ships were imposed and changes in ambient sulfur dioxide (SO_2) concentrations are expected and in Central London where congestion charging and local bus fleet management has perturbed oxides of nitrogen (NO_x) emission sources. The changes in concentrations and emissions are then explained in respect to implementation of policy which would be difficult to detect with other EDA techniques where no meteorological normalisation is performed.

6.5 Methods

6.5.1 Data

6.5.1.1 Port of Dover SO_2

Hourly SO_2 concentrations were analysed from the Port of Dover, a major port located in Kent in the South East of England. Two air quality monitoring sites, Dover Docks and Dover Langdon Cliff's SO_2 data were queried from the Kent Air Quality database.[20] A nearby meteorological site, Langdon Bay located to the west of the port was used to provide surface meteorological observations and were accessed from NOAA's Integrated Surface Database (ISD) (Figure 6.2(a)).[21] The monitoring sites had different commissioning and decommissioning dates and neither site is still operating (Table 6.1). SO_2 observations are available between March 2001 and December 2012. The data capture rates for SO_2 at Dover Langdon Cliff and Dover Docks for their online period were 92 and 82 % respectively. These monitoring sites are of interest because marine fuels in British and European waters have been subject to a series of sulfur content fuel limits. The introduction and continued enforcement of these sulfur fuel limits were expected to influence ambient SO_2 concentrations. The details of these interventions are discussed further in Section 6.6.1.2.

Table 6.1: Details of the air quality monitoring sites in Dover and London used in this analysis.

Location	Site name	Site type	Latitude	Longitude	Elevation	Date start
Dover	Langdon Bay	Meteorological	51.133	1.350	117	1973-03-08
Dover	Dover Langdon Cliff	Urban background	51.132	1.339	98	2001-03-17
Dover	Dover Docks	Urban industrial	51.127	1.336	6	2006-11-17
London	London Heathrow	Meteorological	51.478	-0.461	25	1948-12-01
London	London Marylebone Road	Traffic	51.523	-0.155	35	1997-01-01

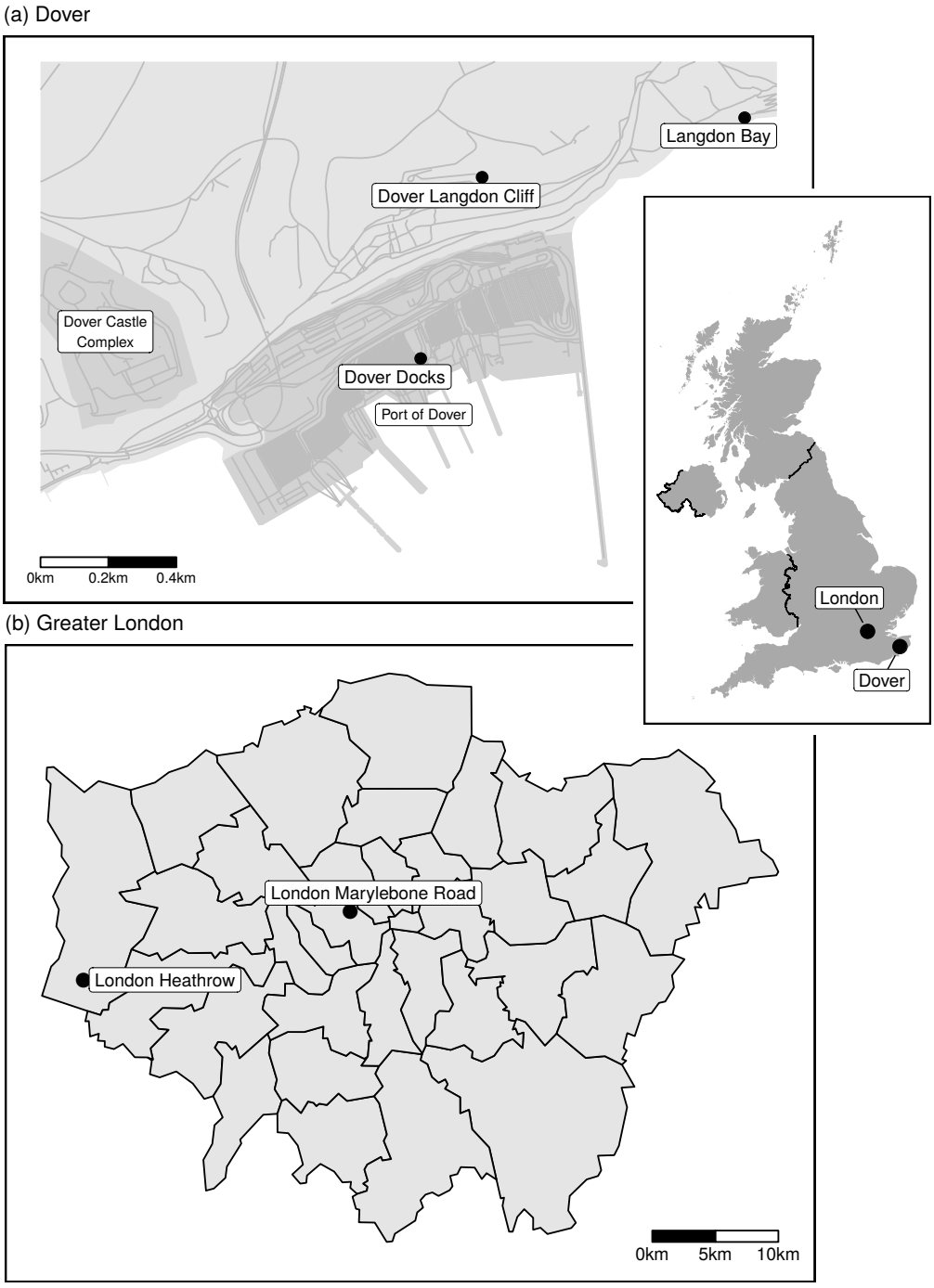

Figure 6.2: Maps of the study sites with a United Kingdom insert for country-scale context. The Port of Dover complex is displayed in (a) and the internal lines indicate roads and Greater London is shown in (b), with the London Boroughs and City of London indicated with internal polygons.

6.5.1.2 London Marylebone Road NO_2 and NO_x

Hourly NO_2 and NO_x data from London's Marylebone Road air quality monitoring site were accessed from **smonitor** Europe, a European database containing the observations and metadata from the AirBase and Air Quality e-Reporting (AQER) repositories.[22,23] NO_x concentrations have been monitored since July 1997 and the final year of reporting sourced from the European data repositories used was 2016. Data capture rates for NO_x and NO_2 for the analysis period were 97 %. London Heathrow, a large airport located at the far west of Greater London was used for surface meteorological observations sourced from NOAA's ISD (Figure 6.2(b)). London Marylebone Road is situated in a complicated central urban environment. The site is located one metre south of the kerb on the A501 trunk road and sits within an irregularly shaped street canyon. London Marylebone Road is a prominent and often analysed site due to its long observational record and diverse suite of pollutants which are monitored at the site.[24]

NO_x and NO_2 concentrations across European cities are a significant issue and many member states are non-compliant to the legal European ambient air quality limits.[25,26] Almost all locations which are non-compliant are classified as roadside (or "traffic-influenced").[27] London has some of the highest roadside concentrations of NO_x and NO_2 in Europe and London Marylebone Road (Figure 6.2(b)) is an often referenced monitoring site for its high concentrations.

To combat the issue of traffic congestion, Greater London authorities imposed the Congestion Charge Zone (CCZ), which was first enforced in February 2003.[28] Since that time, the London Low Emission Zone (LEZ), and the Emissions Surcharge (better known as the T-Charge) have also been implemented to combat air pollution.[29] The details and start dates of these various measures are displayed in Table 6.2. All these interventions are significant investments with large amounts of planning and resources to execute and maintain.

Table 6.2: Details of interventions within Greater London to counter traffic congestion.

Name	Abbreviation	Start date	Area covered	Operation
Congestion Charge Zone	CCZ	2003-02-17	Central London	07:00–18:00 Mo-Fr
London Low Emission Zone (first phase)	LEZ	2008-02-04	Greater London	24/7
London Low Emission Zone (second phase)	LEZ	2012-01-03	Greater London	24/7
Emissions Surcharge	T-Charge	2017-10-23	Central London	07:00–18:00 Mo-Fr
Ultra Low Emission Zone (planned)	ULEZ	2019-04-08	Central London	24/7

6.5.2 Modelling and the hyperparameters

For both examples, the meteorological normalisation procedure was conducted in the same way and the **rmweather** R package (version 0.1.2) was used for this process.[30,31] The number of trees for the random forest models was fixed at 300, the minimal node size was five, and the number of variables split at each node was the default for regression mode: the rounded down square root of the number of independent variables which in these examples was three (**rmweather**'s function arguments `n_trees`, `min_node_size`, and `mtry` respectively). The independent variables used were: Unix date (number of seconds since 1970-01-01) as the trend term, Julian day as the seasonal term, weekday, hour of day, air temperature, relative humidity, wind direction, wind speed, and atmospheric pressure. Training was only conducted on observations which had non-missing wind speed and the pollutant being modelled. Three hundred predictions were used to calculate the meteorologically normalised trend. The normalised trends were aggregated to monthly resolution for presentation in Section 6.6. A conceptual representation of the meteorological normalisation processes is displayed in Figure 6.3.

For the Dover SO_2 examples, models were calculated using the full observational set, but after investigating the models (discussed in Section 6.6.1.1), the observations were filtered to wind directions which were sourced from the port and these models are the ones which were used for the time series analysis (Section 6.6.1.2). For observations at London Marylebone Road, no filtering was undertaken. In the case of London Marylebone Road, there are

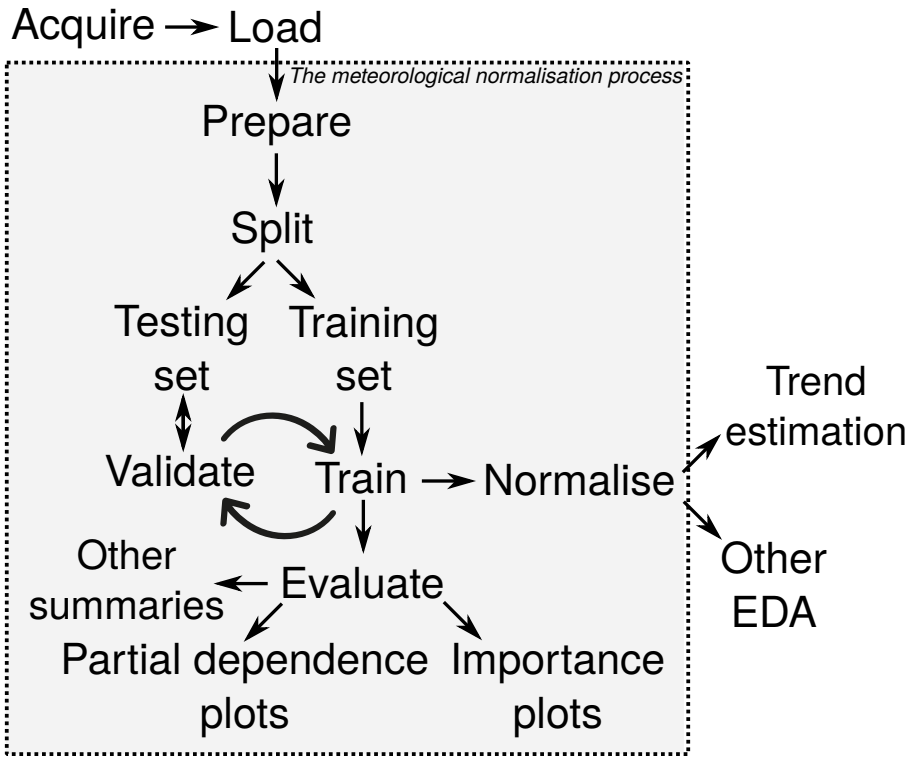

Figure 6.3: The framework for the meteorological normalisation technique. The training and validation phase is iterative to ensure the model does not overfit and adequate performance is achieved. After the technique has been completed, other analyses are conducted on the normalised time series.

a large number of potential events which could influence pollutant concentrations and emissions. To objectively identify events, the meteorologically normalised time series were tested for breakpoints or changes in structure. The structural change algorithm is described in Zeileis et al. [32, 33] and was implemented with the **strucchange** R package.

The random forest algorithm does not directly offer the ability to determine error or uncertainty of estimates. However, uncertainty is important to consider in many situations. To enable uncertainty to be evaluated for the case studies, 50 random forest models were grown for each example with the hyperparameters described above, but with randomly sampled (bootstrapped) input sets. The bootstrapping of the observational data ensured the models were grown on different training sets. The importance values (a measure of the variables' strength or influence on prediction), partial dependencies, and predictions for each of the 50 models were then summarised. The importance measure used was variable permutation difference which is not subjected to a scaling procedure.[34] The summaries used from the "ensemble of the ensembles" were the mean, and the 2.5 % and 97.5 % quantiles of the 50 estimates *i.e.* a range that spans the 95 % confidence interval in the mean. The model performance statistics for the four sets of models are displayed in Table 6.3.

Table 6.3: Mean random forest model performance statistics four the four sets of models grown for the analysis.

Location	Model	n	R^2
Dover	Dover Docks SO_2	34224	0.67
Dover	Dover Langdon Cliff SO_2	53535	0.63
London	London Marylebone Road NO_2	131677	0.82
London	London Marylebone Road NO_x	131677	0.83

6.6 Results and discussion

6.6.1 Port of Dover SO_2

6.6.1.1 Models

The random forest models grown for SO_2 at the two Dover sites had R^2 values of 63 and 67 % (Table 6.3), therefore, the models had moderate explanatory ability for Dover's SO_2 concentrations. However, it should be noted that predicting concentrations over such short time periods with intermittent source strength is challenging and data capture was less than ideal for these monitoring sites. The moderate performance can be explained by SO_2 at this location containing large amounts of variation due to ship movements and if winds were in a favourable direction to transport emissions from the port complex to the monitoring sites (southerlies). Indeed, wind direction was the most important variable for SO_2 explanation for the random forest models (Figure 6.4).

Partial dependence plots of decision tree models allow the learning process to be interpreted and a data user to examine how variables are being handled in the predictive model. Figure 6.5 demonstrates a two-way partial dependence plot for SO_2 concentrations at Dover Landon Cliff using wind direction and date (the trend term) as the independent variables. The feature which is most clear is the band of increased SO_2 dependence between 150 and 210 degrees. Outside of this band of southerly winds, there were low levels of dependence on SO_2 concentrations. The Dover Landon Cliff monitoring site was located north of the Port of Dover docks and very slightly to the east (Figure 6.2(a)). The partial dependence on wind direction is consistent with this location and indicates that wind direction was handled sensibly in the random forest model. This observation can be confirmed further with a bivariate polar plot of mean SO_2 concentrations by wind direction and speed at the monitoring site (Figure 6.6). The first sulfur content fuel change in mid-August 2006 can also be seen in the two-way

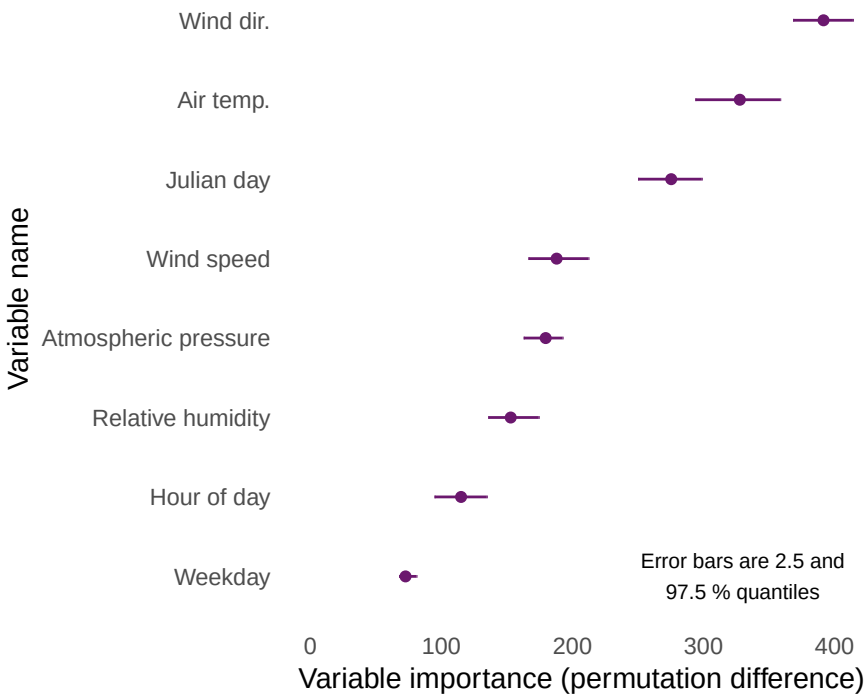

Figure 6.4: Variable importance plot for SO_2 at Dover Langdon Cliff between 2001 and 2010 calculated by 50 random forest models.

partial dependence plot as a clear reduction in SO_2 dependence when winds were sourced from the port (the south; discussed further in Section 6.6.1.2; Figure 6.5).

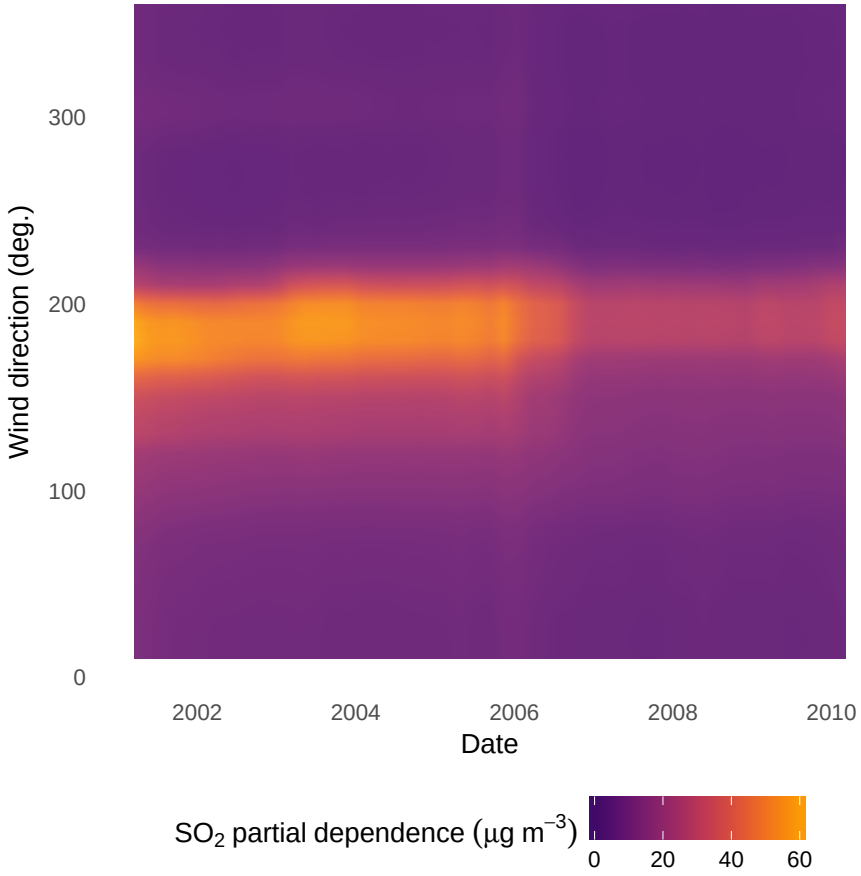

Figure 6.5: Partial dependence of wind direction and date on SO_2 concentrations at Dover Landon Cliff between 2001 and 2010. The Dover Landon Cliff monitoring site was located north of the Port of Dover (Figure 6.2(a)).

Another clear feature isolated by the partial dependence plots was that SO_2 concentrations increased with increasing air temperature at the Dover monitoring sites (Figure 6.7). This relationship was an unexpected outcome because generally, pollutant concentrations are inversely related to air temperature because emissions are more efficiently diluted during warmer periods owing to increased thermal turbulence. For some sources such as heating, emissions are greater at lower temperatures, but when considering

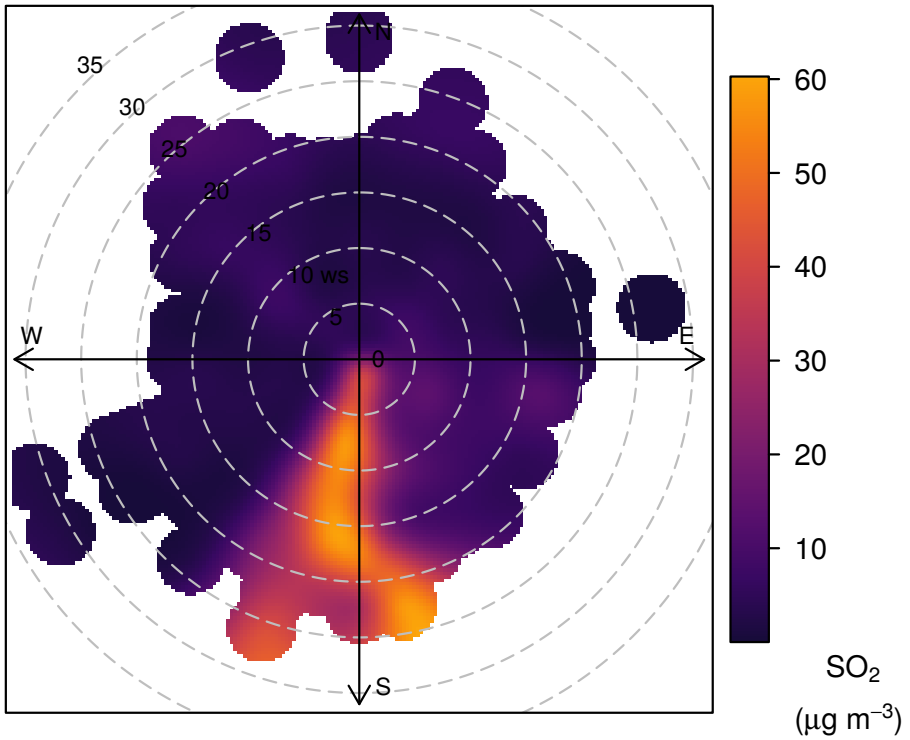

Figure 6.6: Bivariate polar plot of mean hourly SO_2 concentrations at Dover Landon Cliff between 2001 and 2010. The Dover Landon Cliff monitoring site was located north of the Port of Dover (for a location map, see Figure 6.2(a)).

shipping emissions, this would be negligible. At Dover, the SO_2 relationship between concentrations and air temperatures was indicative of convective thermal mixing being an important physical process which resulted in SO_2 emitted by ships to be mixed towards the measurement site at the cliff top. This turbulent mixing at high temperatures resulted in high SO_2 concentrations at the surface and this feature cannot be easily observed in the hourly observational data. The illumination of such physical processes is a major advantage of the random forest algorithm compared to other machine learning methods such as support vector machines (SVM) or artificial neural networks (ANNs) because they do not offer the same amount of model legibility.

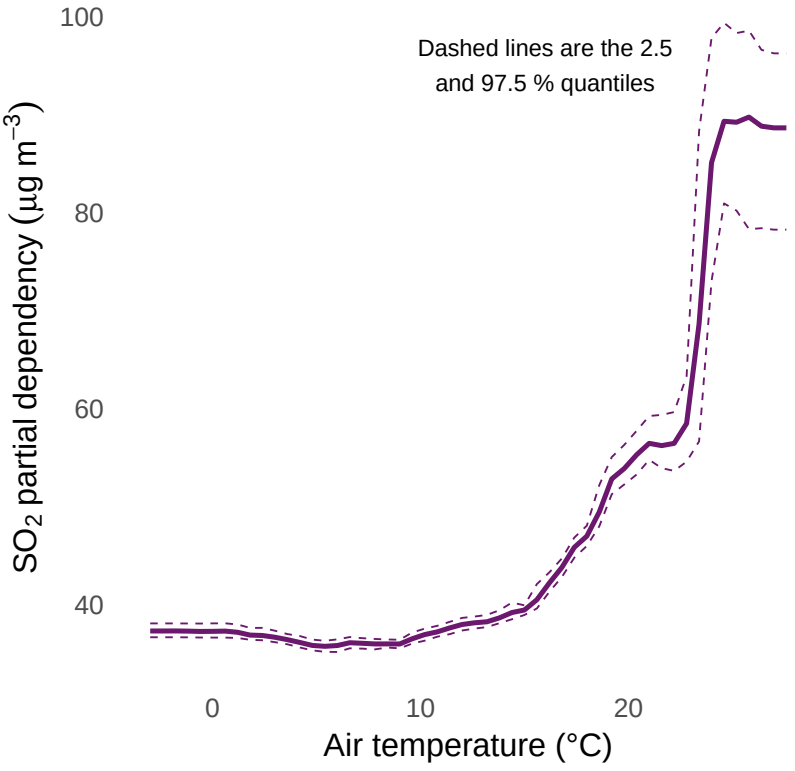

Figure 6.7: Partial dependence of SO_2 on air temperature at Dover Landon Cliff between 2001 and 2010 calculated by 50 random forest models.

6.6.1.2 Influence of sulfur fuel limits on SO_2 concentrations

Since the early 2000s, there has been a number of increasingly stringent sulfur based fuel limits imposed on ships operating in British and European Union (EU) waters due to their status as Sulfur Emission Control Areas (SECAs) or Emission Control Areas (ECAs). The most important events for sulfur control were implemented on August 11, 2006 and January 1, 2010. In August 2006, the MARPOL Annex IV regulations were applied which introduced a 1.5 % sulfur limit on fuel oils used by vessels moving between EU ports.[35] The pre-August 2006 sulfur content for British vessels has been estimated at 2.7 % which represents a reduction in sulfur content of 44 %.[36] At the start of 2010 an additional limit was imposed for all vessels at berth where such vessels were required to be operated with maximum fuel sulfur content of 1 %. These changes should be evident in the SO_2 time series of the nearby ambient monitoring sites. However, if a time series is plotted, the influence of these changes are subtle and not clear due to the high amounts of variation within SO_2 concentrations (Figure 6.8).

The meteorologically normalised SO_2 time series for the Dover sites are displayed in Figure 6.9, after the observations were filtered to wind directions which came for the port, hence the tight 95 % confidence intervals. The dates when changes in sulfur fuel content were implemented are displayed as vertical lines in Figure 6.9 and the influence of sulfur fuel changes are clear (compared with Figure 6.8).

At Dover Langdon Cliff, the monitoring site which was online during the MARPOL 1.5 % fuel sulfur limit transition during August 2001 shows the shift in ambient SO_2 very clearly (Figure 6.9). The mean meteorologically normalised SO_2 concentrations for the pre- and post-fuel change periods were 48 and 26 $\mu g\,m^{-3}$ respectively. This difference represented in percentage change is 45 % and the corresponding estimated change in sulfur fuel content was 44 %. This extremely good agreement between sulfur content fuel changes and normalised ambient SO_2 concentrations suggests that the Port of Dover activities and ship movements remained constant during the

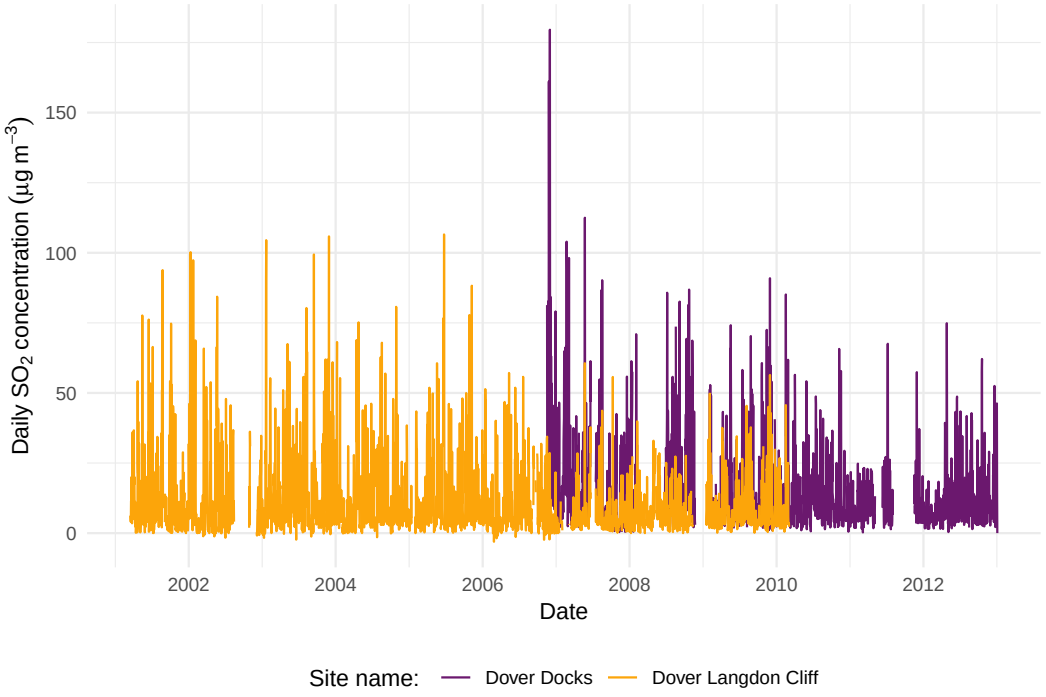

Figure 6.8: Daily SO_2 concentrations at two monitoring sites in Dover between 2001 and 2012.

transition phase and the source of SO_2 at this location was almost exclusively from the port.

The second sulfur fuel content change was implemented on January 1, 2010 and this intervention is also clearly displayed in the meteorologically normalised SO_2 concentrations of the Dover Docks monitoring site (Figure 6.9). The percentage change in fuel sulfur content was 33 % and the percentage change in ambient SO_2 concentrations was 32 %. Like the previous intervention, these two percentage changes match almost exactly, which is somewhat surprising because the intervention was applied only to berthed vessels which would only make up a component of the Port of Dover activities.

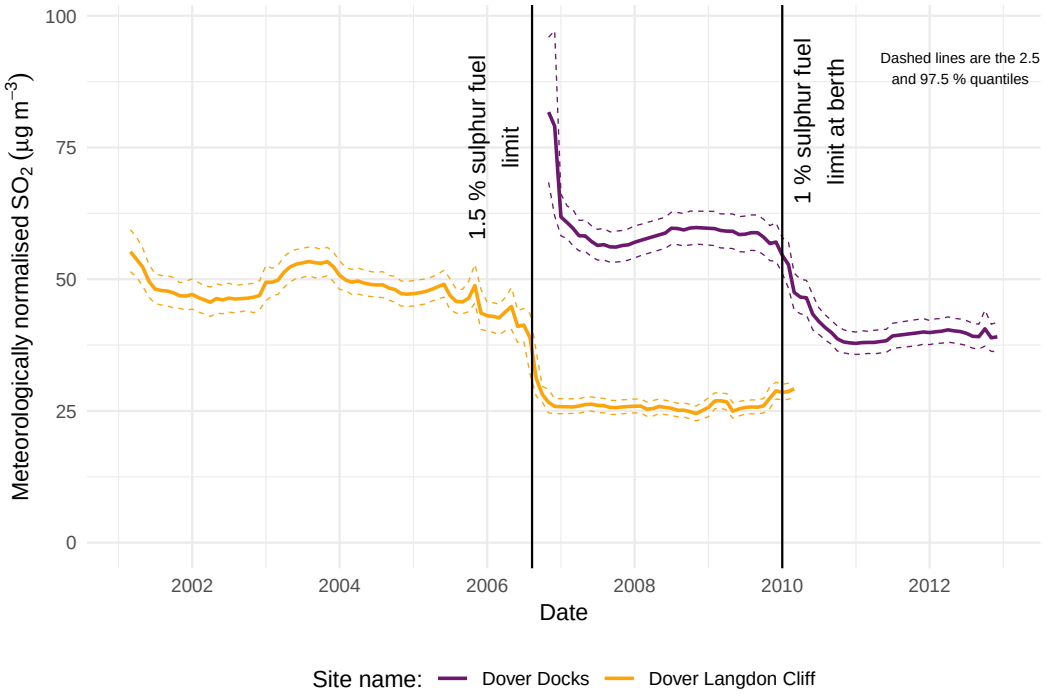

Figure 6.9: Meteorologically normalised SO_2 concentrations at two monitoring sites in Dover between 2001 and 2012 as calculated by 50 random forest models. The vertical lines show the start dates of when changes in marine sulfur fuel content were implemented.

6.6.2 London Marylebone Road NO_x

6.6.2.1 Models

The random forest models of NO_x and NO_2 at London Marylebone Road performed well and had R^2 values of 82 and 83 % respectively (Table 6.3). This good performance can be explained by hour of day being a very good predictor for traffic flows and therefore traffic-sourced pollutants (Figure 6.10). The performance of the random forest models would be rather difficult to achieve with dispersion or deterministic models in such a complicated location. For example, the dispersion models evaluated in Carslaw et al. [37] struggled to represent the street canyon environment, even when traffic information was taken into account. The importance plots for the London Marylebone Road models also show that wind direction is the most important variable to predict NO_2 and NO_x concentrations. London Marylebone Road is located in a street canyon and is subjected to complex flows, including ventilation, vortices, and leeward accumulation of pollutants, (primarily) dependent on wind direction.[10,38] This complexity is demonstrated in the importance of wind direction in explaining NO_x and NO_2 concentrations (Figure 6.10) and this has been noted before at this location.[39,40]

6.6.2.2 Changes in primary NO_2

Using the predictive models for meteorological normalisation results in very clear and almost noiseless meteorologically normalised trends shown in Figure 6.11. It is immediately clear that NO_x and NO_2 are not behaving the same way at this monitoring location. This is because of changes in vehicular primary (directly emitted) NO_2 during the analysis period (1997–2016).[26,41,42] The vertical lines on Figure 6.11 show the breakpoints identified by structural change analysis after the meteorological normalisation procedure.

NO_x concentrations decreased after the introduction of a bus lane adjacent to the monitoring site in 2001 but have remained near constant since

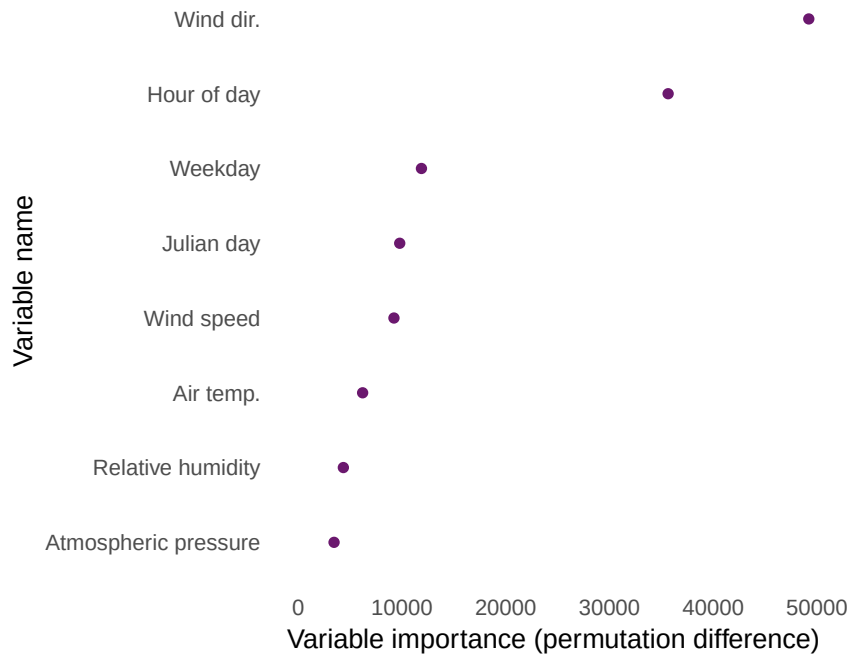

Figure 6.10: Variable importance plot for 50 NO_2 random forest models for London Marylebone Road. The uncertainty among the importances of the 50 models was very small and therefore the quantiles are not shown. The importances for the NO_x models were very similar.

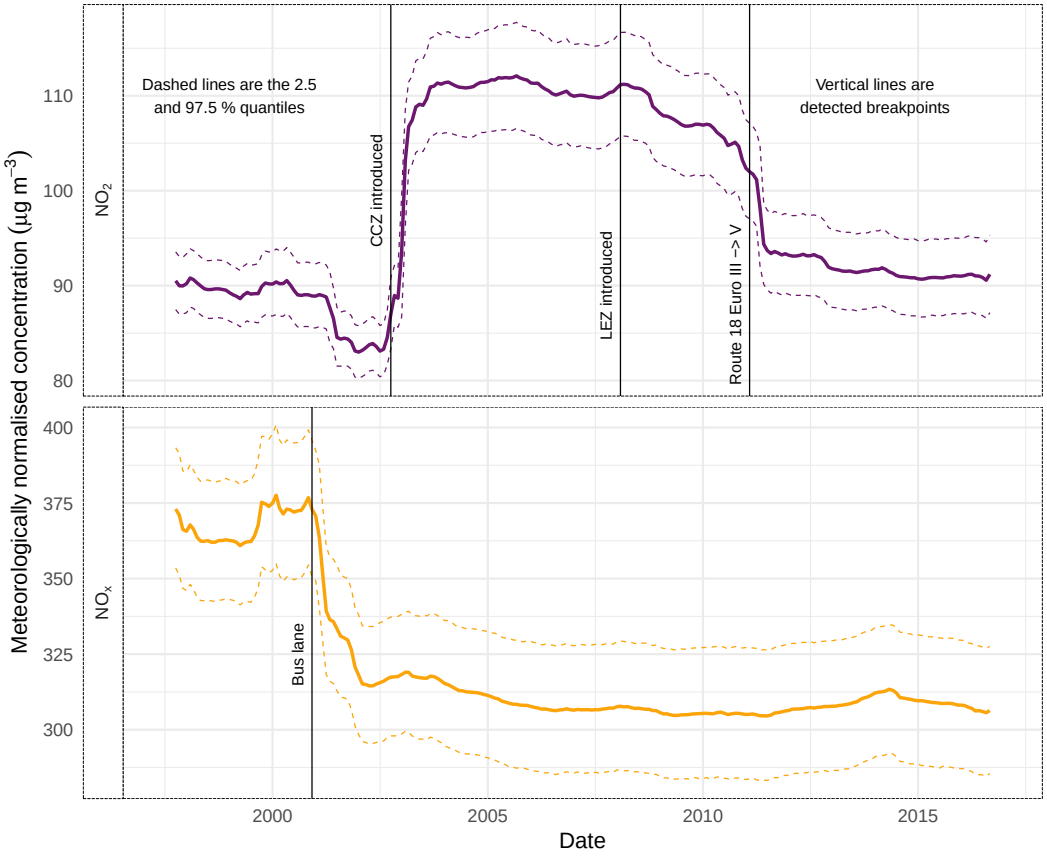

Figure 6.11: Meteorologically normalised NO_x and NO_2 at London Marylebone Road between 1997 and 2016 as calculated by 50 random forest models (for each pollutant). The vertical lines on show the breakpoints identified by structural change analysis.

the introduction of the CCZ in February 2003 (Figure 6.11 and Table 6.2). Despite the progressively stringent vehicular emission controls being applied across Europe between 2003 and 2016 (the last year of data in analysis), they have had little effect to NO_x at London Marylebone Road. This observation could be, at least partly, explained by the disconnect between laboratory testing and real-world emissions of NO_x which become a public issue after the diesel emission scandal in September 2015.[43,44] However, heavy duty vehicles are also very important to consider alongside passenger vehicles at this Central London location.[45,46]

NO_2 concentrations at London Marylebone Road have increased since 1997 and were at their maximum between 2002 and 2008 (Figure 6.11). The changes observed can be explained by changes to the vehicle fleet using the adjacent A501 road resulting from the introduction of congestion charging, London's Low Emission Zone, and evolution of the local bus fleet. The rapid increase of NO_2 concentrations was observed in the meteorologically normalised time series between July 2002 and July 2003 (Figure 6.11). The CCZ was introduced in mid-February 2002; right in the middle of the period of increasing NO_2 and within six months of the suggested breakpoint (October 2012). The increase in NO_2 concentrations was due to increased primary NO_2 because no change in the meteorologically normalised NO_x was observed at the same time.

The implementation of the CCZ was accompanied with a retrofitting programme of Euro III local buses with continuously regenerating diesel particulate filters (CRDPF, also known by their commercial name: CRT filters). CRDPF are passive devices and have two components: an upstream oxidation catalyst and a particulate matter (PM) filter. The oxidation catalyst oxidises NO within the exhaust stream to NO_2 and this NO_2 is then used as a PM oxidant in the filter-proper. The observations show that these retrofitted passive devices were not optimised because much of the generated NO_2 was not reduced within the PM filter and was therefore emitted into the roadside atmosphere and thus significantly increased ambient NO_2

concentrations (Figure 6.11).

NO_2 concentrations remained approximately stable until February 2008 when London's Low Emission Zone (LEZ) was introduced and NO_2 concentrations began to decrease (Figure 6.11). The second NO_2 breakpoint was detected for February 2008 giving some evidence that the LEZ reduced NO_2 concentrations at London Marylebone Road (although no corresponding change in NO_x was observed). However, during this period the local bus fleets were also being progressively replaced with newer buses compliant to the later Euro IV, V, and VI heavy vehicle emission standards (Finn Coyle, Tom Cunnington, and Gabrielle Bowden (Transport for London), personal communication, March 2018) as well of natural vehicle turnover removing older and more polluting vehicles from the in-service fleet. The third NO_2 breakpoint identified coincided with route 18, the bus route with the highest peak vehicle requirements (PVR), shifting from Euro III to Euro V vehicles in late 2010 (Figure 6.11). After 2011, NO_2 concentrations continued to decline with the introduction of Euro VI and hybrid buses servicing the 453, 27, and 205 routes. By the end of 2016, NO_2 had declined to almost pre-CCZ concentrations. The features displayed in the normalised time series were not clear in the raw concentration data (displayed in Figure 6.12) and the breakpoints identified were unable to be resolved without the meteorological normalisation technique.

The tandem use of the meteorological normalisation procedure and breakpoint analysis is powerful and can reveal many changes, but in many cases there may not be sufficient information or metadata to help explain the changes observed. In this Central London example, many of the factors driving pollutant concentrations are known due to the site's prominence.

London Marylebone Road also monitors ozone (O_3), something which is rare for roadside monitoring locations in Europe. The NO_2, NO_x, and O_3 complement allows for the estimation of primary NO_2 with an independent method by determining the total oxidant (OX; NO_2 + O_3) within NO_x.[47,48] Figure 6.13 shows monthly estimates of the primary NO_2 fraction at London

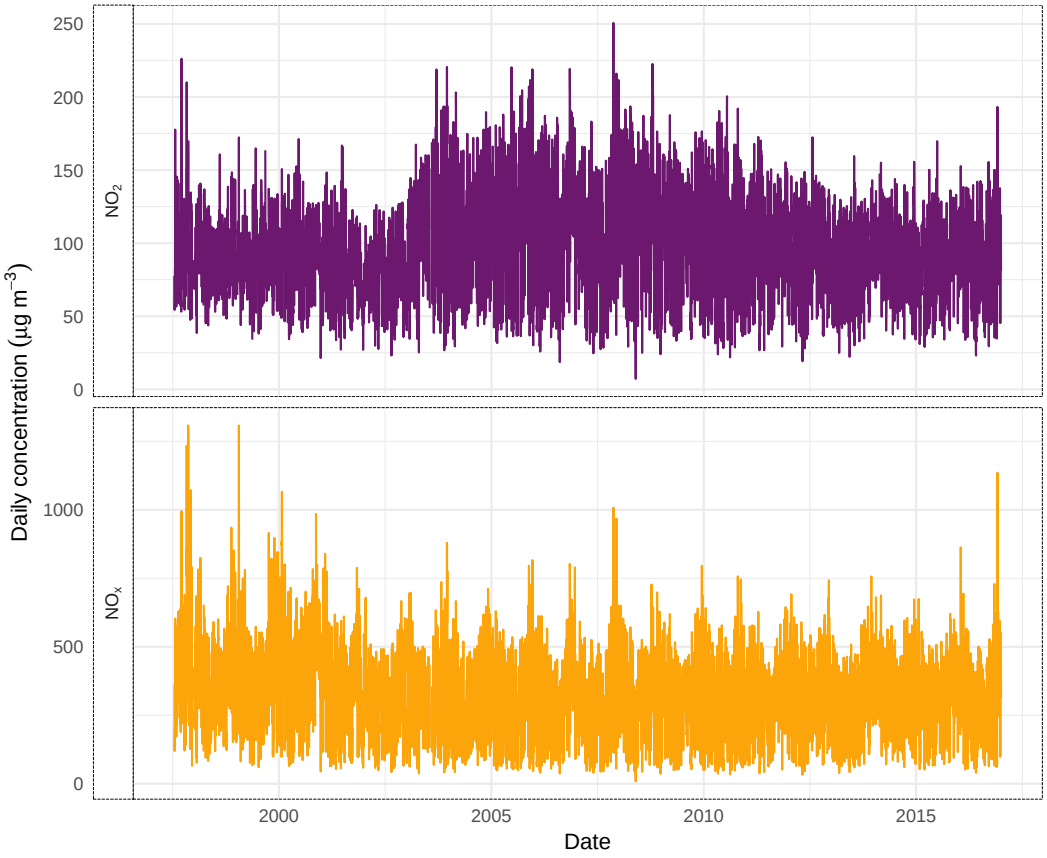

Figure 6.12: Daily NO_2 and NO_x concentrations at London Marylebone Road between 1997 and 2016.

Marylebone Road with robust linear regression. Figure 6.13 is consistent with Figure 6.11 with a rapid increase in primary NO_2 during 2002 and a reduction, but at a slower rate after 2008 thus further confirming and validating that the trends observed in Figure 6.11 are driven by changes in primary NO_2 emissions. The reason why the trend is similar in Figure 6.13 and Figure 6.11 is that at this particular site increased emissions of primary NO_2 were sufficient to have a measurable effect on ambient concentrations.

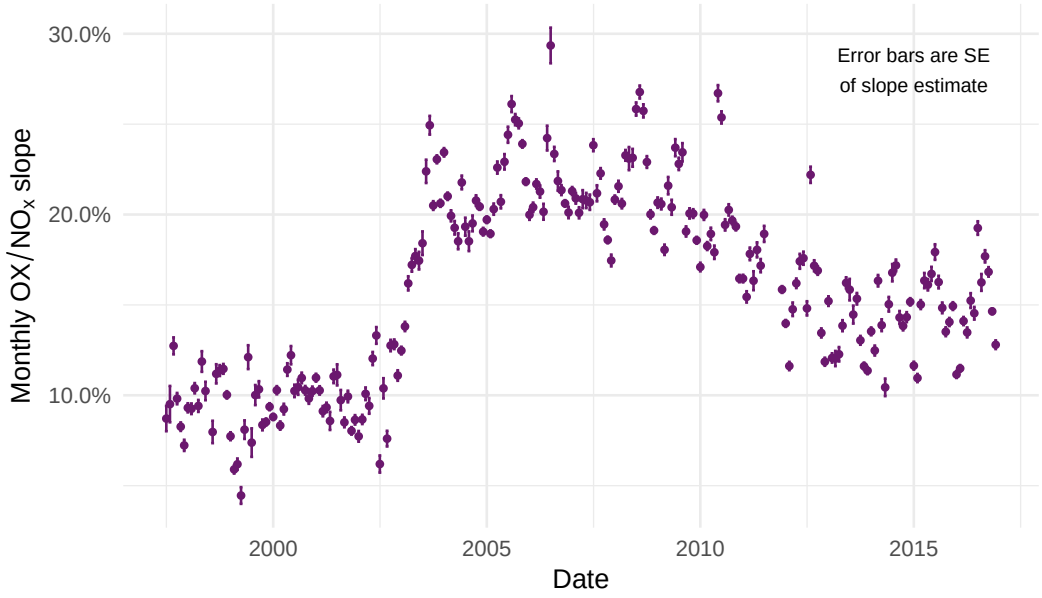

Figure 6.13: Monthly total oxidant (OX; NO_2 + O_3)/NO_x slope at London Marylebone Road between 1997 and 2016. Slope and errors were calculated with robust linear regression.

6.7 Conclusions

Controlling for changes of meteorology is an important component to consider when conducting air quality data analysis over time. A meteorological normalisation technique using random forest was used to investigate interventions in routine air quality monitoring data from two areas. The interventions applied to marine fuel content changes were explored in Dover, a port city in the South East of England and the interventions

were represented in the meteorologically normalised time series almost exactly. The non-black box nature of the random forest models was used to investigate the dependence of pollutant concentrations on meteorological variables such as air temperature and wind direction which highlighted the benefit of the technique where physical and chemical atmospheric processes can be illuminated, understood, and explained.

In the example of the implementation of congestion charging in Central London, very clear changes in primary NO_2 emissions were displayed in the meteorologically normalised time series. The performance of these roadside models was high due to the models' ability to use wind direction and hour of day very effectively, something which dispersion or deterministic models struggle with when used for modelling street canyon environments. The case studies presented are both examples where there is significant ability to cross check the observed features with available information on changes in the sites' local environments to validate the outputs.

The meteorological normalisation technique is very relevant for exploring the influence of interventions or management activities on local air quality. The combination of a non-parametric method, the lack of need for specialised measurements, and the effective use of proxy variables lends the technique to a wide range of air quality data analysis applications.

6.8 References

[1] freepik.com. *FlatIcon*. 2017. URL: www.freepik.com.

[2] Stull, R. B. *An Introduction to Boundary Layer Meteorology*. London: Kluwer Academic Publishers, 1988, 666pp.

[3] Monks, P., Granier, C., Fuzzi, S., Stohl, A., Williams, M., Akimoto, H., Amann, M., Baklanov, A., Baltensperger, U., Bey, I., Blake, N., Blake, R., Carslaw, K., Cooper, O., Dentener, F., Fowler, D., Fragkou, E., Frost, G., Generoso, S., Ginoux, P., Grewe, V., Guenther, A., Hansson, H., Henne, S., Hjorth, J., Hofzumahaus, A., Huntrieser, H., Isaksen, I., Jenkin, M., Kaiser, J., Kanakidou, M., Klimont, Z., Kulmala, M., Laj, P., Lawrence, M., Lee, J., Liousse, C., Maione, M., McFiggans, G., Metzger, A., Mieville, A., Moussiopoulos, N., Orlando, J., O'Dowd, C., Palmer, P., Parrish, D., Petzold, A., Platt, U., Pschl, U., Prvt, A., Reeves, C., Reimann, S., Rudich, Y., Sellegri, K., Steinbrecher, R., Simpson, D., Brink, H. ten, Theloke, J., Werf, G. van der, Vautard, R., Vestreng, V., Vlachokostas, C., and Glasow, R. von. Atmospheric composition change - global and regional air quality. *Atmospheric Environment* 43.33 (2009), pp. 5268–5350. DOI: 10.1016/j.atmosenv.2009.08.021. URL: http://www.sciencedirect.com/science/article/B6VH3-4X3N46N-1/2/1db0fa3c5afafc9418ab227802a71755.

[4] Anh, V., Duc, H., and Azzi, M. Modeling anthropogenic trends in air quality data. *Journal of the Air & Waste Management Association* 47.1 (1997), pp. 66–71. DOI: doi:10.1080/10473289.1997.10464406. URL: https://doi.org/10.1080/10473289.1997.10464406.

[5] Libiseller, C., Grimvall, A., Waldén, J., and Saari, H. Meteorological normalisation and non-parametric smoothing for quality assessment and trend analysis of tropospheric ozone data. *Environmental Monitoring and Assessment* 100.1 (2005), pp. 33–52. DOI: 10.1007/s10661-005-7059-2. URL: http://dx.doi.org/10.1007/s10661-005-7059-2.

[6] Wise, E. K. and Comrie, A. C. Extending the Kolmogorov–Zurbenko Filter: Application to Ozone, Particulate Matter, and Meteorological Trends. *Journal of the Air & Waste Management Association* 55.8 (2005), pp. 1208–1216.

DOI: 10.1080/10473289.2005.10464718. URL: http://dx.doi.org/10.1080/10473289.2005.10464718.

[7] Derwent, R., Middleton, D., Field, R., Goldstone, M., Lester, J., and Perry, R. Analysis and interpretation of air quality data from an urban roadside location in Central London over the period from July 1991 to July 1992. *Atmospheric Environment* 29.8 (1995), pp. 923–946. DOI: 10.1016/1352-2310(94)00219-B. URL: http://www.sciencedirect.com/science/article/pii/135223109400219B.

[8] Grange, S. K., Carslaw, D. C., Lewis, A. C., Boleti, E., and Hueglin, C. Random forest meteorological normalisation models for Swiss PM_{10} trend analysis. *Atmospheric Chemistry and Physics* 18.9 (2018), pp. 6223–6239. DOI: https://doi.org/10.5194/acp-18-6223-2018. URL: https://www.atmos-chem-phys.net/18/6223/2018/.

[9] Carslaw, D. C., Ropkins, K., and Bell, M. C. Change-Point Detection of Gaseous and Particulate Traffic-Related Pollutants at a Roadside Location. *Environmental Science & Technology* 40.22 (2006), pp. 6912–6918. DOI: 10.1021/es060543u. URL: http://dx.doi.org/10.1021/es060543u.

[10] Carslaw, D. C. and Carslaw, N. Detecting and characterising small changes in urban nitrogen dioxide concentrations. *Atmospheric Environment* 41.22 (2007), pp. 4723–4733. DOI: 10.1016/j.atmosenv.2007.03.034. URL: http://www.sciencedirect.com/science/article/pii/S1352231007002919.

[11] Lyubchich, V., Gel, Y. R., and El-Shaarawi, A. On detecting non-monotonic trends in environmental time series: a fusion of local regression and bootstrap. *Environmetrics* 24.4 (2013), pp. 209–226. URL: http://dx.doi.org/10.1002/env.2212.

[12] Barmpadimos, I., Hueglin, C., Keller, J., Henne, S., and Prévôt, A. S. H. Influence of meteorology on PM_{10} trends and variability in Switzerland from 1991 to 2008. *Atmospheric Chemistry and Physics* 11.4 (2011), pp. 1813–1835. URL: http://www.atmos-chem-phys.net/11/1813/2011/.

[13] Friedman, J., Hastie, T., and Tibshirani, R. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Second. Vol. 1. Springer series in statistics Springer, Berlin, 2001.

[14] Breiman, L. Random Forests. *Machine Learning* 45.1 (2001), pp. 5–32. DOI: 10.1023/A:1010933404324. URL: http://dx.doi.org/10.1023/A:1010933404324.

[15] Tong, W., Hong, H., Fang, H., Xie, Q., and Perkins, R. Decision Forest: Combining the Predictions of Multiple Independent Decision Tree Models. *Journal of Chemical Information and Computer Sciences* 43.2 (2003), pp. 525–531. DOI: 10.1021/ci020058s. URL: http://dx.doi.org/10.1021/ci020058s.

[16] Ziegler, A. and König, I. R. Mining data with random forests: current options for real–world applications. *WIREs Data Mining and Knowledge Discovery* 4.1 (2013), pp. 55–63. DOI: 10.1002/widm.1114. URL: https://doi.org/10.1002/widm.1114.

[17] Jones, Z. and Linder, F. Exploratory Data Analysis using Random Forests. 73rd annual MPSA conference, April 16-19, 2015, Chicago, United States of America. 2015. URL: https://pdfs.semanticscholar.org/e7b7/3565b07a7f1369a20b1055f222423f0feb34.pdf.

[18] Breiman, L. Bagging predictors. *Machine Learning* 24.2 (1996), pp. 123–140. DOI: 10.1007/BF00058655. URL: http://dx.doi.org/10.1007/BF00058655.

[19] Henneman, L. R., Holmes, H. A., Mulholland, J. A., and Russell, A. G. Meteorological detrending of primary and secondary pollutant concentrations: Method application and evaluation using long-term (2000–2012) data in Atlanta. *Atmospheric Environment* 119 (2015), pp. 201–210. URL: http://www.sciencedirect.com/science/article/pii/S1352231015302521.

[20] Ricardo Energy & Environment. Kent Air Quality database. 2018. URL: http://www.kentair.org.uk.

[21] NOAA. Integrated Surface Database (ISD). 2016. URL: https://www.ncdc.noaa.gov/isd.

[22] Grange, S. K. ***smonitor**: A framework and a collection of functions to allow for maintenance of air quality monitoring data*. 2018. URL: https://github.com/skgrange/smonitor.

[23] Grange, S. K. *Technical note: **smonitor** Europe*. Tech. rep. Wolfson Atmospheric Chemistry Laboratories, University of York, 2017. DOI: https://doi.org/10.13140/RG.2.2.20555.49448/1. URL: https://doi.org/10.13140/RG.2.2.20555.49448/1.

[24] Jeanjean, A. P. R., Buccolieri, R., Eddy, J., Monks, P. S., and Leigh, R. J. Air quality affected by trees in real street canyons: The case of Marylebone neighbourhood in central London. *Urban Forestry & Urban Greening* 22 (2017), pp. 41–53. DOI: https://doi.org/10.1016/j.ufug.2017.01.009. URL: http://www.sciencedirect.com/science/article/pii/S1618866716303740.

[25] Weiss, M., Bonnel, P., Kühlwein, J., Provenza, A., Lambrecht, U., Alessandrini, S., Carriero, M., Colombo, R., Forni, F., Lanappe, G., Le Lijour, P., Manfredi, U., Montigny, F., and Sculati, M. Will Euro 6 reduce the NO_x emissions of new diesel cars? — Insights from on-road tests with Portable Emissions Measurement Systems (PEMS). *Atmospheric Environment* 62 (2012), pp. 657–665. DOI: https://doi.org/10.1016/j.atmosenv.2012.08.056. URL: http://www.sciencedirect.com/science/article/pii/S1352231012008412.

[26] Grange, S. K., Lewis, A. C., Moller, S. J., and Carslaw, D. C. Lower vehicular primary emissions of NO_2 in Europe than assumed in policy projections. *Nature Geoscience* 10.12 (2017), pp. 914–918. URL: https://doi.org/10.1038/s41561-017-0009-0.

[27] European Environment Agency. *Air quality in Europe — 2016 report*. EEA Report. No 28/2016. 2016. URL: http://www.eea.europa.eu/publications/air-quality-in-europe-2016.

[28] Atkinson, R., Barratt, B., Armstrong, B., Anderson, H., Beevers, S., Mudway, I., Green, D., Derwent, R., Wilkinson, P., Tonne, C., and Kelly, F. The impact of the congestion charging scheme on ambient air pollution concentrations in London. *Atmospheric Environment* 43.34 (2009), pp. 5493–5500. URL: http://www.sciencedirect.com/science/article/pii/S1352231009006268.

[29] Transport for London. Driving. 2018. URL: https://tfl.gov.uk/modes/driving/.

[30] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2018. URL: https://www.R-project.org/.

[31] Grange, S. K. ***rmweather**: Tools to Conduct Meteorological Normalisation on Air Quality Data*. R package version 0.1.2. 2018. URL: https://CRAN.R-project.org/package=rmweather.

[32] Zeileis, A., Leisch, F., Hornik, K., and Christian, K. **strucchange**: An R Package for Testing for Structural Change in Linear Regression Models. *Journal of Statistical Software* 7.2 (2002), pp. 1–38. URL: http://www.jstatsoft.org/v07/i02/.

[33] Zeileis, A., Kleiber, C., Krämer, W., and Hornik, K. Testing and dating of structural changes in practice. *Computational Statistics & Data Analysis* 44.1–2 (2003), pp. 109–123. URL: http://www.sciencedirect.com/science/article/pii/S0167947303000306.

[34] Wright, M. N. and Ziegler, A. **ranger**: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* 77.1 (2017), pp. 1–17. DOI: 10.18637/jss.v077.i01. URL: http://dx.doi.org/10.18637/jss.v077.i01.

[35] International Maritime Organization. Revised MARPOL Annex VI. Annex VI of MARPOL addresses air pollution from ocean-going ships. 2005. URL: http://www.imo.org/en/OurWork/Environment/PollutionPrevention/AirPollution/Pages/Air-Pollution.aspx.

[36] Entec. Defra UK Ship Emissions Inventory—Final Report. Doc Reg No. 21897-01. 2010. URL: https://uk-air.defra.gov.uk/assets/documents/reports/cat15/1012131459_21897_Final_Report_291110.pdf.

[37] Carslaw, D., Apsimon, H., Beevers, S., Brookes, D., Carruthers, D., Cooke, S., Kitwiroon, N., Oxley, T., Stedman, J., and Stocker, J. Defra Phase 2 urban model evaluation. UK AIR: Air Information Resource. 2013. URL: https://uk-air.defra.gov.uk/library/reports?report_id=777.

[38] Catalano, M., Galatioto, F., Bell, M., Namdeo, A., and Bergantino, A. S. Improving the prediction of air pollution peak episodes generated by urban transport networks. *Environmental Science & Policy* 60 (2016), pp. 69–83. URL: http://www.sciencedirect.com/science/article/pii/S1462901116300594.

[39] Charron, A. and Harrison, R. M. Fine ($PM_{2.5}$) and Coarse ($PM_{2.5–10}$) Particulate Matter on A Heavily Trafficked London Highway: Sources and Processes. *Environmental Science & Technology* 39.20 (2005), pp. 7768–7776. DOI: 10.1021/es050462i. URL: http://dx.doi.org/10.1021/es050462i.

[40] Westmoreland, E. J., Carslaw, N., Carslaw, D. C., Gillah, A., and Bates, E. Analysis of air quality within a street canyon using statistical and dispersion modelling techniques. *Atmospheric Environment* 41.39 (2007), pp. 9195–9205. URL: http://www.sciencedirect.com/science/article/pii/S1352231007006863.

[41] Carslaw, D. C. Evidence of an increasing NO_2/NO_x emissions ratio from road traffic emissions. *Atmospheric Environment* 39.26 (2005), pp. 4793–4802. DOI: https://doi.org/10.1016/j.atmosenv.2005.06.023. URL: http://www.sciencedirect.com/science/article/pii/S1352231005005443.

[42] Carslaw, D. C., Murrells, T. P., Andersson, J., and Keenan, M. Have vehicle emissions of primary NO_2 peaked? *Faraday Discussions* 189.0 (2016), pp. 439–454. DOI: 10.1039/C5FD00162E. URL: http://dx.doi.org/10.1039/C5FD00162E.

[43] Brand, C. Beyond 'Dieselgate': Implications of unaccounted and future air pollutant emissions and energy use for cars in the United Kingdom. *Energy Policy* 97 (2016), pp. 1–12. URL: http://www.sciencedirect.com/science/article/pii/S030142151630341X.

[44] Schmidt, C. W. Beyond a One-Time Scandal: Europe's Onging Diesel Pollution Problem. *Environmental Health Perspectives* 124.1 (2016), A19–A22. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4710587/.

[45] Laybourn-Langton, L., Quilter-Pinner, H., and Ho, H. Lethal and illegal: Solving London's air pollution crisis. Institute for Public Policy Research.

2016. URL: http://www.ippr.org/read/lethal-and-illegal-solving-londons-air-pollution-crisis.

[46] Greater London Authority. London Atmospheric Emissions Inventory (LAEI) 2013. 2017. URL: https://data.london.gov.uk/dataset/london-atmospheric-emissions-inventory-2013.

[47] Jenkin, M. E. Analysis of sources and partitioning of oxidant in the UK—Part 2: contributions of nitrogen dioxide emissions and background ozone at a kerbside location in London. *Atmospheric Environment* 38.30 (2004), pp. 5131–5138. URL: http://www.sciencedirect.com/science/article/pii/S1352231004006193.

[48] Carslaw, D. C. and Beevers, S. D. Estimations of road vehicle primary NO_2 exhaust emission fractions using monitoring data in London. *Atmospheric Environment* 39.1 (2005), pp. 167–177. URL: http://www.sciencedirect.com/science/article/pii/S1352231004008775.

Chapter 7

Summary and conclusions

Poor air quality is the leading environmental cause of premature death and exposure to poor air quality is a global issue with 91 % of the human population being exposed to polluted air. Continuous air quality monitoring networks began to be commissioned in the mid-twentieth century to help understand air quality and what gives rise to polluted conditions. Air quality monitoring networks have now generated a very large, and continuously growing observational record. However, this "routine" data has traditionally been underutilised with the primary use being one of simple summaries to check if areas are compliant to legal or guideline values. Undoubtedly a much greater amount of information is contained in these data sets which should be leveraged to improve understanding of this prominent environmental issue. This idea has motivated the work presented in this thesis.

This thesis presents four related research components with the overarching theme of further leveraging air quality monitoring data to develop novel data analytic approaches. These novel approaches aid with extracting additional information on air pollutant emission sources and the physical and chemical atmospheric processes which give rise to elevated pollutant concentrations. All the research presented in this thesis used data from Europe and therefore has a strong focus on transportation activities because these are the activities which dominate air quality issues in the European context.

7.1 Contributions

Chapter 2 presents the development of a new method to extend the established bivariate polar plot visualisations in the **openair** R package with pair-wise statistics. This development was motivated by the need to have readily available tools to be able to extract specific source information in situations where complex source behaviours are present. In these complex situations, emission processes often compete with one another which makes analysis difficult. However, the new technique using wind speed, wind direction and the relationships between two pollutants together was shown to aid source apportionment. Examples of use were given using central London locations using 2013 as the analysis year. The examples demonstrated the utility of the synergy between bivariate polar plots and pair-wise statistics for illuminating black carbon (BC) and particulate matter (PM_{10} and $PM_{2.5}$) sources, with a particular emphasis of partitioning natural and anthropogenic contributions. Using the new approach, the road traffic contribution to $PM_{2.5}$ loading in a complex street canyon central London environment was achieved. Furthermore, the amount of black carbon (BC) composing the $PM_{2.5}$ fraction was quantified. Such an approach is useful in many other environments. These changes have been incorporated in the **openair** package and have seen use by others in other studies.

To allow for productive and standardised data analysis, the **smonitor** Europe database was designed and commissioned (Chapter 3). This database became the primary data source for the other data analysis activities for this thesis. **smonitor** Europe was populated primarily with data sourced from the European AirBase and Air Quality e-Reporting (AQER) repositories with supplementary meteorological data from NOAA's Integrated Surface Database (ISD). The database proved to scale well and in 2018, it contained 4.1×10^9 air quality observations for 12 800 monitoring sites. This database has also seen use in other research activities outside this thesis. It is likely that a public API (application programming interface) will be

developed in the future to enable others to have the same access to the observations presented in this thesis.

Arguably the most important contribution of this thesis was the analysis of European primary NO_2 in the form of roadside NO_2/NO_x emission ratios (Chapter 4). The pan-European analysis of roadside observations demonstrated for the first time that directly emitted NO_2 from road vehicles is decreasing Europe-wide. This offers an optimistic outlook for NO_2 concentrations at European roadside environments when compared to currently used and established emission inventory-based models and emission factors. This approach is unique and suggests that compliance to European NO_2 ambient air quality standards will be achieved faster than expected. This finding is potentially a very important because many European countries and urban areas are debating whether or not to impose very disruptive and expensive low emission zones or similar interventions to reduce roadside NO_2 concentrations. The investment required for such management may not be justified because the observations suggest that the outlook for roadside NO_2 is better than thought and the European passenger vehicle market has shifted post-diesel emission scandal. Both of these factors will most likely combine and result in significantly reduced NO_2 concentrations across Europe's roadside environments without needing to resort to the introduction of low emission zones.

When interventions like low emission zones are implemented, understanding their efficacy for improving air quality is of principal interest. The quantification of a change due to an intervention is however very difficult due to the complex nature of the atmospheric processes and is a general issue across the atmospheric sciences. Methods to robustly detect changes caused by management and intervention activities need to be improved. To address this issue, a machine learning modelling framework called meteorological normalisation was developed (Chapter 5 and 6). Meteorological normalisation allows the use of generally available surface meteorological observations to produce a trend component which displays pollu-

tion concentrations in "average" weather conditions. An R package called **rmweather** was developed containing these tools and is available on R's central repositories (CRAN) with an open source licence.

The meteorological normalisation method allowed for PM_{10} concentrations across Switzerland to undergo a formal trend analysis to investigate if the management of PM has been effective at reducing PM_{10} concentrations (Chapter 5). The technique also suggested two regimes which operate in Switzerland giving rise to elevated PM_{10} concentrations: the first during stagnant, cold, and high emission wintertime conditions, and the second in warm, deep boundary layer conditions where rates of secondary generation of PM were high. However, more interesting was the method's ability in robustly detecting and quantifying changes in air pollutant concentrations. Changes in SO_2 concentrations near a major British port (Port of Dover) were represented almost exactly by changes in the allowed sulfur content in marine fuels (Chapter 6). With an analysis related to the European roadside primary NO_2 study, changes in the NO_2/NO_x emission ratio were very clearly identified at Central London's prominent London Marylebone Road monitoring site (Chapter 6). The features identified at London Marylebone Road outlined issues with retrofitting heavy vehicles (buses) with poorly optimised emission control technology where additional primary NO_2 was directly emitted into the roadside atmosphere. However, despite London's efforts and the tightening of the Euro vehicular emissions standards, NO_x concentrations have remained stable since 2001 at the London Marylebone Road site.

The observations from London Marylebone Road connects to the European roadside NO_2 analysis. Although the features found at this particular Central London monitoring site will not be typical for other European cities, they do demonstrate processes which alter the roadside NO_2/NO_x emission ratio over time. In the London example, despite the introduction of a low emission zone, congestion charging, and the European-scale tightening of the Euro vehicular emission standards, roadside NO_2 and NO_x concentra-

tions have not significantly improved. Therefore, in respect to air quality the returns on such interventions have been small which may also indicate low emission zones are not a highly productive investment if the objective is a significant improvement of roadside air quality.

The work presented here has been published in peer reviewed scientific journals between 2016 and 2019. All articles, with the exception of the recently published (three months before thesis submission) meteorological normalisation intervention analysis have been cited by others. The earliest polar plot enhancements paper has seen the greatest number of citations while the European primary NO_2 trends paper was reported in a number of popular media sources at the time of publication and has since been cited in a British parliamentary document, thus demonstrating public and political interest. It is possible to partially count the number of downloads of the **rmweather** R package and currently this package is seeing approximately 80 downloads a week. I hope the research presented is continued to be used and improved upon in the future.

7.2 Future directions

The work presented here could be extended in the future in a number of ways. In the case of the bivariate polar plots, a technique which has been implemented, but not yet explored is quantile regression. Quantile regression is useful in determining how different components of distributions are behaving and would most likely be another technique which can aid source apportionment. There are a number of commercially available air quality instruments which report tens or hundreds of variables, most notably metal monitors and PM monitors which bin PM into a large number of size fractions. Using the correlation statistic with polar plots for all variables in a "polar plot correlation matrix" could offer an interesting visualisation for those investigating large number of pollutants. Such a matrix could also be combined with an unsupervised learning method such as clustering to sug-

gest groupings of atmospheric pollutants and therefore illuminate different emission sources. The clustering could identify pollutants with similar behaviours and could give very good evidence for specific emission sources.

Much of the research presented in this thesis relied on standardised access to a collection of publicly available data sets. The importance of this component needs to be emphasised because it allowed for efficient and productive data analysis activities to be conducted. Air quality is a data intensive domain and therefore careful thought was put into this process and resulted in the **smonitor** Europe database (Chapter 3). The next step for **smonitor** Europe is to create a publicly accessible API so others can benefit from the same access which was available for this thesis's research. This is not a big task, but it does require a financial commitment to allow for the set-up and maintenance of a publicly accessible web server.

The trends of the European roadside NO_2/NO_x emission ratio requires more investigation to understand the mechanisms responsible to explain why the peak emission ratio was experienced in 2010 and has subsequently decreased. An effective way of addressing this uncertainty is to conduct dedicated vehicle emissions measurements of NO and NO_2 for a wide range of vehicle types, ages, and engine sizes. Chapter 4 gives a number of reasons to explain the features observed, but there is a lack of evidence for many of these suggestions. The nature of the trend was consistent among Europe's urban areas however, and therefore the most important explanations will most likely arise from activities at a European level such as the Euro vehicular emission standards.

In the future, it is very likely that the trend will enter a new phase due to the change of the European passenger vehicle market after the fallout by the Volkswagen diesel emission scandal. Diesel-fuelled passenger vehicles are no longer as popular as they once were in Europe and their decreasing market share is predicted to continue with gasoline-powered, hybrid, and plug-in hybrid vehicles growing in popularity to make up the shortfall. Mazda also plans to put the first gasoline-fuelled homogeneous charge com-

pression ignition (HCCI) engine in 2019 called Skyactiv-X, perhaps giving rise to a revolution in the technology used for internal combustion engines used in passenger cars which will influence roadside NO_x (and other pollutants) concentrations. Somewhat frustratingly, some European states such as Belgium are no longer reporting NO_x in their air quality data submissions. This is because there is not a legal requirement to do so and have taken the decision to only report the pollutants which are legally required of them which could result in a future where the method used to calculate the NO_2/NO_x emission ratio in Chapter 4 will be unable to be used.

The research presented in this thesis used surface based air quality monitoring data but there are alternative or complementary data sources which could be leveraged to help explore and understand air quality. Most notably, are the vast data sets created by satellite remote sensing. There is significant potential for the earth observation the air quality communities to work together in the future to investigate the issues explored throughout this thesis.

The meteorological normalisation technique presented in Chapter 5 and 6 lends itself to a large number of potential applications involving trend analysis or intervention exploration, most notably exploring the efficacy of low emission zones or congestion charging on improving air quality. This is relevant for EU member states because many are developing such interventions to mitigate high concentrations of air pollutants, especially NO_2. Development in the machine learning domain is extremely rapid and it is likely that new and better algorithms will become available in the future which would allow the technique perform better. A major advantage of the random forest algorithm is the non-black box nature of the method because the ability to evaluate the learning process which can be used to investigate physical and chemical atmospheric processes. The ability and methods used to evaluate the learning process will also likely improve in the future.

Another application of the random forest models has come to light since publication of the meteorological normalisation technique in the form of its

use in air quality forecasting. Once a model has been trained, the result is a predictive model which can explain pollutant concentrations at a particular location based on meteorological and time variables. The explanatory ability of such models can be rather impressive and therefore, if high quality weather forecast data is available, the models can be used to predict pollutant concentrations into the future. This is a very worthy avenue of future use of such models because it does not require the very large computing resources needed to create "bottom-up" models for prediction. The same philosophy could also see use in air quality time series interpolation where missing observations due to instrument downtime could be filled with modelled data for the purpose of ensuring a complete time series, something which is often a prerequisite for modelling.

7.3 Final remarks

European air quality is improving over time in almost every respect. However, it seems that that because air pollutants have no threshold value where their negative health effects are the same as a pollutant concentration of zero, work must continue to decrease air pollutant concentrations to their lowest possible levels. The work here contributes a few small components towards this goal, but it is hoped that the research allows others to use routinely collected air quality monitoring data, from the past and the future, in a more effective way to understand the issues surrounding air quality and to reduce the negative effects of this prominent environmental issue.