University of Glasgow

Statistical Modelling of Air Quality in Aberdeen

Cillian Francis Doherty

Submitted in fulfillment of the requirements for the Degree of Master of Science in Statistics

School of Mathematics and Statistics

College of Science and Engineering

University of Glasgow

April 2017

Abstract

This thesis focuses on modelling air pollution in Aberdeen. It takes into account how traffic and meteorological variables affect the Nitrogen Dioxide concentrations at a number of different sites throughout the city during the year 2014. The aim of the thesis is to build a regression model of spatial and temporal concentration variations and use inverse regression to develop a tool to identify control mechanisms that will help manage Nitrogen Dioxide concentrations in an urban setting. This is of particular importance to the Scottish Environment Protection Agency (SEPA).

Chapter 1 focuses on the motivation for carrying out such a study, as well as the aims and objectives. The data are introduced in this Chapter. These include data from different AURN (Automatic Urban Road Network) sites in Aberdeen, as well as diffusion tube data, traffic counts from different locations as well as meteorological data recorded at Dyce Airport.

Chapter 2 covers the temporal modelling of air quality in Aberdeen using time series analysis. Time series methodology is explored which includes an initial exploration of the model variables using linear regression; followed by residual diagnostics; time series regression; the definition of autocorrelation function (ACF), partial autocorrelation function (PACF) and stationarity; the exploration of seasonality and harmonic regression, and ends with generalized additive model methodology. This spans from 2006-2015.

Chapter 3 investigates the spatial modelling of air quality in Aberdeen. This is done through numerical and graphical summaries. Methods used to explore NO_2 data are presented. This includes geostatistical modelling. Two different models are investigated. Model parameters are estimated, using maximum likelihood estimates and restricted maximum likelihood estimates. This is followed by prediction of future values, using a statistical technique known as Kriging.

Chapter 4 uses inverse regression to estimate road traffic flows required to achieve compliance with national air quality objectives. This Chapter also presents the usefulness of inverse regression.

Chapter 5 ends with a discussion on what further work can be done, and any conclusions for this thesis. It looks at the strengths and weaknesses of each Chapter in turn.

Table of Contents

List of Tables

List of Figures

Acknowledgement

I would like to thank my advisers Duncan Lee and especially Prof. Marian Scott. I have been very fortunate to receive her help and guidance. Thanks for the countless changes of "data is" to "data are", Marian. Thank you to the guys in 420 who helped make the days go quicker. Thank you to SEPA for providing the data. And thank you to the University for funding me. Most of all thank you to my family for supporting me for the duration of the thesis.

Declaration

I have prepared this thesis myself; no section of it has been submitted previously as part of any application for a degree. I carried out the work in it, except where otherwise stated.

Signed: _____

Date: _____

Chapter 1: Introduction

1.1 Air pollution and standards already in place

Air pollution can be categorized into airborne particles and gases, making it a challenge to monitor. There was an article [1] in 2014 which introduced the issue of air pollution on Scottish streets. This highlights the importance of the issue as well as the fact a respected media agency is taking an interest. Also highlighted are the adverse health effects which are associated with high levels of air pollution and the streets in Scotland which are most affected by said air pollution. The pollutants which are focused upon most are Nitrogen Dioxide and Particulate Matter 10 (NO_2 [2] and PM_{10} respectively). Streets in Glasgow and Aberdeen are among the top of the list of the most polluted streets in Scotland according to 2014 article [1] and a 2016 article [3] by the BBC.

Air pollution can be described as particulates and harmful materials which are introduced into Earth's atmosphere (which in turn can be described as a layer of gases which surround the earth), leading to disease, allergic reaction, ill health/death in humans, damage to other living organisms such as plants and animals, and damage to the natural or built environment. It can be caused by natural sources, and may come from anthropogenic sources. Earth's atmosphere is a naturally occurring system of gases which is essential to support and maintain life on Earth. Estimates from a 2014 World Health Organisation report and the International Energy Agency have air pollution causing approximately 7 million deaths worldwide in 2012 [4][5].

For the purpose of this thesis we look at Nitrogen Oxides, mainly NO_2. NO_2 is formed during high temperature combustion, as well being generated in the form of electric discharge during thunderstorms. NO_2 is visible to the human eye in the form of a brown haze [6] dome above a city or a plume downwind of cities, although this is likely to be due to airborne particles than NO_2. They also carry an odour which is known to be sharp.

The main anthropogenic sources for NO_2 are stationary sources such as power plants, factories for manufacturing and waste incinerators. There are also mobile sources which include motor vehicles, marine vessels and aircraft [7]. The main source of interest in this thesis is motor vehicles, and this is further broken down into different motor vehicle classes.

Air pollution can have an adverse effect on human health [3][8][9]; this includes - and is not

exhaustive of - the following in the long term; mortality, ischaemic heart disease, stroke, chronic obstructive pulmonary disease (COPD), and lung cancer, as well as acute lower respiratory infections in children. Asthmatics in particular may be affected by short term exposure of high levels of air pollution although the general population may suffer to a lesser degree by experiencing a dry throat and sore eyes. There are also agricultural [10] and economic [11] effects which contribute to the argument that air pollution needs to be constantly monitored and reduced to levels which are healthy for humankind and the Earth.

One of the most notable periods of air pollution in Scotland or England was in 1952 when what is known as The Great Smog or The Big Smoke took place in London [12]. A huge cloud of smoke descended over London for four days, which was so thick that people could not see more than a few feet in front of them. This caused the transport system to come to a halt and between 4000 and 12,000 casualties [13]. It was this event which led to the monitoring and investigation of air pollution levels, and ultimately to an increase in public awareness of the health effects of pollution and the resulting research and regulation. After The Big Smoke occurred the UK government took reactionary measures and as a result the Clean Air Acts of 1956 and 1958 were drawn up and passed [12].

Throughout the world, there are various air pollution control technologies and strategies in place to reduce air pollution. These range from land-use planning to the development of the use of cleaner power sources such as wind, solar and hydro power, which do not cause air pollution. For mobile sources, conversion to cleaner fuels or electric vehicles is taking place. There is strong evidence to suggest that this is better than using fossil fuels and this transition is as a result of public opinion as well as that of the scientific community being firmly in favour that an increase in and long term exposure to air pollution can have a negative effect on health. Studies which have been done on air pollution show the effect it can have on human health. Some notable ones are that of Dockery *et al.* [14], Pope III *et al.* [15], and Dominici *et al.* [16]. These have respectively concluded that; air pollution is statistically significantly associated with mortality, and positively associated with lung cancer deaths and cardiopulmonary disease; air pollution was associated with cardiopulmonary and lung cancer mortality; and there is an increased risk of hospital admission for cardiovascular and respiratory diseases with short-term exposure to air pollution.

Across the world air pollution is being measured due to an increased level of awareness of air pollution and the effects it has on human health. Within Europe this is done by a partnership of two agencies or networks; the European Environment Agency (EEA) [17] and the European Environment Information and Observation Network (EIONET) [18]. The EIONET within any co-operating countries supports the collection and organization of available data. This is passed on to the EEA which provides information to government bodies and institutions as well as making it available to the general public; this is the case in Scotland for example. The EEA explores the data in the hope of understanding the environment it is related to and providing information which could be used in the change of policy. Relevant data are accessible to governing bodies and politicians, as well as the public, which means that there is available information regarding the state of the environment. It should also be mentioned that both the UK government and the environmental agencies monitor air quality.

There are a number of regulations in place and different types of air quality standards. These include, although are not limited to, the U.S National Ambient Air Quality Standards, the E.U. Air Quality Directive, the North American Air Quality Index, the Air Quality Health Index (Canada), and TA Luft (Germany). At a European level, the European Union sets regulations which its member states must adhere to. Within these countries they each have their own governing bodies that are subject to a fine from the EU if they do not comply with said regulations. Within Scotland, the Scottish Government has outlined a set of air quality guidelines derived from the EU, of which it tries to achieve across the country. At a UK level, the Department for Environment, Food and Rural Affairs (Defra) [19] and the Scottish government run Scottish Air Quality [20] each regulate and monitor air quality standards in the UK and Scotland respectively. These limits are given in terms of annual means, and are the same for the EU, the UK and Scotland. The NO_2 concentration limit is $40\mu gm^{-3}$ for each zone.Defra has responsibility for the Air Quality Strategy which has been set out in the different regions of England, Scotland, Wales and Northern Ireland.

Table 1.1.1 below shows the air quality objectives i.e. metrics of compliance with standards. This is different to the air quality standards which are numerical concentration thresholds over specified averaging periods;

Pollutant	Applies	Objective	Concentration measured as	Data to be achieved by and maintained thereafter
Nitrogen Dioxide	UK	$200\mu gm^{-3}$ not to be exceeded more than 18 times a year	1 hour mean	31 December 2005
	UK	$40\mu gm^{-3}$	Annual mean	31 December 2005

Table 1.1.1: National Air quality objectives for the protection of human health [24]

This table is an extension of the fact that the UK has an NO_2 annual concentration limit of $40\mu gm^{-3}$.

There has been an established relationship between air pollution and meteorological data for a number of years. Air pollution studies which include meteorological effects tend to include as a covariate ambient temperature. It was meteorological effects which contributed to the Great London Smog [12] and it has been proven recently that the effect of temperature on morbidity rates is a continually important problem [21]. In light of this, temperature and other meteorological factors have been included when analyzing the NO_2 data from diffusion tubes and AURN sites in a temporal sense.

1.2 Introduction to the data

1.2.1 Diffusion tube and AURN monitoring site data

The NO_2 monitoring site data are provided by SEPA, and can also be found on the Scottish Air Quality website (which is a government run website) [20]. This website is an easy-to-use, well developed interface and has real time data available for the user. Historically, the data date as far back as the mid-1980s for some sites, and although some sites are now closed, any running ones have data available for today. These data go through an encompassing system of verification and checking to ensure that the data are as accurate and close to real-time as possible. There are a number of different methods available for monitoring air quality, with automatic monitoring sites being one of the more accurate methods as it sets a limit on the amount of human error which can be included while providing a high temporal resolution. There are also summary statistics available to coincide with the real time data and there are a number

of pollutants which are also measured at over 90 different sites as well as NO_2. These include PM_{10}, $PM_{2.5}$, Ozone, Carbon Monoxide (CO) and Sulphur Dioxide (SO_2). All pollutants measured are measured in micrograms per cubic meter μgm^{-3}.

There are five Automatic Urban Road Network (AURN) sites which have available data from 2006 to the present day in Aberdeen. These can be seen in the table below (Table 1.2.1). These sites are not uniformly distributed throughout the city, as can be seen from Figure 1.2.1, and hence may not give a representative spatial representation of NO_2 pollution across Aberdeen. The sites are classified by the Scottish Air Quality website according to the environment in which they are situated, with the website having 12 different classifications and two of these being applicable to the five sites in Aberdeen. This is also shown in the table below (Table 1.2.1). The classification which is observed most often is the "roadside" classification. This is described by Scottish Air Quality as "a site sampling between 1 m of the kerbside of a busy road and the back of the pavement. Typically this will be within 5 m of the road, but could be up to 15 m." The other environmental classification of interest which applies only to the station at Errol Place, is "urban background" classification. This is described as "an urban location distanced from sources and therefore broadly representative of city-wide background conditions e.g. urban residential areas."

Site	**Classification**
Anderson Drive	Roadside
Errol Place	Urban Background
King Street	Roadside
Union Street	Roadside
Wellington Road	Roadside

Table 1.2.1: Environmental Classification for each AURN site

Below is a table with the basic summary of the actual concentrations (in μgm^{-3}) for each site during 2014;

Site / Statistic	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Anderson Drive	0	10	19	25.75	36	316
Errol Place	-2	10	19	23.64	32	182
King Street	0	15	25	29.81	40	172
Union Street	0	26	46	52	72	411
Wellington Rd	0	21	42	50.23	71	262

Table 1.2.2: Basic Statistics of NO_2 at AURN sites for 2014 ($\mu g m^{-3}$)

The figure below shows the spatial distribution of the 5 AURN sites. It is clear that they are not distributed uniformly throughout the city, and they almost fall in a linear pattern from South to North, with the exception of Anderson Drive. In table 1.2.2 the minimum value at Errol Place is negative, this reflects some residual quality assurance issues.

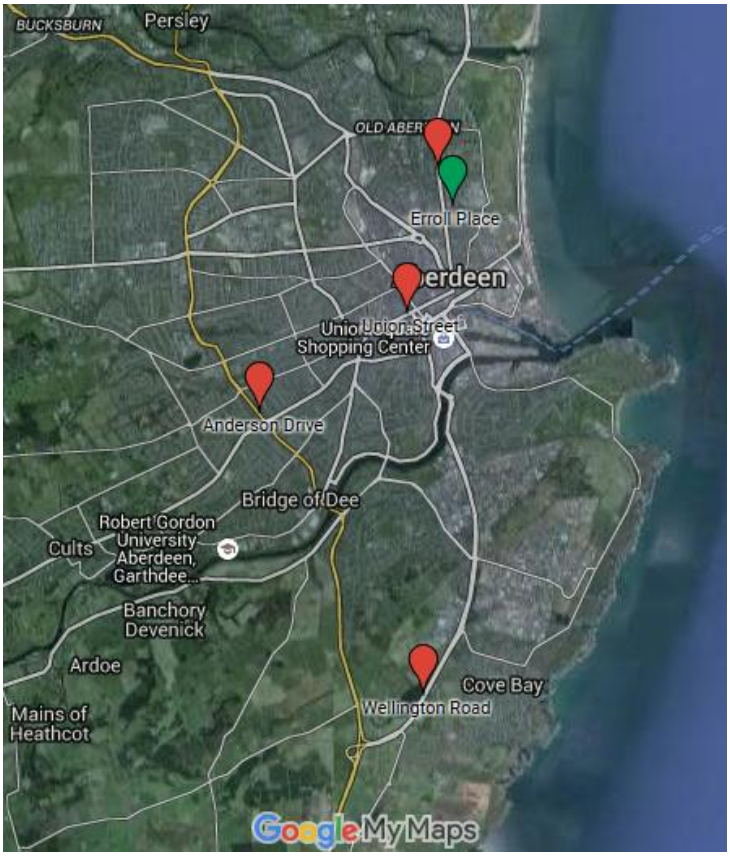


Figure 1.2.1: Map Showing locations of AURN sites in Aberdeen

Local authorities (LAQM) manage diffusion tubes. These diffusion tubes collect NO_2 data, as well as other air pollutant data, although the focus in this thesis is on NO_2 data. These data are looked at in this thesis alongside the AURN data. According to the LAQM website [25], diffusive samplers are widely used as an indicative monitor of ambient NO_2 in the context of review and assessment. Diffusion tubes are particularly useful [25]:

- when simple indicative techniques will suffice;
- to give an indication of longer term average NO_2 concentrations;
- for indicative comparison with the Air Quality Strategy Objectives based on the annual mean;
- for highlighting areas of high NO_2 concentration.

Some known limitations to passive diffusion tubes are that they have positive bias caused by in within tube chemical reactions. Positive bias is also caused by wind turbulence when using diffusion tubes. The diffusion tubes having this positive bias, although giving a "conservative" estimate of actual air concentrations, can also provide useful means for comparing the air quality in different areas [75].

There are 51 locations in Aberdeen City which are diffusion tube monitoring sites. They are useful for identifying areas of high NO_2, particularly when monitoring traffic emissions, where the concentration does not vary much from day to day. A map showing the locations is found in section 3.2. It is also of interest the sites where the 51 loactions are located. The monitoring sites are in the locations that they are because of known links between air pollution and poor health. This is known as preferential sampling and can be used to assess environmental compliance in polluted areas [76].

The temporal frequencies of the AURN data and the diffusion tube data are an hourly and daily (although presented annually) frequency respectively i.e. the concentration of NO_2 is recorded every hour of every day for the AURN sites while the concentration is recorded daily for the diffusion tube sites, and presented as an annual average. For time series analysis and modelling the data are aggregated to a daily level for the AURN sites i.e. an average is taken for each 24 observations (corresponding to 24 hours in a day). For the spatial analysis and modelling, the

data for the diffusion tube sites are already at an annual level, while the AURN data are further aggregated to an annual level and one year is used for modelling, namely 2014.

1.2.2 Meteorological Data

In addition to the NO_2 data there are meteorological data available, also from SEPA. This same data can be found on the Weather Underground website which is a commercial weather service founded in 1995 and is a part of The Weather Channel Companies. The data themselves consist of a number of meteorological variables, recorded hourly at Dyce airport, which is located to the North-West of Aberdeen. Not having meteorological recordings at each AURN site is a limitation to the analysis of the data, although the data recorded at Dyce can be applied to each site. These are the most accurate data available. There are data available at Dyce airport as far back as 2000 and is an hourly temporal format.

From the meteorological variables which are available to analyse, the following are selected; wind speed, wind direction, cloud cover, rainfall, temperature, humidity, pressure at mean sea level. Meteorological variables have been shown to have an effect on air pollution and hence NO_2 [26] [27]. Wind speed is measured in kilometer per hour (kmh^{-1}), wind direction in degrees; cloud cover in oktas; rainfall in millimetres (mm); temperature in ℃; humidity measures the amount of water vapour in the air and is taken as a percentage; and pressure is measured in pascals (Pa).

1.3.3 The traffic data

The traffic data are taken from a data set provided by SEPA and were originally in the form of annual figures for different roads and at different road lengths. They are counts recorded at specific road links, and are known as ‘count points’. Having this in an annual format was not consistent with the aggregated data for the NO_2 concentrations and the meteorological measurements such as wind speed or temperature. In order to have all variables on the same temporal level, the traffic count data are disaggregated, using figures from the Department for Transport, as well as figures from SEPA for different roads around Scotland. Each disaggregation process is described below and how these approaches are used in conjunction with one another to obtain the data in the format preferred for analysis is explained;

Data provided by SEPA

– This data provided by SEPA depicts the average number of different vehicle types per day travelling past certain count points over a number of years. These data are manipulated so that instead of showing all count points in Scotland overall years between 2000 and 2015, they depicted the years 2006 – 2014 at Anderson Drive specifically. The vehicle types are as follows; motorcycles, cars and taxis, buses, light goods vehicles, and different types of heavy goods vehicles. A sample of the observations looks like the following:

Year	Motorcycles	Cars Taxis	Buses Coaches	LGVs	All HGVs	All Motor Vehicles
2006	41	7213	14	1659	228	9155
2007	42	6931	14	1692	232	8912
2008	44	7029	13	1790	216	9092
2009	28	6181	11	787	430	7437
2010	80	6460	8	1381	251	8179

Table 1.3.3.1: Average Vehicle counts per day during different years in Aberdeen

1st dataset from Department of Transport (tra0306) [28]

– These data (the 1st of two data sets from the Department of Transport) are the average distribution by day of the week, for Scotland i.e. these data are of values at or around 100. This value of 100 is taken as an index, that is, an average day is taken as 100. These data range from the years 2006 – 2014. A sample of these data can be seen in table 1.3.3.2:

Day of the week	2006	2007	2008	2009	2010
Sunday	82	81	78	81	80
Monday	101	103	103	103	104
⋮	⋮	⋮	⋮	⋮	⋮
Saturday	85	81	103	106	109

Table 1.3.3.2: Sample of average distributions of traffic count by different days of the week for the years 2006 - 2010

2nd dataset from Department of Transport (tra0307) [28]

– These data (the 2nd of two data sets from the Department of Transport) are the average daily traffic flows by month of the year, for Scotland. Similar to the previous data set it has values at or around 100. This value is taken as an index also, that is, the average daily traffic flow in a month is taken as 100. This data also ranges from the years 2006 – 2014. A sample of these data can be seen in table 1.3.3.3:

Month	2006	2007	2008	2009	2010
January	91	90	90	90	90
February	94	94	95	94	93
⋮	⋮	⋮	⋮	⋮	⋮
December	93	93	92	92	91

Table 1.3.3.3: Sample of average daily traffic flows by month of the year for the years 2006 - 2010

The datasets are used in conjunction to create the disaggregated data. It is possibly best to explain how they are used by using the previous three tables and providing an example.

Let α be the outcome of interest, i.e. a particular vehicle type during a particular day, month and year. It follows that;

$$\alpha = \frac{\left(\frac{x_{i,j}}{100}\right) \times y_{k,l}}{100} \times z_{m,n} \qquad (1.3.3.1)$$

where $x_{i,j}$ is an entry from table 1.3.3.1, $y_{k,l}$ is an entry from table 1.3.3.3, and $z_{m,n}$ is an entry from table 1.3.3.2. Say the outcome of interest, α, is the number of motorcycles which passed the count point at Anderson Drive on the 1st January 2006. In this case; $i = j = k = l = m = n = 1$ i.e. the first entry in each of the respective tables is used in equation *1.3.3.1* (since $i, k \; and \; m =$ row entries and $j, l \; and \; n =$ column entries) and α can be found as the following;

$$\alpha = \frac{\left(\frac{41}{100}\right) \times 91}{100} \times 82 \approx 31$$

So we can say 31 motorcycles passed the count point at Anderson Drive on the 1st January 2006. This process described is reiterated for every date between the 1st January 2006 and 31st December 2014, and for all vehicle classes. It is pseudo data in a sense, so it has its limitations, although it is useful still.

1.3 Aims

The aims of the thesis are as follows;

- To model the temporal patterns in the NO_2 data recorded at the AURN sites;
- To model the spatial patterns in the data recorded at both the AURN sites and the diffusion tube for the year 2014;
- To model the effects of covariates such as meteorology and traffic on NO_2 concentrations;
- To use inverse regression to explore traffic conditions to meet certain NO_2 conditions.

The aims are carried out in chronological order. Chapter 2 covers the temporal patterns of the NO_2 data and Chapter 3 the spatial patterns of the NO_2 data. Both Chapters 2 and 3 model the effects of the covariates mentioned above and Chapter 4 focuses on inverse regression. Inverse regression explores under what conditions are the model covariates so that a set level of NO_2 is not exceeded. This is a useful tool for advising on policy, and is discussed in more detail in Chapter 4.

Chapter 2: Temporal Modelling of Air Quality in Aberdeen

At each of the AURN monitoring sites the NO_2 concentration is recorded. This is recorded at an hourly rate, as mentioned previously. In this Chapter, these hourly data are aggregated to a daily level, followed by a temporal modelling and analysis on the data. The Chapter begins with numerical and graphical summaries of the data, then the data is modelled by a linear method followed by a general additive method.

2.1 Methodology

One general approach which is taken to modelling data has three stages, these are;

1. Model identification and selection
2. Parameter estimation
3. Model checking e.g. assumptions including independence of residuals.

If at this last step the model is not useful i.e. the estimation is inadequate, the process is repeated in order to build a better model. Stationarity and seasonality (as discussed below) must be identified, and an autocorrelation plot is used for this.

2.1.1 Exploring model variables using linear regression

A brief outline of a simple regression model is given where y_t is the response variable which in this case is $\log(NO_2)_t$, where $t = 1, \dots, T$ is an index for time. The log transformation is taken due to the nature of the data – it is slightly skewed and so a transformation is needed to help normalise it. This is discussed in more detail later. Assuming that the response variable is being influenced by a series of explanatory variables $x_{k,t}$ where $k = 1, \dots, K$ and $t = 1, \dots, T$, the relationship between NO_2 and the explanatory can be described by a linear regression model;

$$y_t = \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_K x_{K,t} + \varepsilon_t \qquad (2.1.1.1)$$

In this model $(\beta_1, \dots, \beta_K)$ are unknown and fixed regression parameters and $\{\varepsilon_t\}$ is the random error term which is assumed to have a mean of zero $\varepsilon \sim N(0, \sigma^2)$. One popular approach for fitting such a model is through Ordinary Least Squares (OLS) whereby the linear model above can also be written in matrix notation:

$$Y = X^T \beta + \varepsilon \qquad (2.1.1.2)$$

and the least squares estimate of β is given by the following;

$$\hat{\beta} = (X^T X)^{-1} X^T Y \qquad (2.1.1.4)$$

When a standard linear regression model which uses the OLS estimation technique is used, there are a number of assumptions made, and these assumptions must hold if parameter estimates are to be accurate. The first assumption is one of homoscedasticity, which means that the errors have constant variance. Other assumptions include a linear relationship between response and covariates; multivariate normality i.e. the data is normally distributed; minimum multicollinearity; and no autocorrelation. Each of these assumptions are checked in the next section, and there is more discussion surrounding them.

2.1.2 Time Series Regression

A time series can be described as a set of observations y_t, with each y being observed at a specified point in time t. Although a time series can be either discrete or continuous, given the nature of the recorded NO_2 concentrations at sites in Aberdeen, a discrete time series is used here. This time series of a discrete nature can be described as one in which the set T_0 of times at which observations are made is a discrete set, which is also the case when observations are recorded at fixed time intervals, for example annual recordings, or monthly recordings.

It is necessary, in order to understand time series regression, that a hypothetical mathematical model is created to represent the data. Once a model is chosen, it is then possible to estimate parameters, check for goodness of fit to the data and then possibly use the fitted model to develop our mutual understanding of the process generating the time series, in this case traffic and meteorological factors in Aberdeen. Once a satisfactory model has been built, it may be used in a variety of ways. This includes separating noise from signals, predicting future values of the time series, and controlling future values.

A model may be used to represent a compact description of data, for example for NO_2 concentrations at Union St, Aberdeen, and coming in the form of a sum of a specific trend $\{m_t\}$, seasonal $\{s_t\}$ and random $\{\varepsilon_t\}$ terms. For this data, it is important to implement seasonal adjustment. That is to recognise the presence of seasonal components and to remove them from the model so they are not confused with long-term trends. An example of an additive model

with a trend component and a seasonal component is below (it also includes a remainder or error term);

$$y_t = m_t + s_t + \varepsilon_t \qquad (2.1.3.1)$$

Autocorrelation plots are useful, particularly in time series modelling, since failing to account for said autocorrelation can lead to the use of incorrect standard errors. Autocorrelation is the correlation between elements in a series and elements in that same series, separated by a given time interval. Using the sample autocorrelation function (acf) and the partial autocorrelation function (pacf) plots are the most common ways for checking if autocorrelation is present in the residuals. These are discussed in more detail below.

There is a class of time series models called autoregressive processes, which are the most common models for correlated data [29]. An autoregressive process can be seen below;

$$y_t = \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + \varepsilon_t \qquad (2.1.2.1)$$

where y_t is the observation, ε_t is the (unobservable) random disturbance (noise) at time t, p is the order of the process and $\beta_1, \ldots, \beta_p$ are the parameters of the model. These processes are used for looking at the residuals of the models and seeing how they may be correlated.

2.1.2 Model Diagnostics and Evaluation

Examining the results after the model has been fit, which are defined by $r_t = y_t - \hat{y}_t$ where $\hat{y}_t$ is the fitted value at time *t*, we can assess the model assumptions. Plotting the residuals against time t, the residuals should have a 0 mean with an equal spread above and below the mean with constant variance. Checking for non-constant variance is done by looking for the presence of a fanning out of the residuals. The assumption of normality must also hold, and this is checked using a histogram or a QQ plot. Finally, assuming that errors are uncorrelated with one another must hold true – which is not the usual case for time series data. These residuals of the model can highlight problems related to assumptions made when modelling.

The AIC stands for Akaike Information Criteria and is a measure of the relative usefulness of statistical models created for a single set of data. When the AIC is provided with a set of models for the data, it compares them with one another. Hence, it is a method of model selection.

R-squared explains the proportion of the variation of the response explained by the independent variables for a linear regression model like the ones created for the AURN sites. Adjusted R^2 adjusts that statistic depending on the number of independent variables in the model.

2.1.4 ACF/PACF and stationarity

A time series $\{Y_t, t = 0, \pm 1, ...\}$ loosely speaking, can be said to be stationary if it has statistical properties similar to those of a certain "time-shifted" series $\{Y_{t+h}, t = 0, \pm 1, ...\}$, for each integer h. Focusing attention to the properties which only depend on the first- and second-order moments of $\{Y_t\}$, we can make this idea precise by defining the following;

Let $\{Y_t\}$ be a time series with $E\{Y_t\}^2 < \infty$. The **mean function** of $\{Y_t\}$ is:

$$\mu_Y(t) = E(Y_t) \qquad (2.1.4.1)$$

The **covariance function** of $\{X_t\}$ is

$$\gamma_Y(r, s) = Cov(Y_r, Y_s) = E[\big(Y_r - \mu_Y(r)\big)\big(Y_s - \mu_Y(s)\big)] \qquad (2.1.4.2)$$

For all integers r and s.

$\{Y_t\}$ is (**weakly**) **stationary** if

(i) $\mu_Y(t)$ is independent of t.

(ii) $\gamma_Y(t + h, t)$ is independent of t for each h. This can also be written as $\gamma_Y(h)$.

It should be noted that there also occurs strict stationarity in time series although, generally speaking, any time series which is strictly stationary, is also weakly stationary. This is because the strict stationarity of a time series applies to at least two variables, whereas the weak stationarity (or just stationarity) of a time series applies to just one variable.

Let $\{Y_t\}$ be a stationary time series. The **autocovariance function** (ACVF) of $\{Y_t\}$ is

$$\gamma_Y(h) = Cov(Y_{t+h}, Y_t) \qquad (2.1.4.3)$$

The **autocorrelation function** (ACF) of $\{Y_t\}$ is

$$\rho_Y(h) \equiv \frac{\gamma_Y(h)}{\gamma_Y(0)} = Cor(Y_{t+h}, Y_t) \qquad (2.1.4.4)$$

The ACVF and ACF provide a very useful measure of the degree of (in)dependence between values recorded in a time series at different times and for this reason are useful for assessing assumptions made of time series models.

2.1.5 Seasonality/Harmonic Regression

Data can be represented as a realisation of the process, i.e. the classical decomposition model;

$$y_t = m_t + s_t + \varepsilon_t \qquad (2.1.5.1)$$

where m_t is a slowly changing function known as a trend component, s_t is a function with known period d referred to as the seasonal component, and ε_t is a random noise component which is stationary as defined by definition 2.1.2.

There is not always an obvious trend component m_t present for data, sometimes there is only a seasonal component and a random noise component, s_t, ε_t respectively, visibly present. It is also of interest, in order to determine the stationarity of the random noise component, to estimate and extract the trend and seasonal component. The theory for these processes can be used to find a suitable probabilistic model for the random noise process ε_t, to analyse the properties of the process, and to use it in conjunction with the trend and seasonal components for possible predicting and emulating of $\{Y_t\}$. There is another approach [30] which applies differencing operators repeatedly to the series $\{Y_t\}$ until the differenced recordings resemble a realisation of a stationary time series $\{W_t\}$.

Harmonic regression is described in more detail here. When the case arises that there appears to be a cyclical or seasonal patterns across time, one or more harmonic functions can be used to capture this seasonality. The equation below, discussed in [31], is the basis for basic harmonic regression;

$$y_t = \beta_0 + A\cos(2\pi wt + \psi) + \varepsilon_t \qquad (2.1.5.2)$$

where y_t is the response variable which in this case is log(NO_2), w is the cycle component which determines the frequency of the wave, t is the time index, β_0 is the intercept term, A is the magnitude of the wave and ψ location of the start of the phase. It is assumed that w and t are known parameters and A and ψ are unknown. Using the angle sum trigonometric identity in the following equation

$$cos(\alpha \pm \beta) = cos(\alpha)cos(\beta) \mp sin(\alpha)sin(\beta) \quad (2.1.5.3)$$

the harmonic regression can be written in terms of the following equation;

$$y_t = \beta_0 + \beta_1 \cos(2\pi wt) + \beta_2 \sin(2\pi wt) + \varepsilon_t \quad (2.1.5.4)$$

Other terms such as linear ones can be included in the model, for example taking the meteorological factors of wind speed and humidity we could have;

$$y_t = \beta_0 + \beta_1 \cos(2\pi wt) + \beta_2 \sin(2\pi wt) + \beta_3(Wind\ Speed) + \beta_4(Humidity) + \varepsilon_t \quad (2.1.5.5)$$

2.1.6 Generalised Additive Model Methodology

It may be of use to initially fit a linear model. A generalised additive model may be of use to fit to the data if the linear model has been proven to not be a good fit for the data since the trend is not linear. It is useful to describe the relationship between the response (log NO_2) and the explanatory variables by some function or functions which take the following form;

$$f(X, \beta) \quad (2.1.6.1)$$

This is a function of the covariates X and their respective coefficients β. The idea of such a model arises from the fact that in many real life situations, effects are not linear [31]. A generalized additive model is described as a generalized linear model with a linear predictor involving smooth functions of covariates [32] [33]. The general structure of the model is something of the following;

$$g(\mu_i) = \boldsymbol{X}_i^* \theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \cdots \quad (2.1.6.2)$$

where

$$\mu_i \equiv \mathbb{E}(Y_i) \text{ and } Y_i \sim \text{some exponential family distribution.}$$

Y_i is a response variable, $\boldsymbol{X}_i^*$is a row of the model matrix for any strictly parametric model components, $\boldsymbol{\theta}$ is the corresponding parameter vector, and the f_j are smooth functions of the covariates, x_k. This model allows for relatively flexible specification of the dependence of the response term on the covariates, using only 'smooth functions' to specify the model, instead of detailed parametric relationships. Any flexibility and convenience comes at the cost of two new

theoretical problems. It is necessary to represent both the smooth functions and to choose their level of smoothness. [34]

GAMs can be represented using penalized regression splines and estimated by penalized regression methods, while using cross validation to estimate the appropriate degree of smoothness for f_j.

Consider a model containing one smooth function and one covariate;

$$y_i = f(x_i) + \varepsilon_i \qquad (2.1.6.3)$$

where y_i is a response variable, x_i a covariate, f a smooth function and the ε_i are i.i.d. $N(0, \sigma^2)$ random variables.

To estimate f, this above equation (2.1.6.3) needs to be represented as a linear model [34], which is done by choosing a basis i.e. defining the space of functions which f (or a close approximation to it) is an element. If $b_j(x)$ is the j^{th} such basis function, then f is expected to have a representation

$$f(x) = \sum_{j=1}^{q} b_j(x)\beta_j, \qquad (2.1.6.4)$$

for some values of the unknown parameters, β_j. Substituting 2.1.6.4 into 2.1.6.3 yields a linear model. For a model containing a smooth function a smoothing parameter must be chosen. The smoothing parameter determines how much of the data is used to fit a model. It follows that this smoothing parameter, λ needs to be chosen often by cross validation. If λ is too high then the data will be over-smoothed, and if it is too low then the data will be under smoothed: in both cases the result is that the spline estimate $\hat{f}$ will not be as close to the true function f as it can be. Ideally we want to choose λ so that $\hat{f}$ is as close as possible to f, a suitable criterion may be to choose λ to minimise

$$M = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i - f_i)^2 \qquad (2.1.6.5)$$

M can be found from the following equation, according to [34];

$$v_g = \frac{n\sum_{i=1}^{n}(y_i - \hat{f}_i)^2}{[tr(\boldsymbol{I}-\boldsymbol{A})]^2} \qquad (2.1.6.6)$$

Where $\hat{f}$ is the estimate from fitting all the data, and $\boldsymbol{A}$ is the corresponding influence matrix (a matrix which yields the fitted value vector when post multiplied by the data vector). This above equation is known as the generalised cross validation score (GCV). This score is known to be computationally efficient and more crucially in this case it can also be shown to minimise $\mathbb{E}(M)$ in the large sample limit. This is of importance since, $\hat{f}^{[-i]} \approx \hat{f}$ with equality in the large sample limit, so $\mathbb{E}(v_g) \approx \mathbb{E}(M) + \sigma^2$ also with equality in large sample limit. Hence choosing λ in order to minimise v_g is a reasonable approach when the ideal is to minimise M. This is all explained much more thoroughly in [34].

A final note of GCV – if GCV is in place, the model needs to be fitted once with the full data for each value of the smoothing parameter, λ. Plots of the sequence number for different values of λ versus GCV can be used to determine an optimal value of λ. It should also be noted that the presence of correlation between errors can cause automatic smoothing selection methods such as GCV to break down.

2.2 Site-by-Site Exploratory Data Analysis

This section discusses the different trends, features and patterns of the NO_2 monitoring site data for all sites in Aberdeen between 2006 and 2014. There are 5 sites being investigated for time series analysis. These sites all belong to the AURN (Automatic Urban Road Network).

Site	**% of days with no data available**
Anderson Drive	8
Errol Place	11
Union St	8
King St	5
Wellington Road	11

Table 2.2.1: Percentage of days with no data available, by site

Also in this section any characteristics of the data are highlighted which might pose an issue when modelling the data. One such issue with data can be the proportion missing. This is evident

when one looks at a particular month where there are no data available. The previous table shows the percentage of missing days with no data available for each site.

Missing Data

Environmental data does not usually come with all data in place, with sometimes seemingly large sections of a data frame missing. As the data we are using comes from automatic network sites, it is natural that there are perhaps proportions of the data missing. This missingness could be due to any number of reasons, including malfunctioning instruments, incorrect calibrations, communication failure across the network monitoring system and in some cases, the locations have become disused for monitoring purposes. Given that there is some missing data across a period of time, the representativeness of the data could be questionable when inference is being made on a model. In reality, this proportion of missing data is relatively small i.e. approximately <11% for each of the sites. The following plot shows the amount of missing data for Anderson Drive. This plot is representative of the amount of missing data at the other sites, relatively little. The white gaps are where 100% of the data for that period is missing. Where there is possible an issue of the site not being functional, this is represented by a white gap over longer period of time. For all of the sites, there is at least some missing data. Quantifying the plots, the x-axis shows the time period from 2006 to 2015, while the y-axis shows the percent of the observed data available for that day. These are in units of days.

From looking at the following plot, it is fair to conclude that there is very little missing data, and there does not appear to be a pattern in those data which are missing. The plot itslf is reflective of the amount of data missing at other sites.

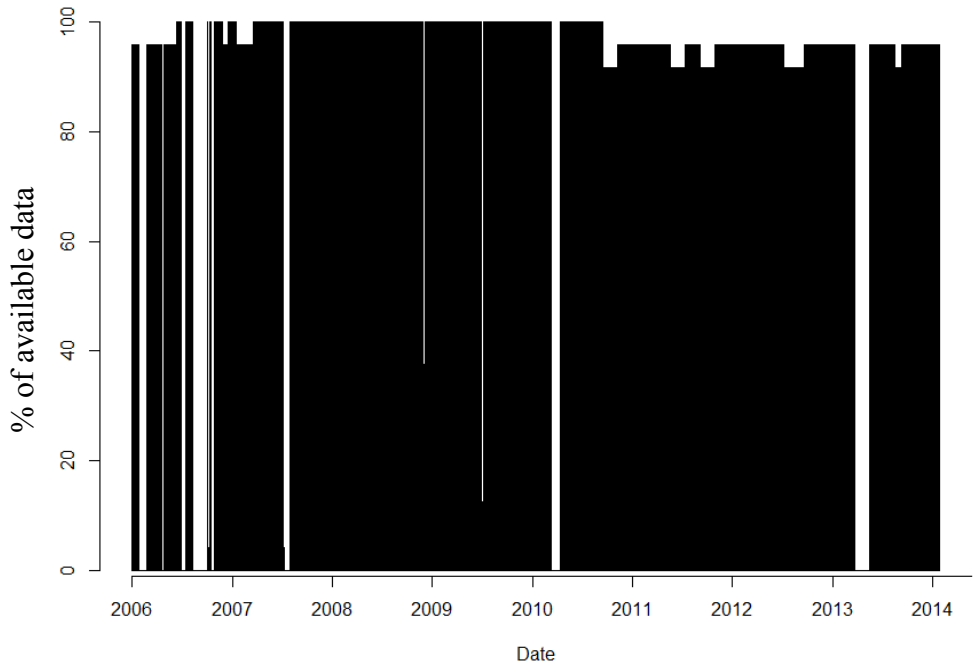


Figure 2.2.1: Bar plot of proportion of observed data at Anderson Drive

Outliers

Outliers may occur in the data, and according to [34] [35], an outlier is an observation point that is distant from other observations. An outlier is usually due to one of two causes; one of these causes is an error in the experiment and the other cause is variability in the data [34]. Outliers can be dealt with by either removing them from the dataset if they are due to experimental error, or by taking a log transformation, which shows the stabalisation in the variance in the distribution of the data. This stabalisation of the variance can be seen in the following histograms; figures 2.3.1 and 2.3.2 respectively. This was the case for the NO_2 datum recorded at the AURN sites. This is discussed further in the next section.

2.3 Graphical summaries of Nitrogen Dioxide Monitoring Site Data

To gain an impression of how the NO_2 data behave over time, the data are plotted against time to give an insight into the overall trend of the data and obtain a subjective comparison between each of the sites and across the years. There appears to be a non-constant variance issue for the time series corresponding to each site, with most of the values clustered around low NO_2 concentrations, with some seasonal effects present at certain sites. This non – constant variance

issue was addressed by applying different transformations to each site such as a log transformation, a square root transformation and an exponential transformation. The log transformation adequately addressed this issue by distributing the distribution in a fashion which more closely resembles a normal distribution. This can be seen by a comparison of the NO_2 data recorded at the different AURN sites before and after a logarithmic transformation was taken, which is shown in Figures 2.3.3 and 2.3.4 respectively.

As mentioned previously regarding outliers, when these log transformations were applied, it was the case for most extreme values that they became integrated into the main body of the distribution.

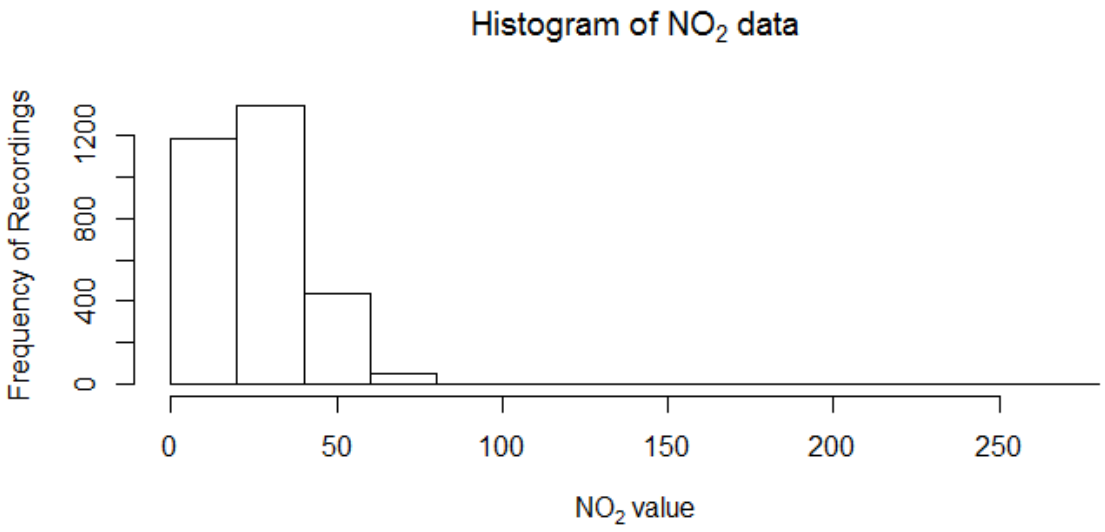


Figure 2.3.1: Histogram depicting the untransformed NO_2 data recorded at Anderson Drive over the years 2006 – 2014. It can be seen from this histogram that the data are heavily skewed.

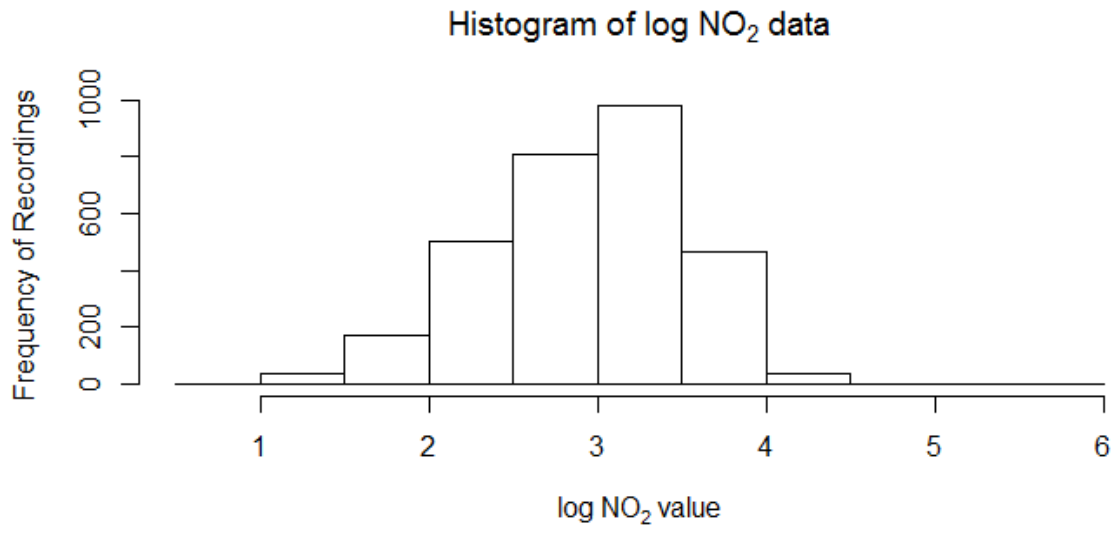
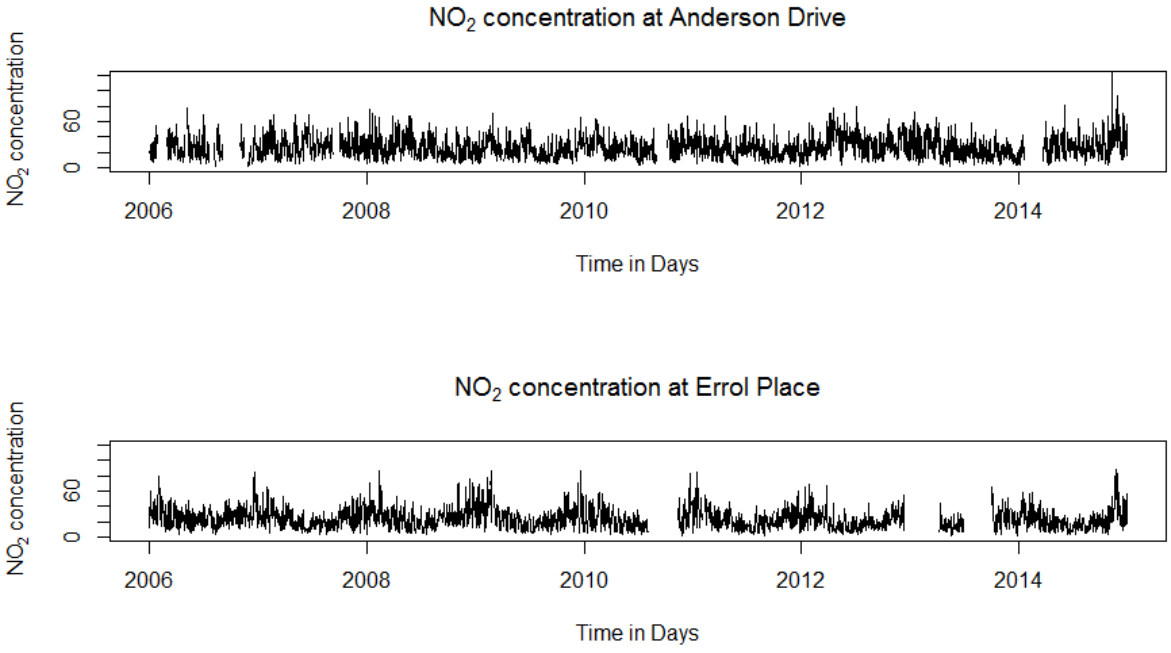


Figure 2.3.2: Histogram showing the log transformed NO_2 data at Anderson Drive for the years 2006 – 2014. The histogram shows, from a comparison with Figure 2.3.1, that a log transformation of the data stabalises the variance.



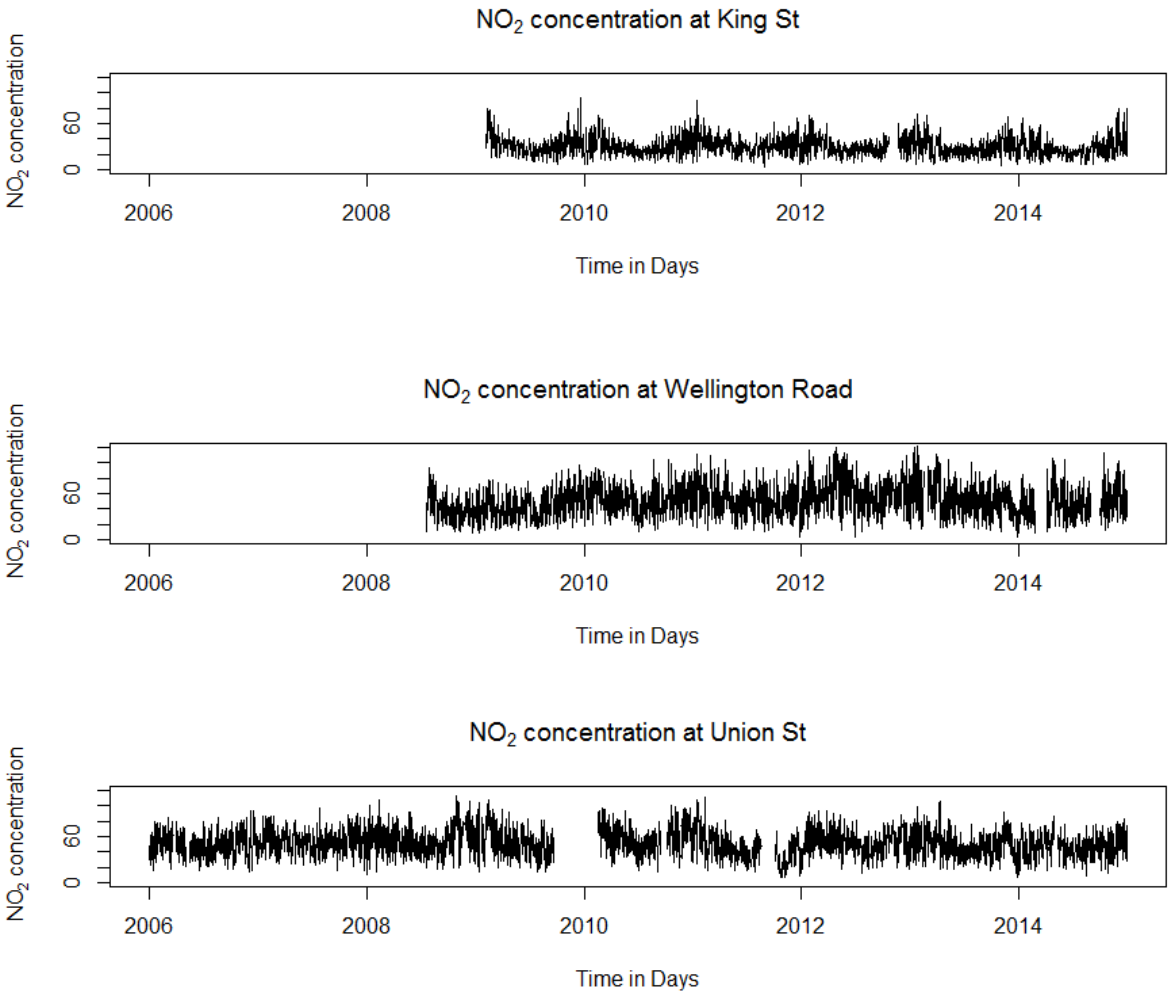


Figure 2.3.3: Time Series of NO_2 for each site between 2006 and 2014 (no transformation)

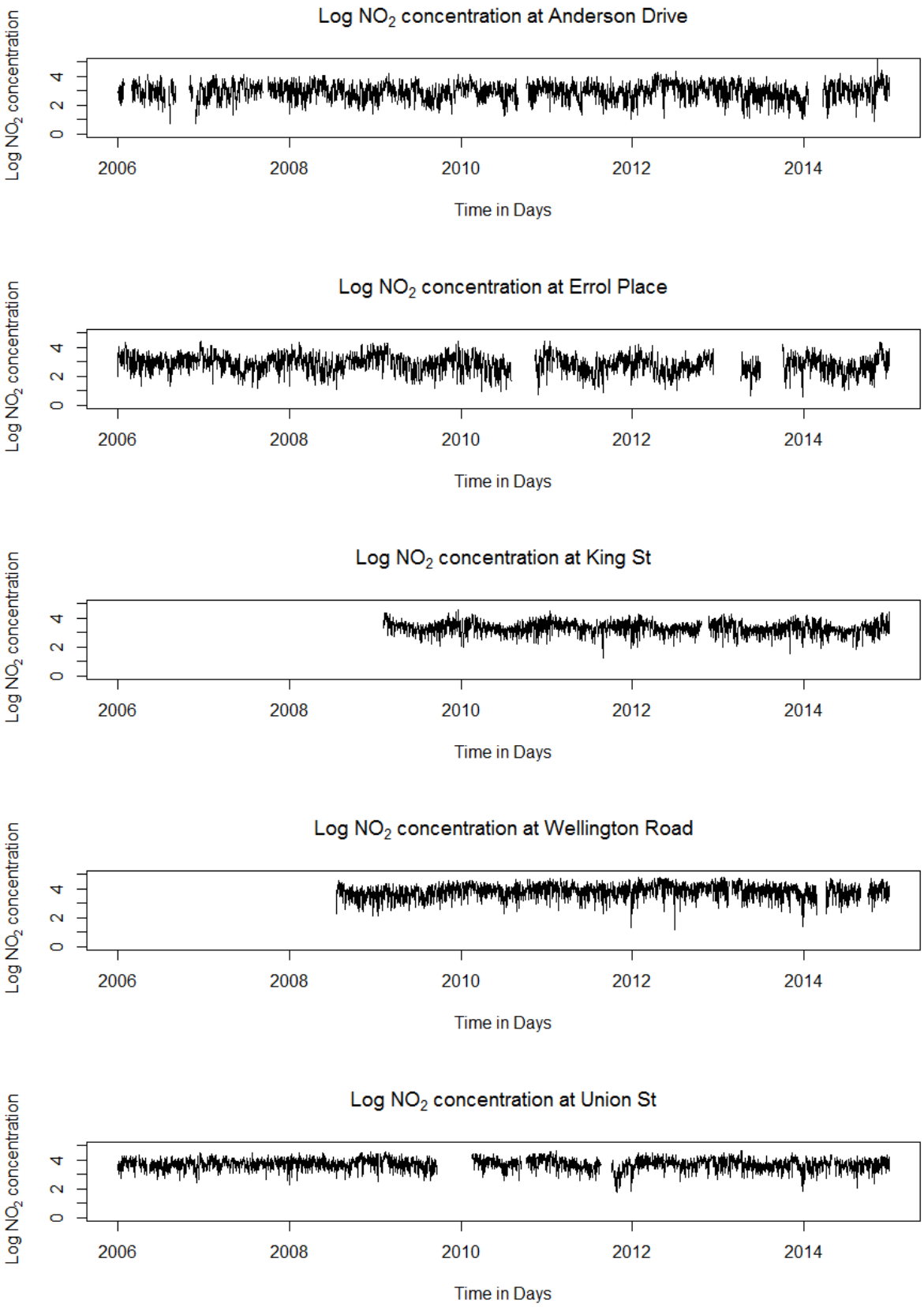
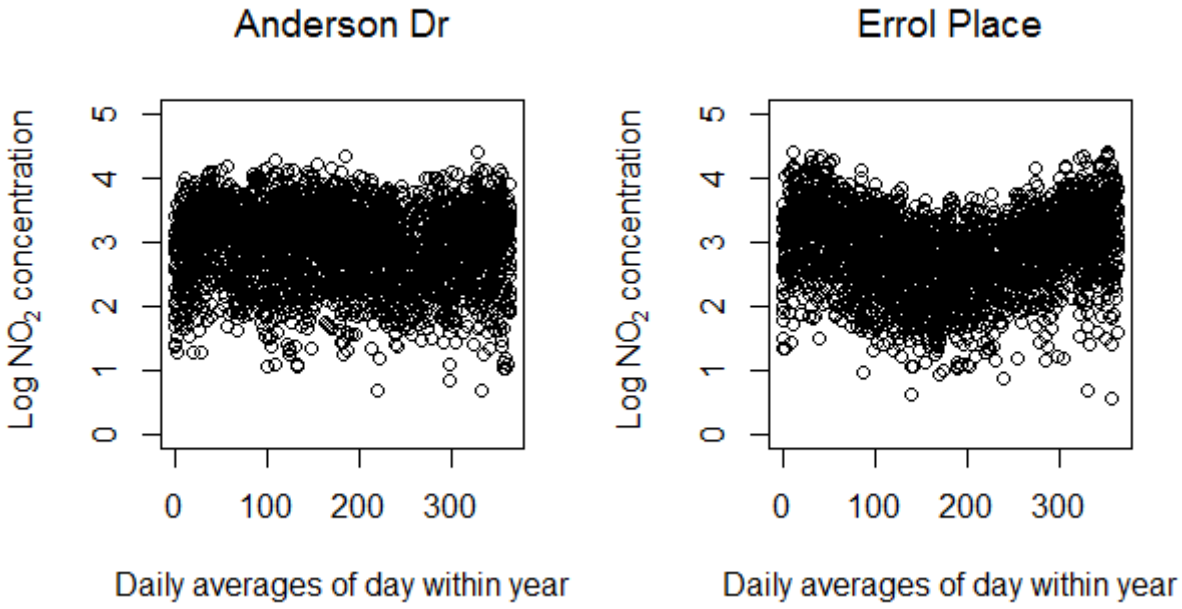


Figure 2.3.4: Time series of log transformed NO_2 for each site between 2006 and 2014

Overall, looking at both Figures 2.3.3 and 2.3.4, depicting both the transformed and untransformed time series data, NO_2 would appear to follow a wave like seasonality with the peaks and dips of each site differing slightly. The wave-like sinusoidal seasonality could be due to weekly or daily variations in log NO_2 concentrations or it could be linked to a covariate effect.

It is probable that sites which are closer to one another would be more correlated than others which are further away. This will be explored later in the Chapter using spatial statistics.

It is also of interest to look at how the data behaves over the course of a year. This is done by numbering the days of the year 1- 366 and plotting them against their respective log NO_2 value. This is done at each site and the plots are as follows in Figure 2.3.5. It can be said, that at Anderson Drive and Wellington Road, there is no upwards or downwards trend, while at Errol Place, King St and Union St there is a quadratic shape running through the data – more so at Errol Place and King St than Union St. This could be due to less traffic on these particular roads in the summer months and this is evidence for seasonal patterns.



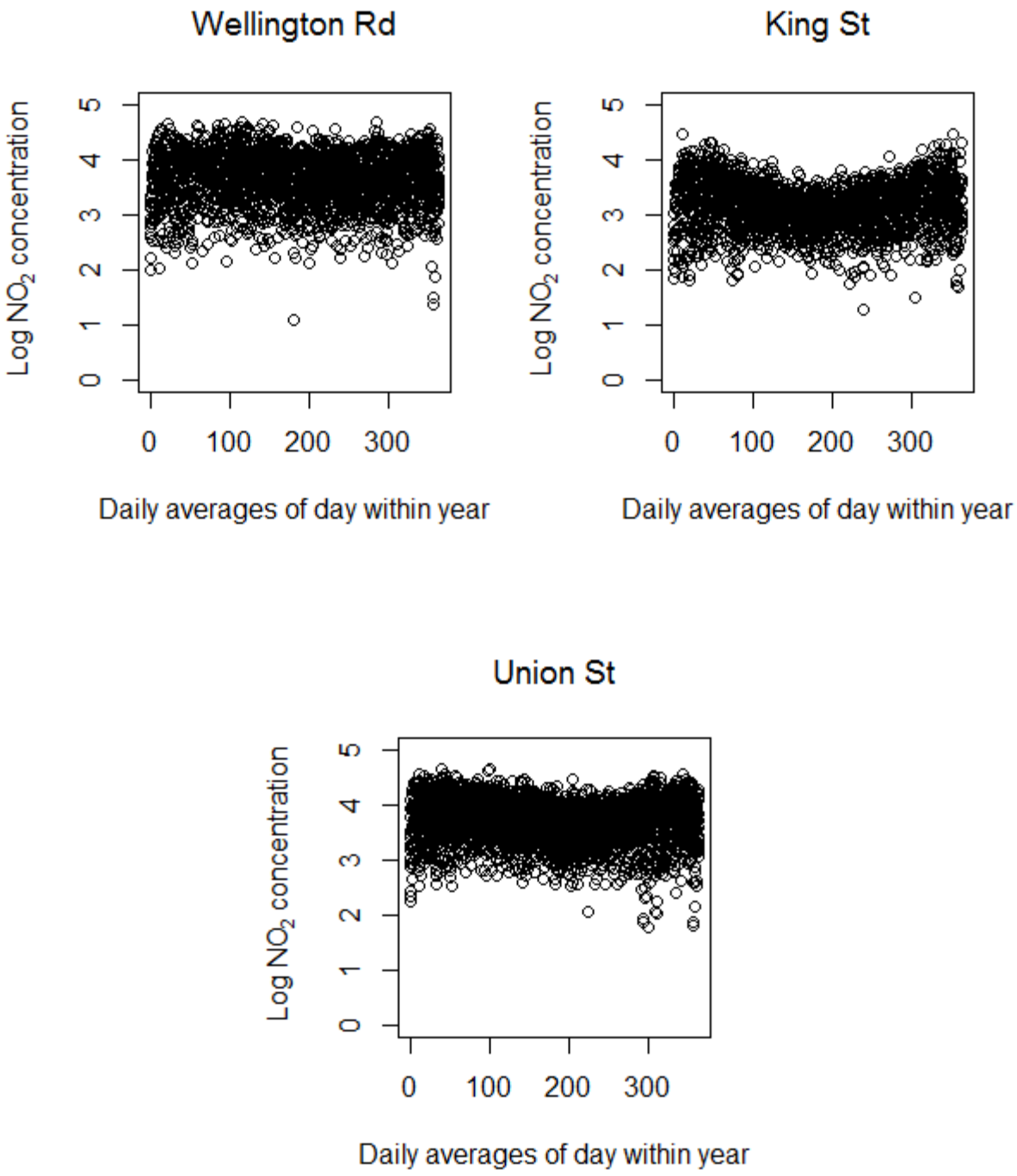
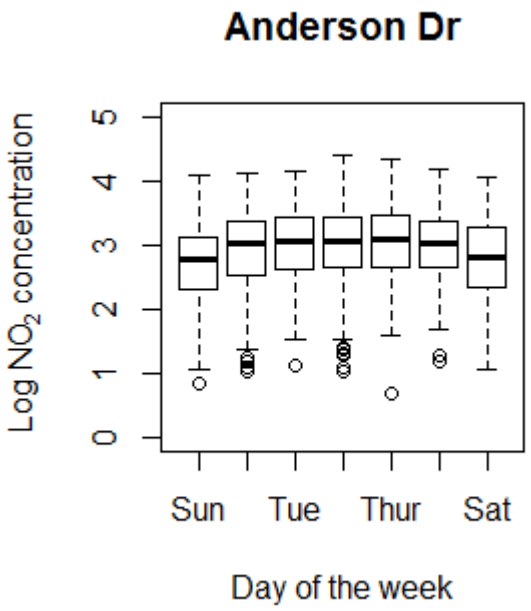
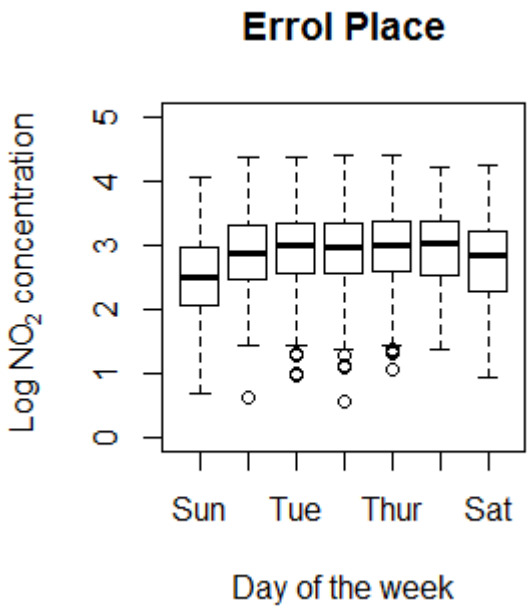
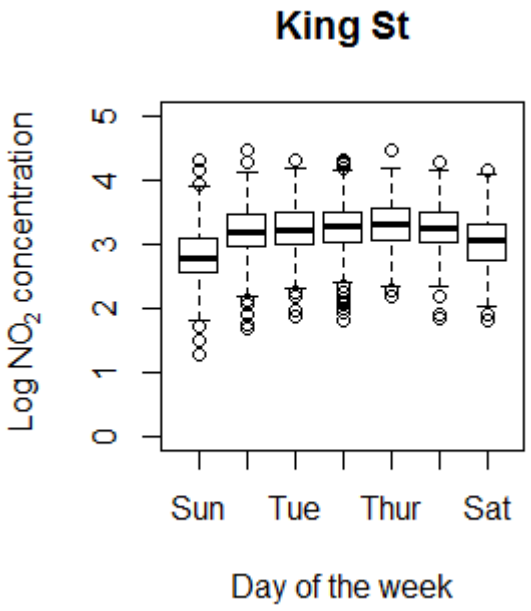


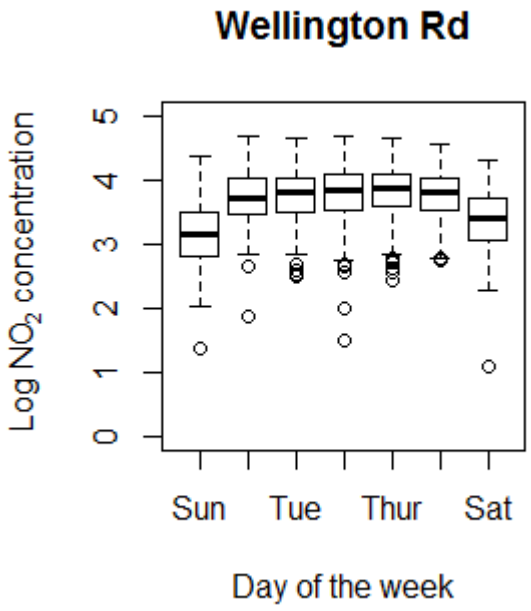
Figure 2.3.5: Daily averages recorded at each site for the years 2006 – 2014.

Looking at the log values of NO_2 over each day of the week is also of interest. Each day is taken individually and a boxplot is created from the log NO_2 values. This is done at each AURN site;









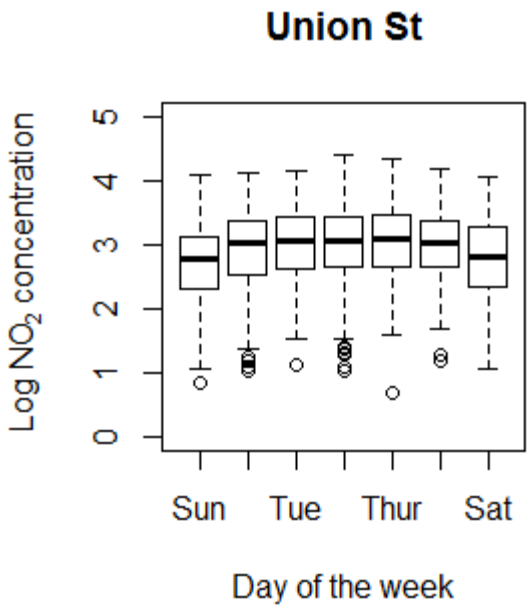
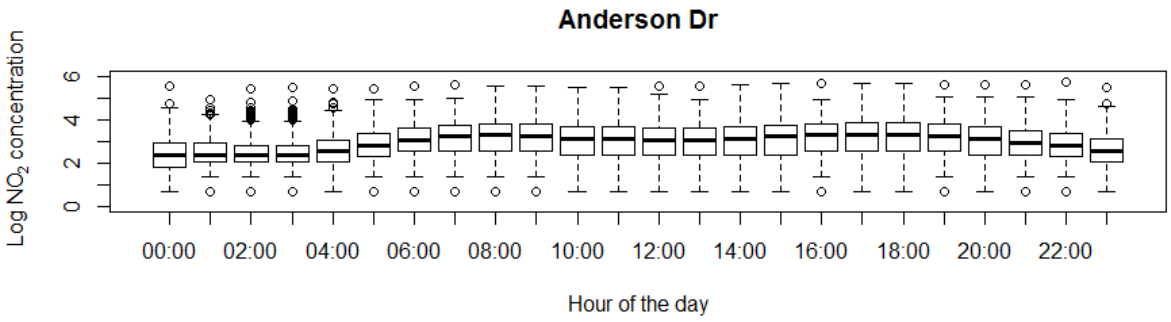


Figure 2.3.6: Daily averages recorded during 2006 – 2014, presented as the daily averages for each day of the week for each site.

It can be seen from these plots in Figure 2.3.6 that for every site, Sunday has a lower median concentration than any other day of the week, and the same can be said for Saturday at every site - except for King Street, which has a median concentration similar to those of the weekdays (Monday – Friday that is). This is indicative of less traffic being on the roads during the weekend, especially on Sundays. There are also outliers shown in each plot, they tend to be low rather than high.

The following Figures show the log NO_2 concentrations at each of the AURN sites, this time focussing on the hour of the day. Doing this can show when, during the day, there are higher concentrations of NO_2, if there are any.



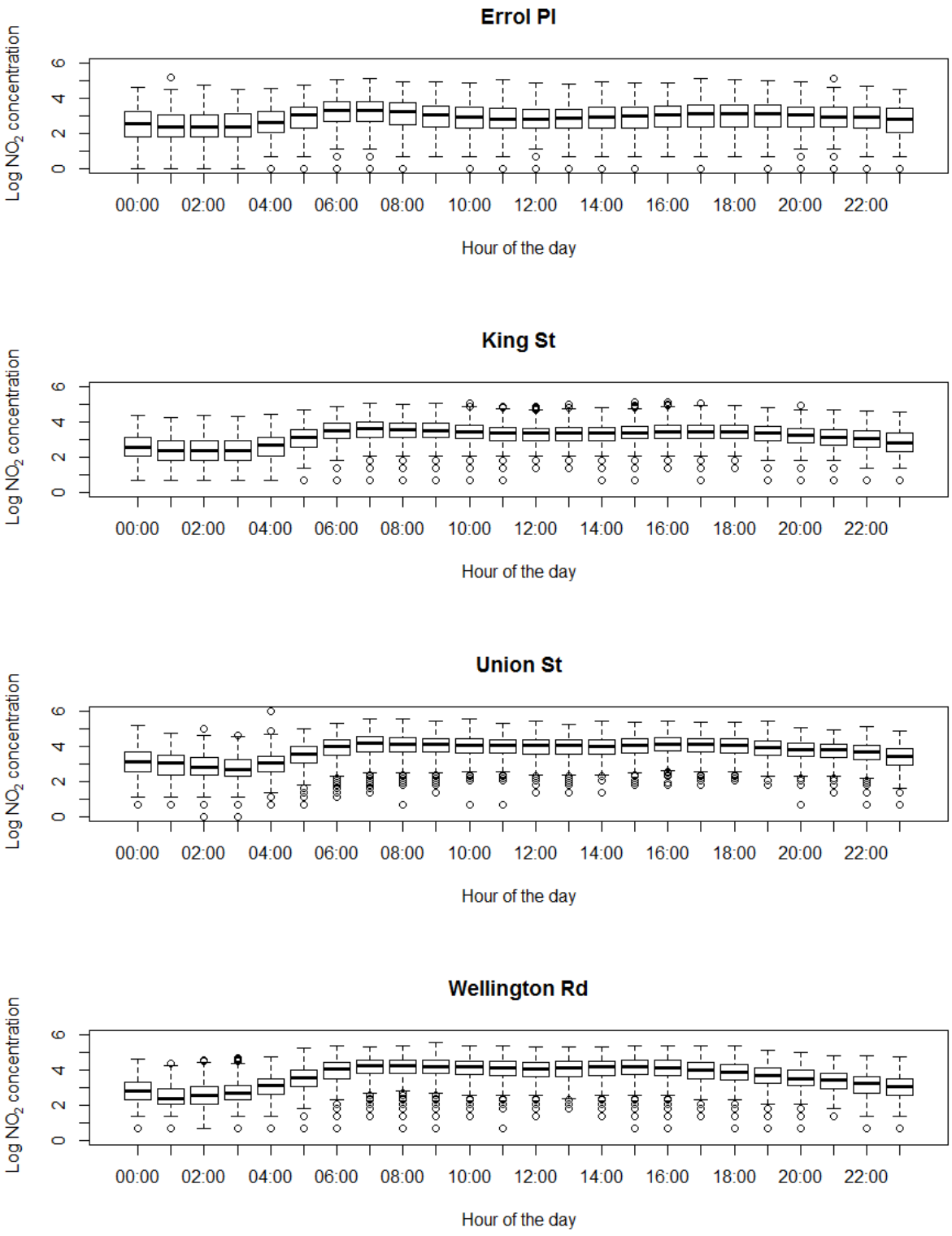


Figure 2.3.7: Boxplots showing the daily average NO_2 concentrations for each of the different sites for the years spanning 2006 – 2014.

As can be seen from all of the plots in Figure 2.3.7, log NO_2 concentrations are higher during the day than they are at night, particularly between the hours of 07:00 and 20:00 for most sites. There are peaks at rush hour times (08:00 - 09:00 and 17:00 – 18:00) for most sites.

2.4 Graphical and Numerical Summaries of Meteorological Data

The meteorological data which are described in section 1.3.2 consist of daily mean values of temperature, humidity, pressure at mean seas level, total rainfall, cloud cover, wind speed, and wind direction at one site in Aberdeen, namely the airport at Dyce, Aberdeen. The table below summarises each of the potential covariates with a number of summary statistics. The summary statistic values are recorded for the time period 2006 – 2014.

Variable	**Min**	**Q1**	**Median**	**Mean**	**Q3**	**Max**	**St. Dev**
Wind Speed (kmh^{-1})	0.34	2.98	4.18	4.50	5.72	13.93	1.99
Wind Direction	28.0	168	208	211	254	351	62
Cloud Cover (oktas)	0	4	6	5	7	8	2
Rainfall (mm)	0.00	0.00	0.01	0.10	0.10	2.56	0.20
Temperature (°C)	-10.4	5.2	8.5	8.7	12.4	22.7	4.7
Relative Humidity (RH)	34	73	81	80	88	100	11
Pressure at MSL (Pa)	955	1002	1011	1010	1019	1043	13

Table 2.4.1: Summary Statistics for Meteorological factors

The plots of the meteorological factors can be seen in Figures 2.4.1 and 2.4.2 below. These explore the individual trends of each of the meteorological factors. Temperature, Humidity and pressure all seem to follow a yearly sinusoidal pattern, while a wind speed, wind direction, cloud cover and rainfall seem to follow a less distinguishable distribution i.e. the points on the plot are more randomly scattered. Rainfall should go through a transformation, possibly log, as the time series suggests the data are clustered around extremely low values.

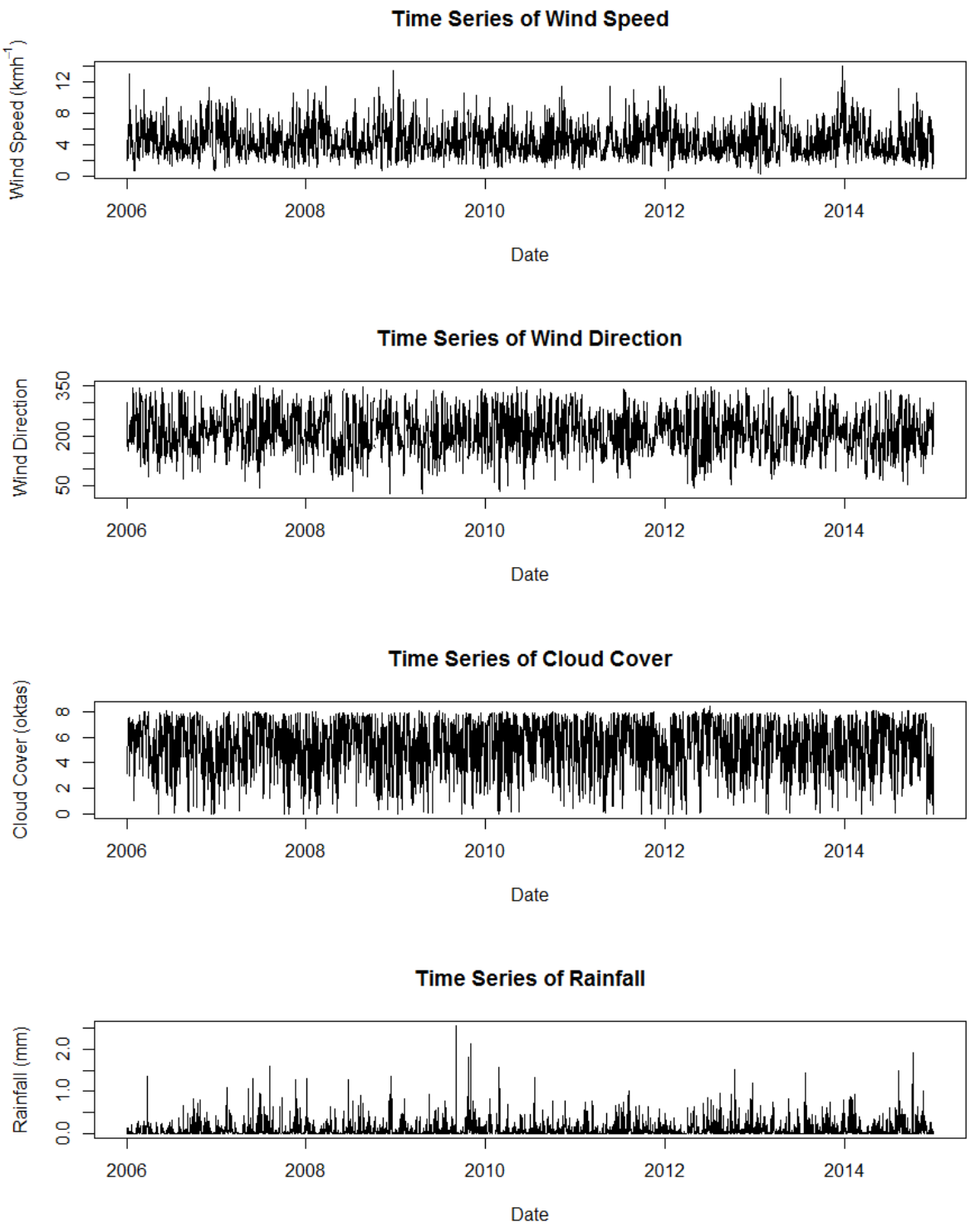


Figure 2.4.1: Time series of meteorological factors for 2006 – 2014 at Dyce, Aberdeen. The four panels correspond to the following variables; wind speed; wind direction; cloud cover and rainfall respectively.

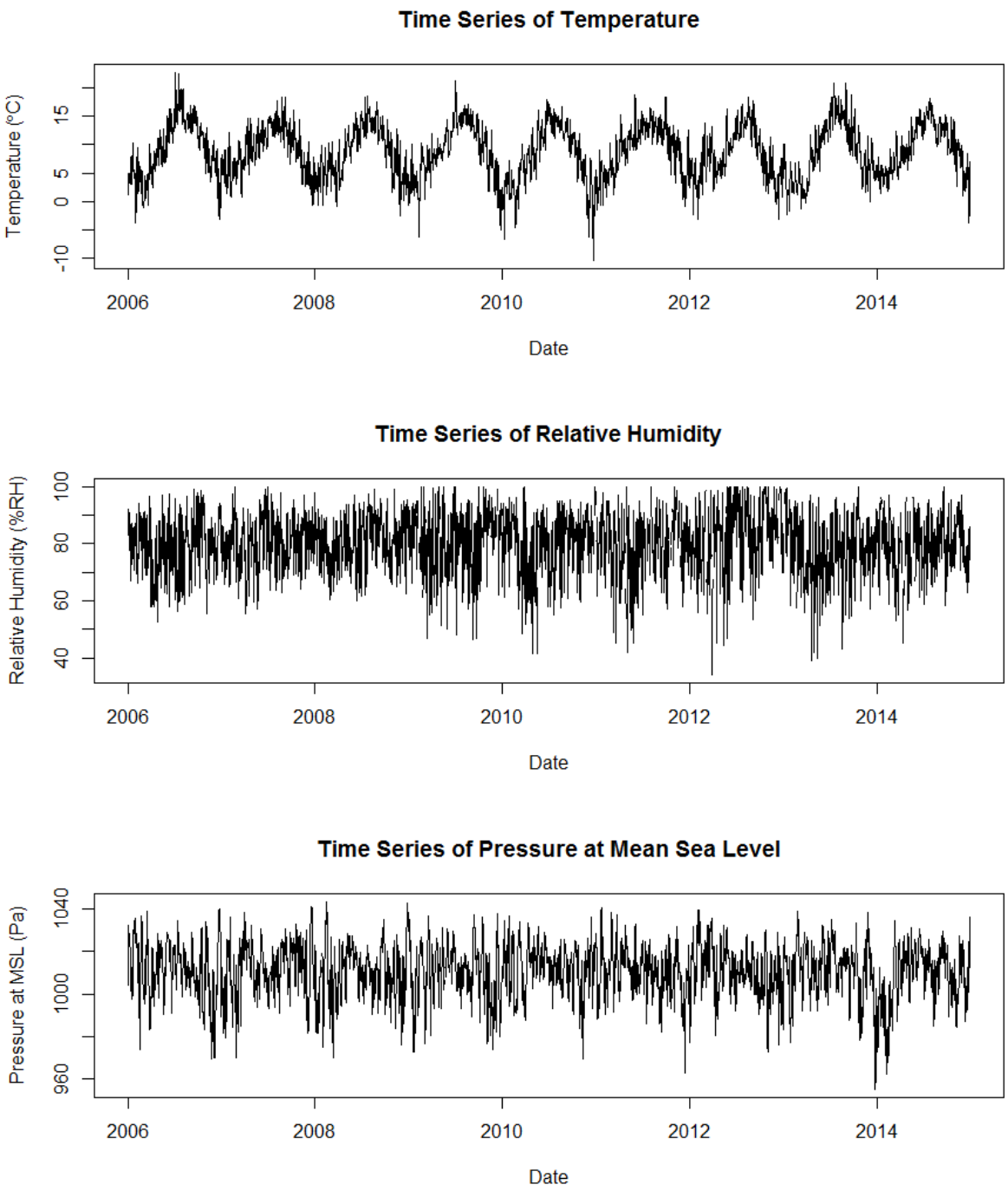


Figure 2.4.2: Time series of meteorological factors for 2006 – 2014 at Dyce, Aberdeen. The three panels correspond to the following variables; temperature; relative humidity and pressure at mean sea level respectively.

2.5 Graphical and Numerical Summaries of Traffic Data

The traffic data as described in section 1.3.3 consist of daily mean values of all motor vehicles (which is the total number of motor vehicles, not including LGVs, HGVs and Buses and coaches), light goods vehicles, all HGVs (heavy goods vehicles), and buses and coaches, at 5 sites in Aberdeen. The table below (Table 2.5.1) summarises each of the potential explanatory variables with a number of summary statistics. The median and mean values for the different vehicles differ between vehicle class within each site as well as between sites. The highest number of vehicles is recorded at Wellington Road, with 15,489 total motor vehicles (this is not including LGVs, HGVs, or buses and coaches). The highest standard deviation for all motor vehicles (not including the other aforementioned classes) also occurs at Wellington Road, with a standard deviation of 1656.

The plots of the different vehicle classes at Union Street over time in the following Figures explore the individual trends. From the plots it is clear that there are different levels of vehicle counts on weekdays and non-weekdays. There is also a consistent quadratic shape from year to year for all of the potential variables. Although not seen here in full, as only Union Street is shown, the explanatory variable all motor vehicles have been consistent for all sites except for Errol Place and Union Street which decrease significantly after 2007 and 2011 respectively. Again, although not shown here, all sites have seen an exponential trend for the number of buses and coaches except for Union Street and Anderson Drive, which show a decrease after 2011, and a completely different pattern due to the relatively low numbers, respectively. The following plots are useful in that they show the layout of the data created from the tables in section 1.3.3, as discussed previously. This is the pattern of the data which are created using the disaggregated approach described in the same section. The data in the following plots reflect the man-made approach to creating the data, as opposed to collecting the data naturally from the respective sites.

	Min	Q1	Median	Mean	Q3	Max	St. Dev
Errol Place							
All Motor Vehicles	3478	4715	5394	5420	5773	8353	1029
Light Goods Vehicles	463	719	831	932	1019	1753	313
All HGVs	414	578	655	650	708	919	103
Buses and Coaches	138	217	284	296	349	567	96
Anderson Drive							
All Motor Vehicles	4761	6405	7399	7204	7964	9194	967
Light Goods Vehicles	647	945	1077	1067	1179	1470	164
All HGVs	310	414	482	468	516	586	62
Buses and Coaches	7	11	12	12	14	17	2
Wellington Rd							
All Motor Vehicles	8916	11435	13376	12888	14195	15489	1656
Light Goods Vehicles	1718	2444	2761	2730	3024	3646	400
All HGVs	1112	1661	1859	1925	2272	2799	390
Buses and Coaches	57	87	98	103	114	174	25
King Street							
All Motor Vehicles	3698	4795	5601	5402	5968	6565	698
Light Goods Vehicles	590	791	914	892	991	1129	120
All HGVs	301	420	477	477	525	646	77
Buses and Coaches	111	165	192	202	234	340	52
Union Street							
All Motor Vehicles	1042	1533	1714	1720	1958	2226	267
Light Goods Vehicles	144	230	267	272	316	420	58
All HGVs	24	43	72	67	87	100	22
Buses and Coaches	126	185	213	211	233	294	24

Table 2.5.1: Summary statistics for traffic variables at 5 different sites in Aberdeen. Vehicle units are vehicles per day or vehicles day^{-1}

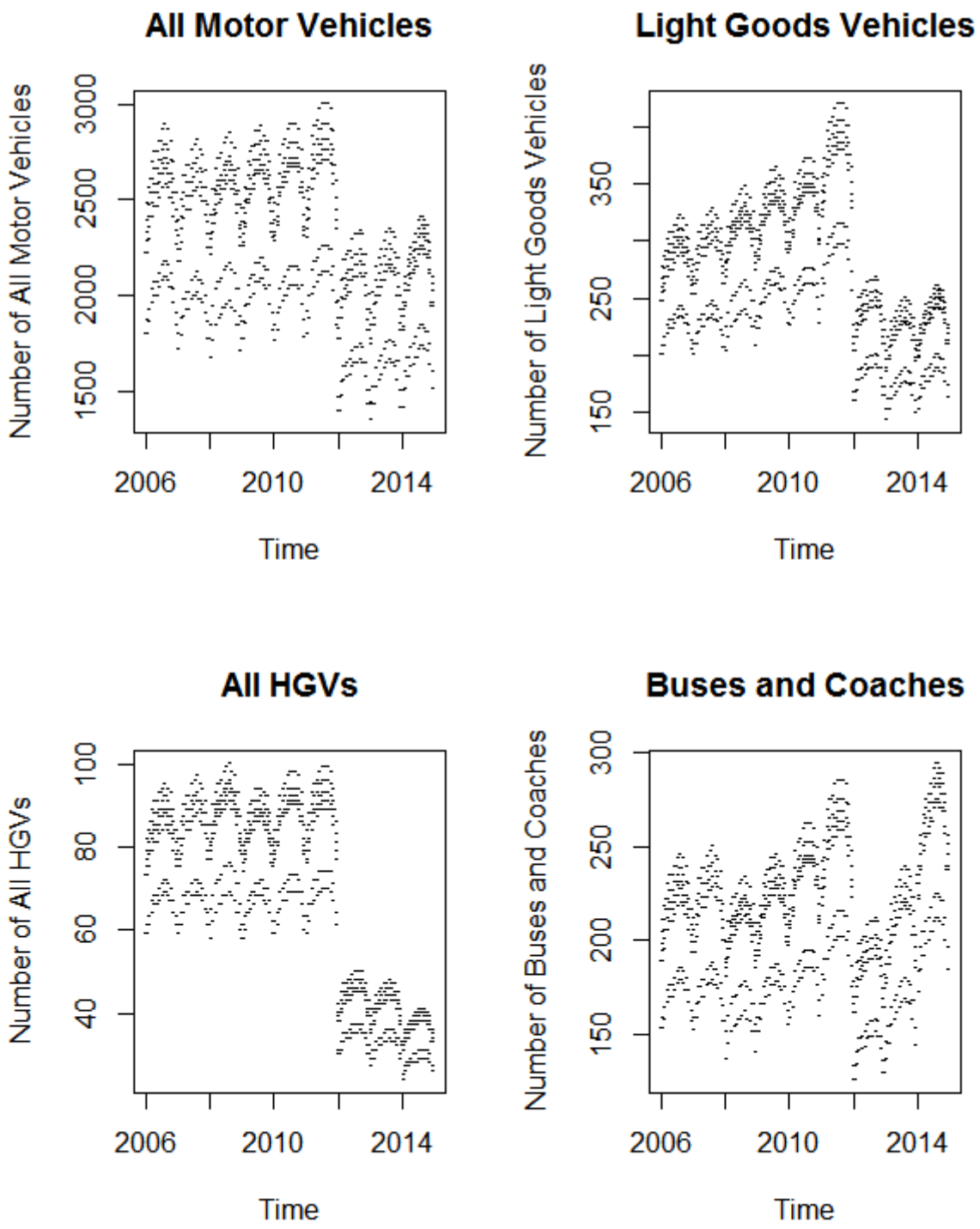


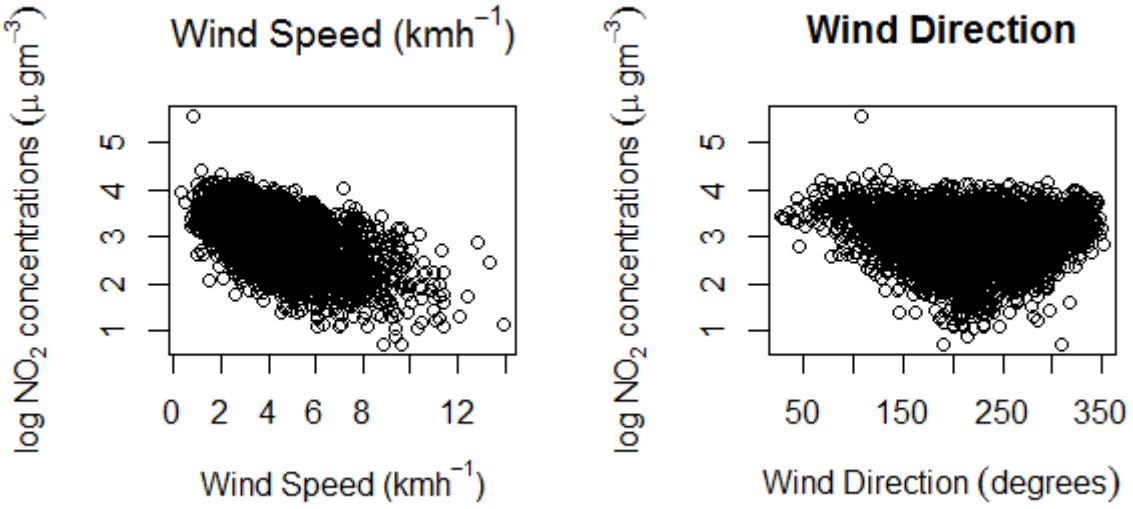
Figure 2.5.1: Time series plots of traffic variables at Union Street between 2006 and 2014. The top left panel shows all motor vehicles, the top right panel shows the Light Goods Vehicles, the bottom left hand panel represents All HGVs, and the bottom right panel shows buses and coaches.

The abrupt change in each of the vehicle class levels at 2012 can be estimated, given the public data available, to be due to a change in the counting method for the number of vehicles before and after 2012. Although these plots show the relationship between different vehicle classes and time, they do not show how the vehicle classes are related to the variable of interest, NO_2. This is looked at in the following section.

2.6 Relationships between NO_2 and potential explanatory variables

NO_2 and the vehicle classes mentioned have a relationship which can be seen from the scatterplots in this section. These relationships are also reflective of the NO_2 relationship with the vehicle classes which are found at other sites such as Anderson Drive, King Street and so on. The scatterplots don't show any relationship between NO_2 and the vehicle class variables, apart from perhaps the HGVs which is in two clusters, although this is more a comment on the HGV class - particularly at Union Street - than it is on the relationship between HGVs and NO_2. i.e. the HGV class is in two clusters as that is how the data happened to be recorded. Also in this section, the relationships between NO_2 and meteorological variables are looked at in scatterplots.

2.6.1 Meteorological covariates



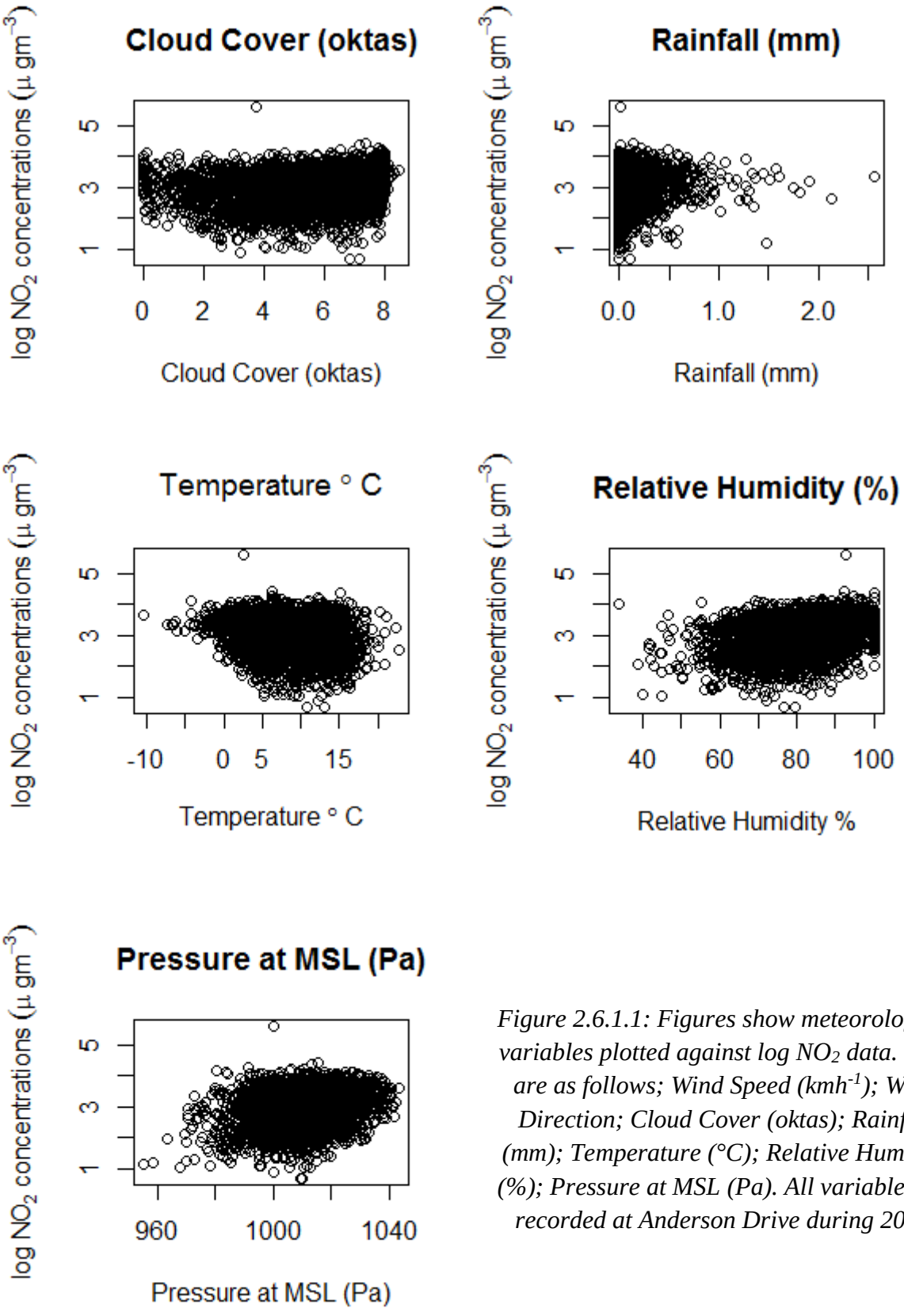


Figure 2.6.1.1: Figures show meteorological variables plotted against log NO_2 data. They are as follows; Wind Speed (kmh^{-1}); Wind Direction; Cloud Cover (oktas); Rainfall (mm); Temperature (°C); Relative Humidity (%); Pressure at MSL (Pa). All variables are recorded at Anderson Drive during 2014.

Looking at Figure 2.6.1.1 the relationships between log NO_2 and the meteorological factors are shown as a visual aide. As wind speed increases, log NO_2 concentration decreases to the first plot, and the relationship can be described as linear. The second plot shows that log NO_2 concentration does not seem to have a clear relationship with wind direction. The change in the concentrations of NO_2 can be put down to more observations being recorded between 100 and 300 degrees, than there are recorded between 0 and 100 degrees, and 300 and 360 degrees.

In the top left plot on the previous page, cloud cover does not seem to have a significant relationship with log NO_2 since as cloud cover increases there does not seem to be any visible change in log NO_2 values. There are lower values of log NO_2 as at cloud cover levels higher than 2 oktas, although this could be due to the fact there are more observations at cloud cover levels higher than 2 oktas. The plot on the top right of the previous page, rainfall and log NO_2 does not have a visible relationship, although rainfall is notoriously difficult to model as there are many low values i.e. days when there were no rain and even with a log transformation the relationship between rainfall and the response is difficult to interpret in comparison to say, temperature and log NO_2. This can be seen from the large amount of low values and very few values of rainfall above 0.5mm.

The third plot on the previous page (depicting temperature) shows as temperature increases there does not seem to be a visible change in log NO_2 concentrations. The scatterplot showing relative humidity shows that as relative humidity increases so too does the log NO_2 concentration. It looks like the relationship could be linear. The final plot in figure 2.6.1.1 shows the relationship between log NO_2 and pressure at mean sea level. This relationship could be described as linear and from the plot, it can be concluded that as pressure increases so too does log NO_2.

It is also of interest to explore collinearity between the explanatory variables. Collinearity occurs when two or more variables in the data set are highly correlated with one another. In order to have the best model possible, it is hopeful that the variables have relatively little collinearity. A pairs plot is seen below of the meteorological variables, and their collinearity (or lack of) can be seen from this in Figure 2.6.1.2.

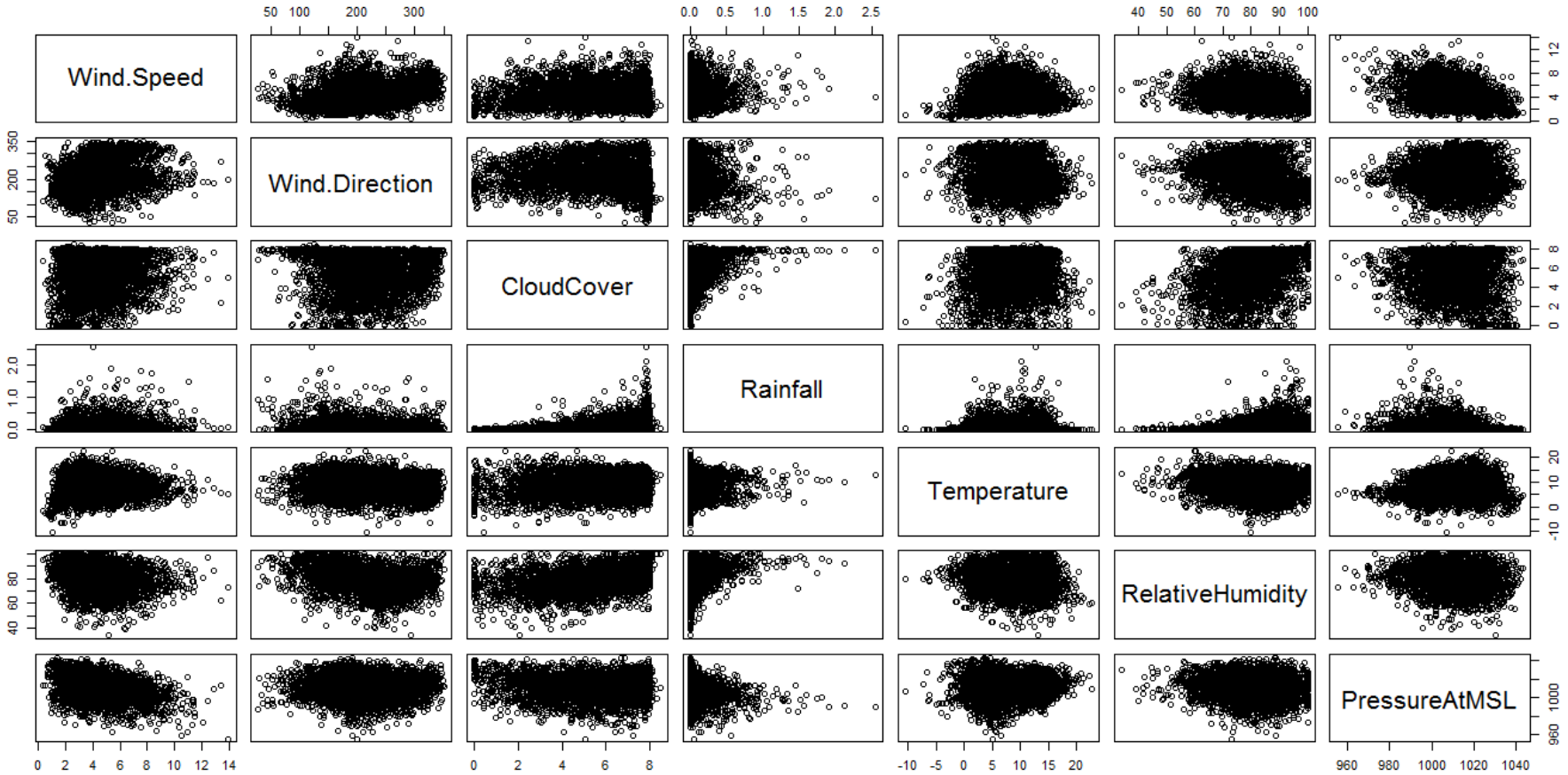


Figure 2.6.1.2: Pairs plot of meteorological variables with labels. It can be seen from these plots that there is not much, if any, collinearity present between the meteorological variables.

2.6.2 Traffic covariates

The following Figures (Figure 2.6.2.1 and Figure 2.6.2.2) shows the collinearity between the traffic explanatory variables, and the relationship between log NO_2 concentration and each of the vehicle classes at Anderson Drive respectively. This is similar to the previous section regarding meteorological variables. From looking at Figure 2.6.2.1 it is clear that there is multicollinearity present for the traffic variables at Anderson Drive and this must be dealt with in an appropriate manner by only including some of the variables. This is done to reduce the amount of noise in the model when it is being built.

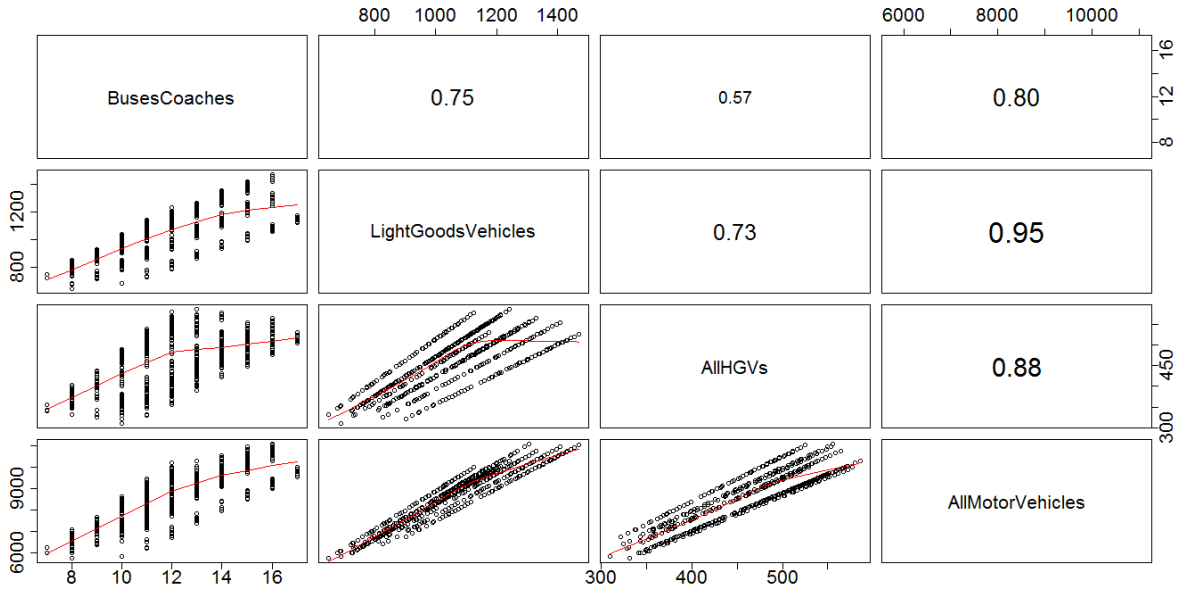


Figure 2.6.2.1: Pairs plot showing the collinearity between the traffic explanatory variables at Anderson Drive for the years 2006 – 2014. The bottom left-hand panels show the visual representation of the data and the top right-hand panels show the correlation between the two variables. The variables are named on the diagonal panels.

The following Figure (Figure 2.6.2.2) shows the traffic explanatory variables relationship with the response variable of interest – log NO_2 concentration. Each of these variables, from this initial look, can be said to have no linear relationship with log NO_2 concentration.

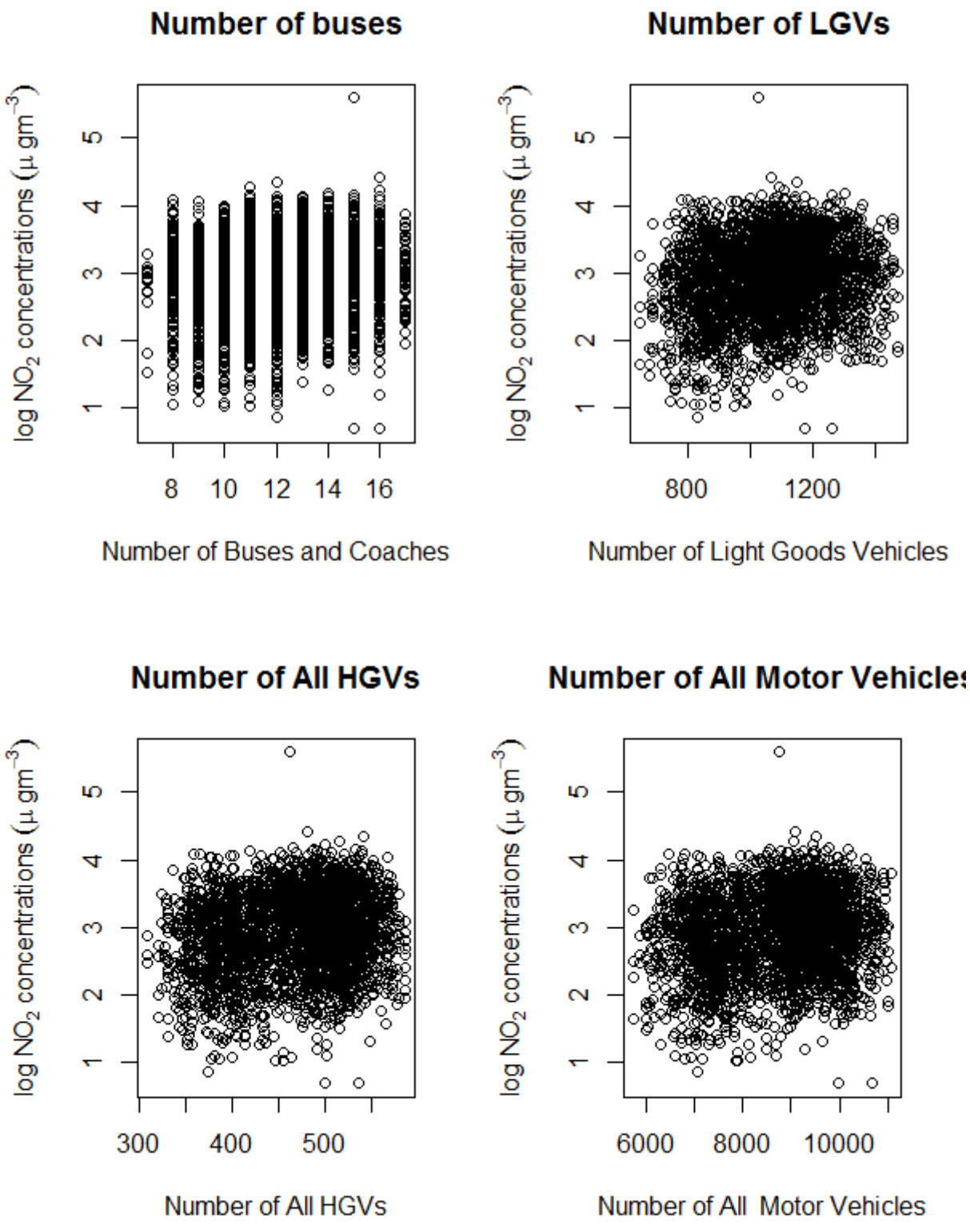


Figure 2.6.2.1: Log NO_2 concentrations vs the diffferent vehicle classes. The panels are labelled as their corresponding vehicle classes. This is recorded at Anderson Drive.

This section and the previous section have covered the exploratory analysis in mostly informal ways, in order to assess the relationship between NO_2 and traffic covariates across a temporal domain. The meteorological variables which will potentially be included in the final model have also been explored in a temporal context, and collinearity between variables has been discussed. Multicollinearity has also been explored and it has shown that some meteorological explanatory variables are correlated with one another. The same has been done for traffic factors which will potentially be included in the final model. Multicollinearity for traffic variable has shown that traffic variables are highly correlated with one another. The next section quantifies to a more accurate degree if and to what extent NO_2 is related to the time, meteorological and traffic factors using more effective and appropriate regression assumptions.

2.7 Exploring Trends and Seasonality using Linear Regression Modelling

This section determines the relationship between log NO_2 and the meteorological and traffic variables, using linear regression. This linear regression technique uses the daily means available for each of the covariates, from each of their respective sites.

The first step is to fit a linear model, assuming that the observations are uncorrelated. Following this a check is carried out of the assumptions and this process is repeated if there is evidence of correlation. The linear regression models are fit using OLS and it is assumed that the errors are uncorrelated and have mean zero.

As shown in the previous section, a sinusoidal pattern was present across the year with a possible weekly effect in log NO_2. In order to model this, harmonic regression is used, which is a type of regression. A harmonic function included regression terms for the pattern over the year as well as the pattern over the week, known as day within year (DWY) and day within week (DWW) respectively. These variables, as well as a continuous year variable, are coupled with the meteorological and traffic explanatory variables and different combinations are fit in order to gain an idea of what variables are related to the NO_2 and if this differed between sites. The models are described in Table 2.7.1. The model equations are then explained as follows:

$$y_t = \beta_0 + \beta_1(year)_t + \beta_2 \cos\left(\frac{2\pi(DWY)_t}{365}\right) + \beta_3 \sin\left(\frac{2\pi(DWY)_t}{365}\right)$$
$$+ \beta_4 \cos\left(\frac{2\pi(DWW)_t}{7}\right) + \beta_5 \sin\left(\frac{2\pi(DWW)_t}{7}\right) + \beta_6(Wind\ Speed)_t$$
$$+ \beta_7(Wind\ Direction)_t + \beta_8(Cloud\ Cover)_t + \beta_9(Rainfall)_t$$
$$+ \beta_{10}(Temperature)_t + \beta_{11}(Humidity)_t + \beta_{12}(Pressure)_t$$
$$+ \beta_{13}(Buses\ and\ Coaches)_t + \beta_{14}(Light\ Goods\ Vehicles)_t$$
$$+ \beta_{15}(All\ HGVs)_t + \beta_{16}(All\ Motor\ Vehicles)_t + \varepsilon_t$$

Equation 2.7.1: All factors included in a non-specific site model

where $y = \log NO_2$ concentration, and $t = 1, \ldots 3287$ since there are 3287 days spanning the years 2006 – 2014.

Statistically significant variables are selected to be in the model for each site, following a process. The process is of the following nature; the statistically insignificant variables are dropped with the most insignificant being dropped first and the least insignificant (while still being classified as statistically insignificant with a p-value < 0.0.5) being dropped from the model last. Different variables being dropped from different locations can be interpreted as them (the variables) not being statistically significant. They may not have an effect on NO_2 concentration, according to the models and tests carried out on the models. The fact there are different variables dropped from different locations means that there may some variance between sites for the same variables.

Model Site	Model Description	Variable(s) removed	n
Full Model at all sites	Year, DWY, DWW, Wind Speed, Wind Direction, Cloud Cover, Rainfall, Temperature, Humidity, Pressure, Buses and Coaches, Light Goods Vehicles, All HGVs, All Motor Vehicles	NA	3287
Anderson Drive	Year, Wind Speed, Wind Direction, Cloud Cover, Rainfall, Temperature, Humidity, Pressure, Buses and Coaches, Light Goods Vehicles, All HGVs, All Motor Vehicles	DWY, DWW	3287
Errol Place	Year, DWY, DWW, Wind Speed, Wind Direction, Cloud Cover, Rainfall, Humidity, Pressure, Buses and Coaches, Light Goods Vehicles, All HGVs, All Motor Vehicles	Temperature	3287
King Street	Year, DWY, Wind Speed, Wind Direction, Cloud Cover, Rainfall, Humidity, Pressure, Buses and Coaches, Light Goods Vehicles, All HGVs, All Motor Vehicles	DWW, Temperature	2191
Wellington Road	Year, DWY, DWW, Wind Speed, Cloud Cover, Rainfall, Temperature, Humidity, Pressure, Buses and Coaches, All HGVs, All Motor Vehicles	Wind Direction, Light Goods Vehicles	2557
Union Street	Year, DWY, DWW, Wind Speed, Wind Direction, Rainfall, Temperature, Humidity, Pressure, Buses and Coaches, Light Goods Vehicles, All HGVs, All Motor Vehicles	Cloud Cover, Pressure	3287

Table 2.7.1: Description of the final linear models at each of the 5 sites

2.8 Modelling Trend, Seasonality and Time Series Errors for Each Site

The fitted model is summarised in this section. Parameter estimates, standard errors and p-values for each site are obtained. These are presented in table 2.8.1.

Union St	Estimate	Standard Error	p-value
Intercept	-1.009e+02	1.479e+01	1.12e-11
Year	5.205e-02	7.352e-03	1.83e-12
DWW	-5.125e-02	6.009e-03	<2e-16
DWY	7.763e-02	9.342e-03	<2e-16
Wind Speed	-1.273e-01	2.956e-03	<2e-16
Wind Direction	1.902e-03	9.668e-05	<2e-16
Rainfall	2.294e-01	3.121e-02	2.59e-13
Temperature	-1.894e-02	1.914e-03	<2e-16
Relative Humidity	-3.410e-03	5.718e-04	2.78e-09
Buses and Coaches	-3.168e-03	4.005e-04	3.64e-15
Light Goods Vehicles	-4.818e-03	5.531e-04	<2e-16
All HGVs	6.308e-03	1.638e-03	0.00012
All Motor Vehicles	9.055e-04	7.301e-05	<2e-16

Table 2.8.1: Estimates, Standard Errors and p-values for final model for Union St

The table above can be summarised by taking individual parameter estimates and explaining their relationship with y, the log NO_2 concentration. Taking year as the variable of interest, provided all other variables are held constant, log NO_2 concentration will increase by $0.052\mu gm^{-3}$ for every year increase. Similarly, taking Wind Speed this time, provided all other variables are held constant, log NO_2 concentration will decrease by $0.13\mu gm^{-3}$ for every kilometre per hour increase in wind speed. One more interpretation from the model is taking the total number of all motor vehicles, provided every other variable is held constant, log NO_2 concentration will increase by $0.0091\mu gm^{-3}$ for every unit increase of all motor vehicles.

Parameter estimates, standard errors and p-values are collected at the other AURN sites in Aberdeen.

2.9 Model Diagnostics

Here we can see the AIC values and R^2 adjusted values for each of the AURN sites;

Site	AIC value	R^2 Adjusted
Errol Place	3998.662	0.3358
Anderson Drive	3003.127	0.5083
King St	698.653	0.5492
Wellington Road	524.7395	0.6664
Union St	789.086	0.5067

Table 2.9.1: Summary of the AIC value and R^2 Adjusted for each model corresponding to a specific site

Looking at the R^2 adjusted values for the models above, Wellington Road is the site for the model which has a response that has the most of its variation explained by the independent explanatory variables.

The following plots are the standardised residuals vs fitted values. The plot of the residuals is reasonably symmetrical, is distributed around zero, and there is no obvious pattern. The normal Q-Q plot shows that the distribution of the data follows the normal distribution with some skewness at the lower tail. This is shown on the 2nd plot in the Figure 2.9.1 as the points mostly follow the normal Q-Q line, with deviations at either end.

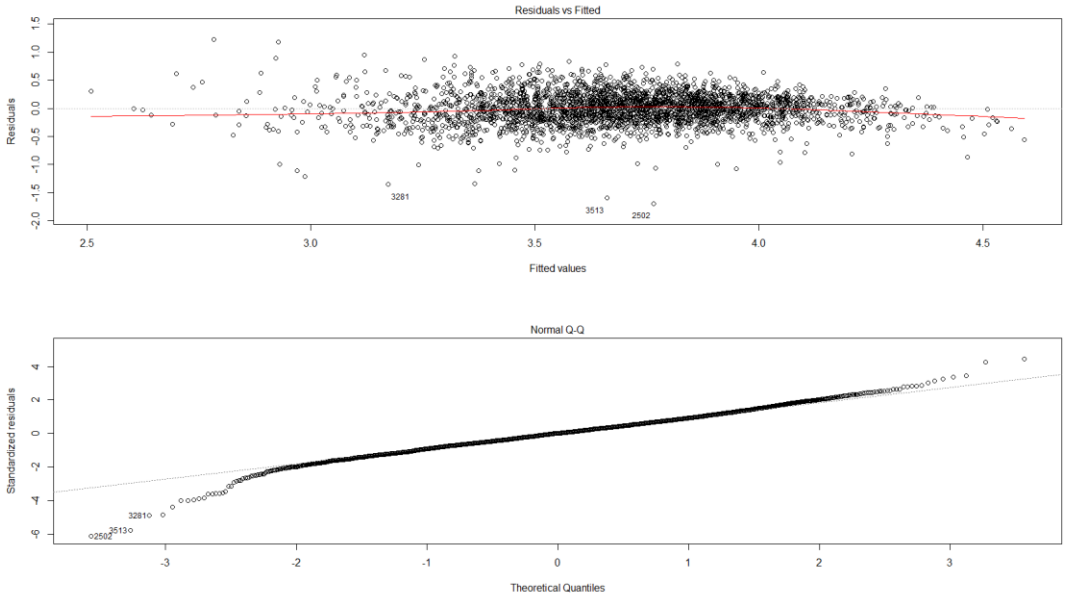


Figure 2.9.1: Log NO_2 Residuals at Union St, residuals vs fitted and normal Q-Q plots

Given the evidence of non linearity in the relationships between log NO_2 and covariates for the model built at Union St, as well as the low R^2 adjusted values for all of the models, it would serve well to ensure independence of the model for Union St, as well as plot some autocorrelation plots to see if there is any correlation present.

Figure 2.9.2 below shows the autocorrelation plot for the residuals from Union St, which indicates that there is some correlation remaining in the residuals at Union Street. The seasonality looks to be a weekly one as the ACF peaks approximately every 7 lags.

There is almost a cyclical trend as at lag 7 there is another peak of an autocorrelation of 0.267, which then decreases from lag to lag until it increases again to an autocorrelation of 0.219 at lag 14. As mentioned previously, this suggests that seasonality is present and NO_2 concentrations are more similar from week to week than they are from day to day. This acf plot also leads to the conclusion that the data come from an underlying autoregressive model. The next step is to estimate the parameters for the autoregressive model which can be found from the "time-shifted" series $\{Y_{t+h}, t = 0, \pm 1, \dots\}$, which is mentioned in section 2.3.1. Looking at the PACF of the residual series below (Figure 2.9.4) and Figure 2.9.3, the suggestion of using an AR(7) series may be appropriate.

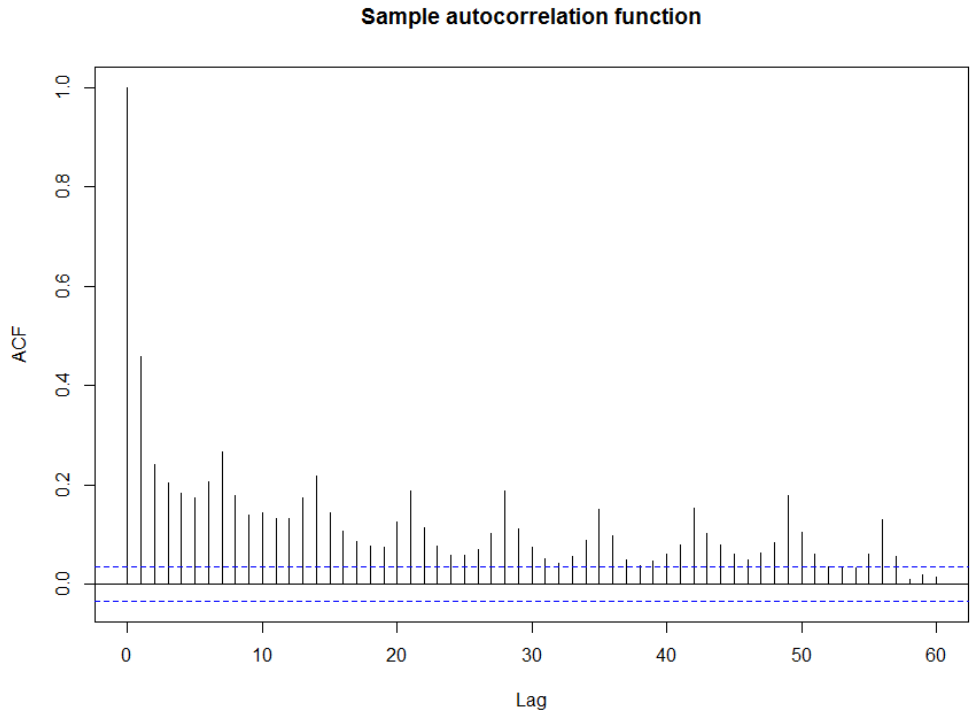


Figure 2.9.2: Autocorrelation plot for Log NO_2 concentrations recorded at Union Street.

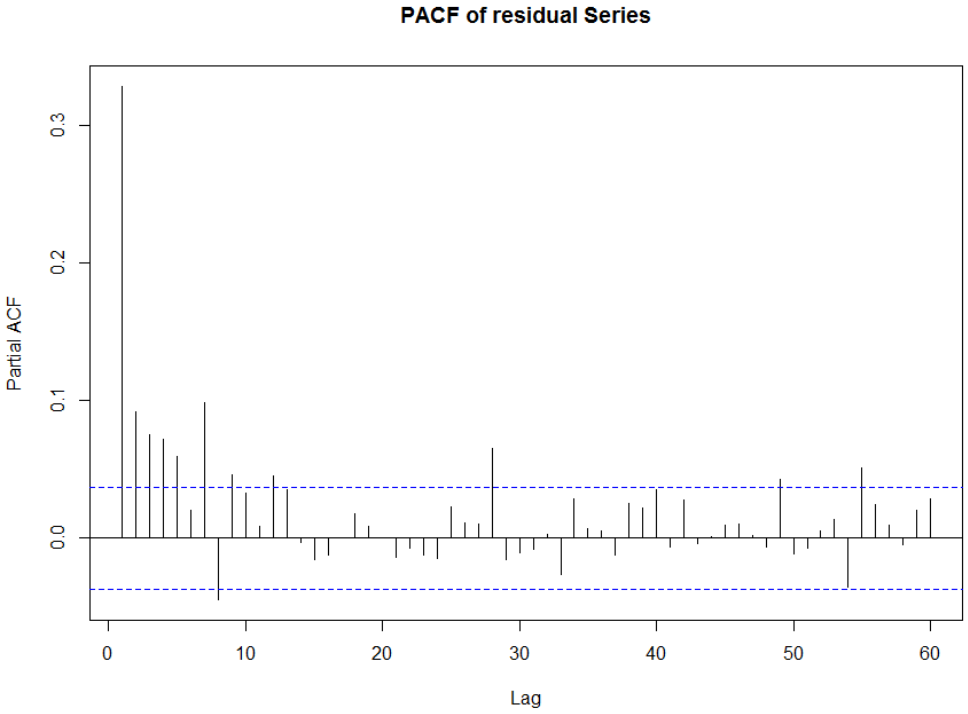


Figure 2.9.3: PACF of residuals from linear model at Union St

Plotting the residuals to ensure independence, we find that an AR(7) process is appropriate as there are no significant lags in either the ACF or PACF plot of the residual series and so the data does not need to be remodelled:

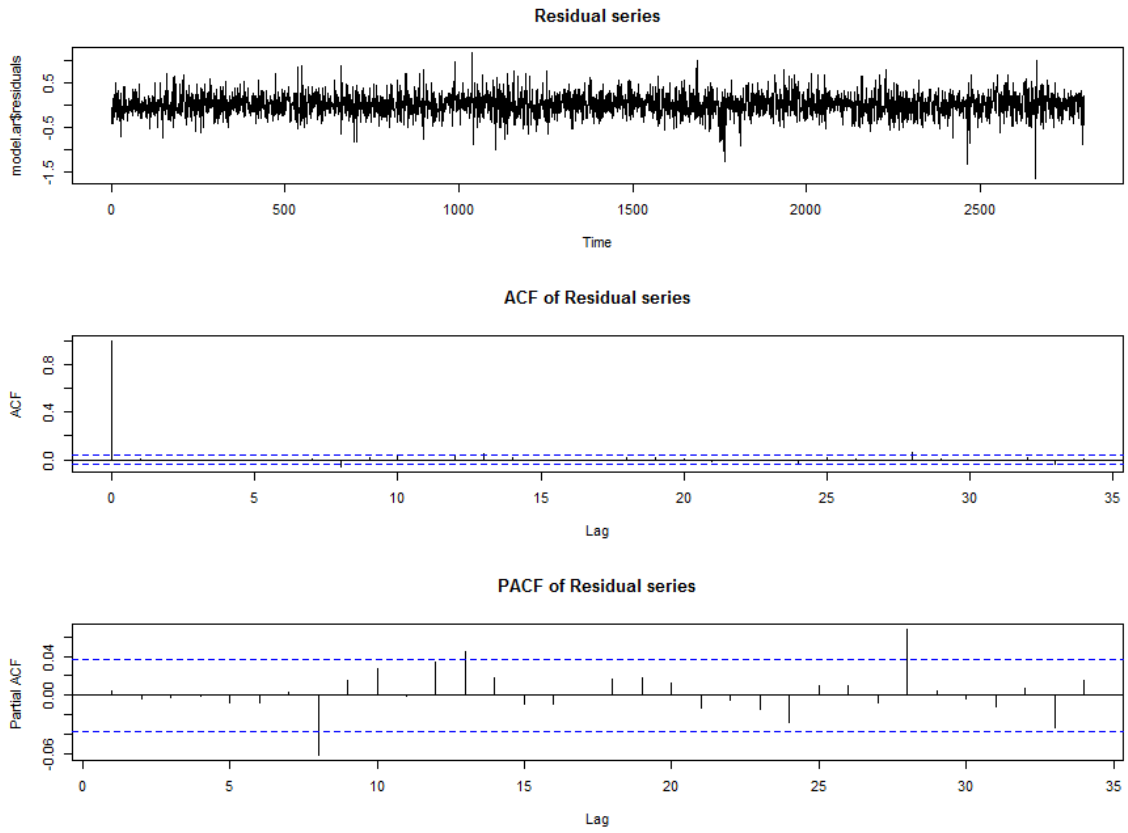


Figure 2.9.4: Time Series, ACF, and PACF of residuals from Union St linear model

2.10 General Additive Models

In this section general additive models are looked at for each site with plots to follow. This is needed as linear models have been proven to not be a good fit to the data. This methodology arises from most effects in real life not being linear. Hence, continuing on from the methodology in section 2.1.6, the results of the general additive models are as follows (in table 2.10.1 below). There are different n's here for the models at different sites than the n's at the same sites for linear models since there are different variables removed for the different models.

Model Site	Model Description	Variable(s) removed	N
Full Model at all sites	Year, DWY, DWW, Wind Speed, Wind Direction, Cloud Cover, Rainfall, Temperature, Humidity, Pressure, Buses and Coaches, Light Goods Vehicles, All HGVs, All Motor Vehicles	NA	3287
Anderson Drive	Year, DWY, Wind Speed, Wind Direction, Cloud Cover, Rainfall, Temperature, Humidity, Pressure, Buses and Coaches, Light Goods Vehicles, All HGVs, All Motor Vehicles	DWW, Buses and Coaches, Light Goods Vehicles, All HGVs	2749
Errol Place	Year, DWY, DWW, Wind Speed, Wind Direction, Cloud Cover, Temperature, Humidity, Pressure, Buses and Coaches, All Motor Vehicles	Rainfall, Light Goods Vehicles, All HGVs	2726
King Street	Year, DWY, Wind Speed, Wind Direction, Cloud Cover, Humidity, Pressure, Buses and Coaches, Light Goods Vehicles, All HGVs, All Motor Vehicles	DWW, Temperature, Rainfall	2893
Wellington Road	Year, DWY, DWW, Wind Speed, Wind Direction, Cloud Cover, Rainfall, Temperature, Humidity, Pressure, Buses and Coaches, All HGVs, All Motor Vehicles	Pressure, Buses and Coaches, All HGVs, Light Goods Vehicles	2015
Union Street	Year, DWY, DWW, Wind Speed, Wind Direction, Cloud Cover, Rainfall, Temperature, Humidity, Pressure, Light Goods Vehicles, All HGVs, All Motor Vehicles	Buses and Coaches	2794

Table 2.10.1: Description of the Generalised Additive models at each of the 5 sites

Site	R^2 Adjusted value	GCV score	AIC value
Errol Place	0.631	0.14384	2451.385
Anderson Drive	0.65	0.12586	2104.82
King Street	0.622	0.064905	195.3558
Wellington Road	0.744	0.059111	20.14053
Union Street	0.649	0.056252	-111.2433

Table 2.10.2: Summary of R^2 Adjusted, GCV, and AIC for each model corresponding to a specific site

The following plots show the smooth fit for the explanatory variables and a 95% confidence interval for each. These are shown for each variable for each final model created for each site. The dashes along the x axis are known as the "rug" and indicate where the sample observations occurred.

An example of the models being used is that which is created for Errol Place. This is seen below;

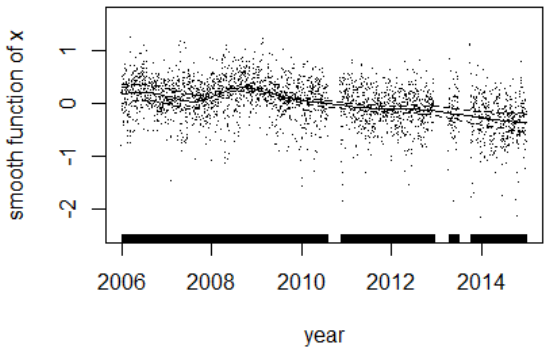
$$\begin{aligned} y_t = \beta_0 + f_1(year)_t + f_2\cos\left(\frac{2\pi(DWY)_t}{365}\right) + f_3\sin\left(\frac{2\pi(DWY)_t}{365}\right) + f_4\cos\left(\frac{2\pi(DWW)_t}{7}\right) \\ + f_5\sin\left(\frac{2\pi(DWW)_t}{7}\right) + f_6(Wind\ Speed)_t + f_7(Wind\ Direction)_t \\ + f_8(Cloud\ Cover)_t + f_9(Temperature)_t + f_{10}(Humidity)_t \\ + f_{11}(Pressure)_t + f_{12}(Buses\ and\ Coaches)_t + f_{13}(All\ Motor\ Vehicles)_t \\ + \varepsilon_t \end{aligned}$$

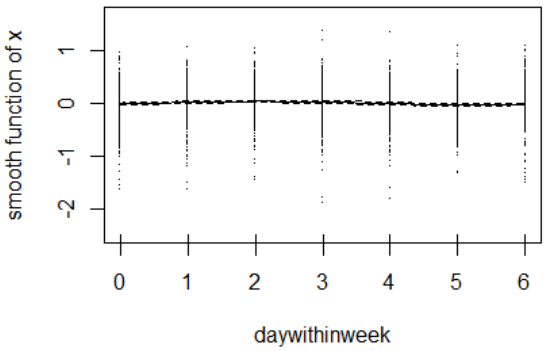
where $y = \log NO_2$ concentration, and $t = 1, \ldots 2726$ since there are 2726 days with available data spanning the years 2006 – 2014 for Errol Place. Each f_x is some smoothing function for $x = 1, \ldots, 13$.

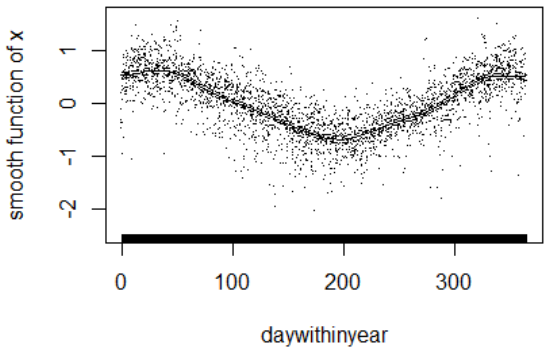
Similarly to the linear models, variables are removed in order of the most statistically insignificant variable being removed first, followed by the next most statistically insignificant variable and so on. This is done until only statistically significant variables are remaining in each model for each site.

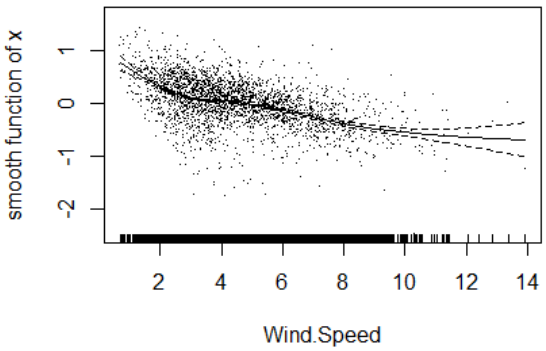
Errol Place

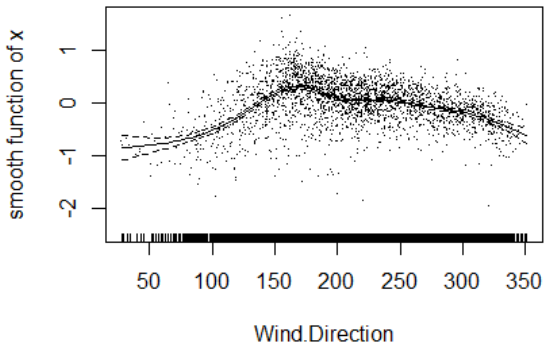
The following plots are for the generalised additive model which was created for Errol Place.

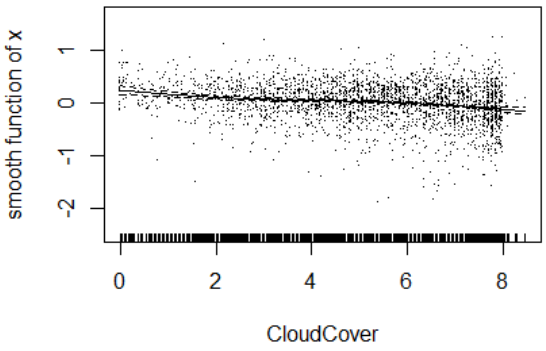












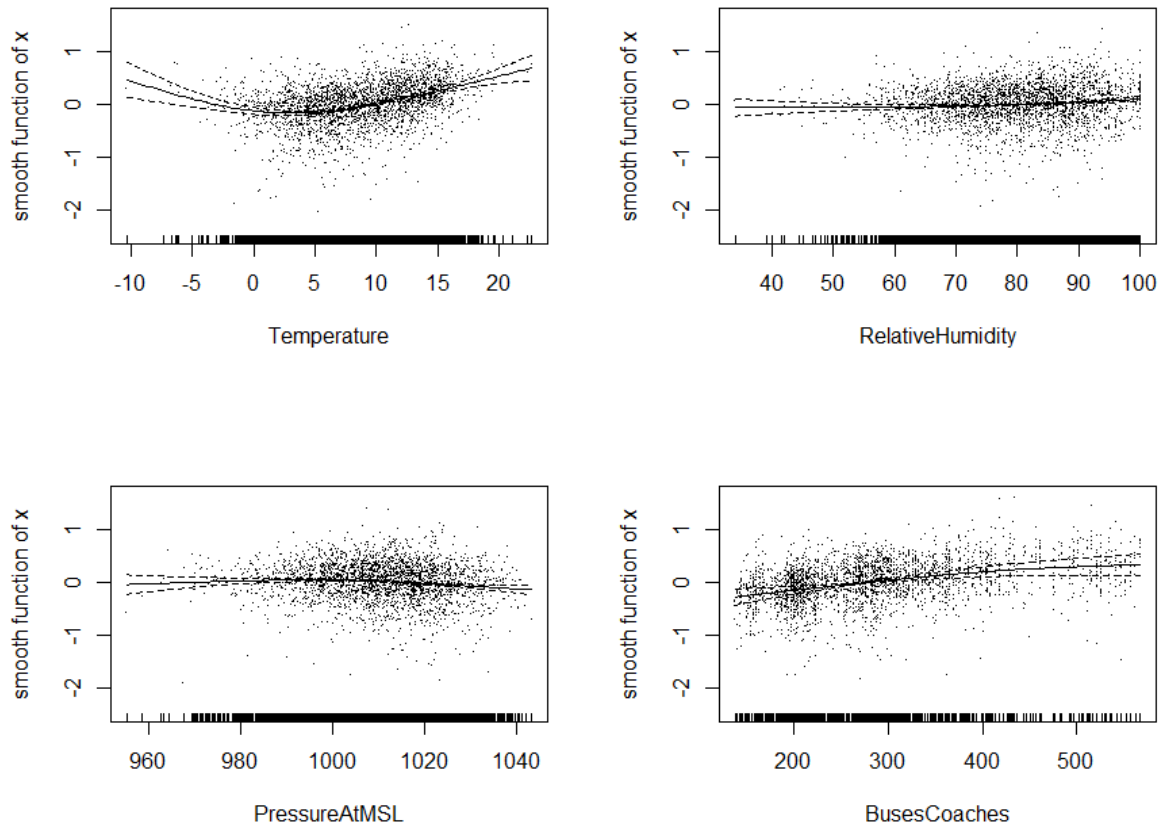


Figure 2.10.1: Plots of the fit of the explanatory variables in the GAM for Errol Place

From the first plot in Figure 2.10.1 (the plot for year) it looks as if there is a decline of the NO_2 concentration between 2006 and the end of 2007, followed by an increase of NO_2 concentration until around the start of 2009, followed by a steady decrease until 2015. The next plot (depicting the day within week variable) suggests that the NO_2 concentration does not vary much from day to day within the week.

The third plot in Figure 2.10.1 (showing the day within year variable) suggests that there is a lower concentration of NO_2 in the middle of the year, between May and August while there is a higher concentration of NO_2 at the start and end of the year, during the Winter months – November to February. The plot following this one (showing wind speed) suggests that as wind speed increases, the concentration of NO_2 decreases, almost linearly.

The fifth plot (showing wind direction) suggests that the concentration of NO_2 is higher when the wind is blowing in a southernly direction compared to any other direction. Most observations

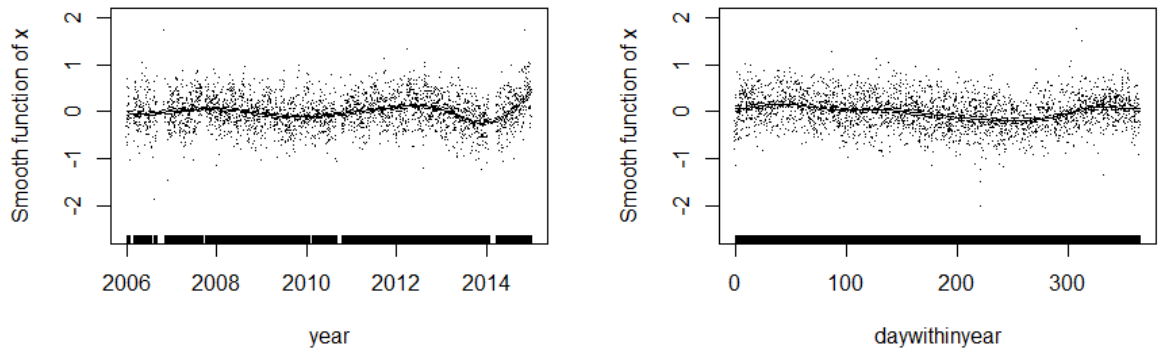
are recorded between East, South and West. The following plot (cloud cover) shows that as cloud cover increases, NO_2 concentration decreases. This relationship is almost linear.

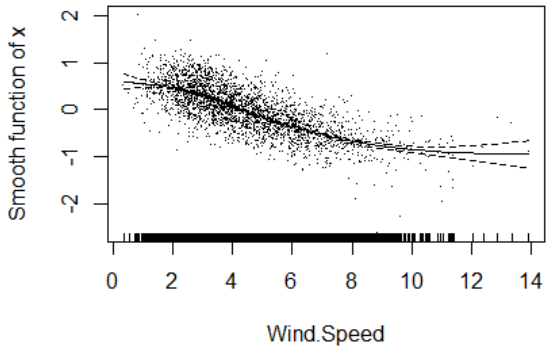
The plot showing temperature suggests a quadratic relationship between temperature and NO_2 concentration, with the concentration being relatively low between temperatures of 0 and 10°C. This occurrence could be due to less observations being recorded out with these temperatures and NO_2 concentrations are not, in fact higher at temperatures lower than 0°C. The plot also suggests that from 5°C and higher the NO_2 concentration increases. The plot showing relative humidity suggests that as relative humidity changes, the NO_2 concentration does not change much, when compared to other covariates in this model.

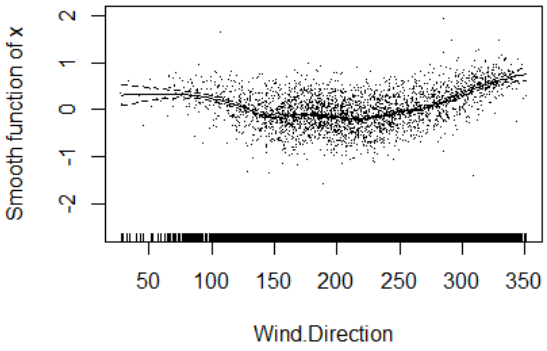
The second last plot (showing pressure) suggests that the pressure at mean sea level does not have much of an effect on the NO_2 concentration. This may be due to the small range of the pressure at mean sea level covariate. The final plot in Figure 2.10.1 shows that the NO_2 concentration increases as the number of buses and coaches increases. This shows that the NO_2 produced by the buses and coaches does increase the amount of NO_2 recorded by the measuring instruments.

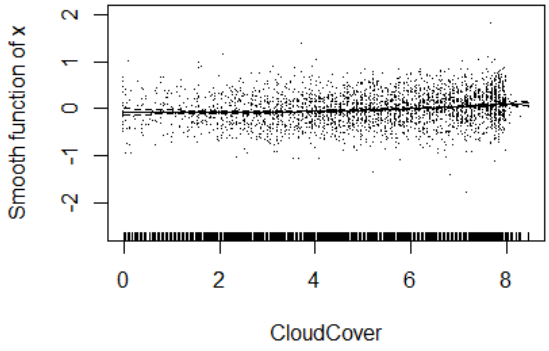
Anderson Drive

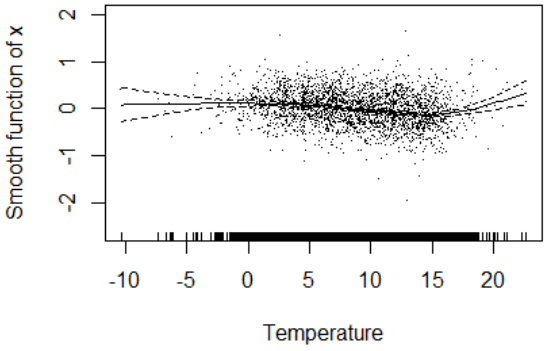
The following plots are for the generalised additive model created for Anderson Drive.

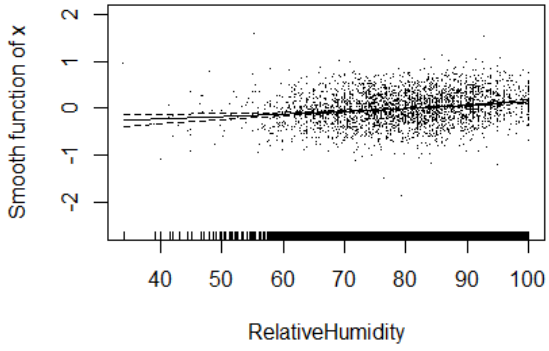


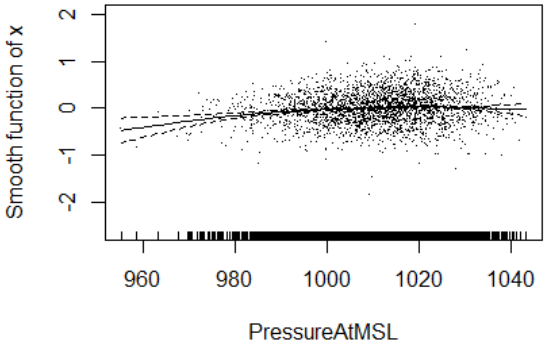

Smooth function of x
-2
-1
0
1
2
0
2
4
6
8
10
12
14
Wind.Speed


Smooth function of x
-2
-1
0
1
2
50
100
150
200
250
300
350
Wind.Direction


Smooth function of x
-2
-1
0
1
2
0
2
4
6
8
CloudCover


Smooth function of x
-2
-1
0
1
2
-10
-5
0
5
10
15
20
Temperature


Smooth function of x
-2
-1
0
1
2
40
50
60
70
80
90
100
RelativeHumidity


Smooth function of x
-2
-1
0
1
2
960
980
1000
1020
1040
PressureAtMSL

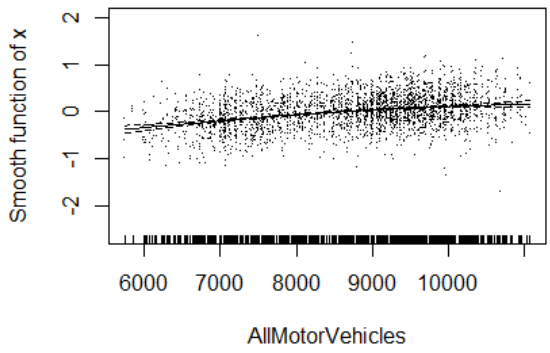


Figure 2.10.2: Plots of the fit of the explanatory variables in the model for Anderson Drive

The first plot of Figure 2.10.2 (showing year) shows that the NO_2 concentration does not change from year to year for the years 2006 – 2012 approximately, whereas after 2012 there is a decrease until 2014, and an increase of the NO_2 concentration after 2014. The plot showing day within year suggests that the NO_2 concentration does not change much depending on the day of the year at Anderson Drive, apart from maybe a slight decrease of NO_2 concentration in the Autumn months.

The third plot of figure 2.10.2 showing wind speed suggests that as wind speed increases, the concentration of NO_2 decreases. This could be due to the NO_2 being dispersed into the wider troposphere. Similar to the plot for wind direction at Errol Place, the next plot suggests the concentration of NO_2 is lower when there is a southern wind when compared to when there is a northern wind for the NO_2 concentration recorded at Anderson drive. This could be due to the weather variables being recorded at the one station of Dyce, instead of multiple recording sites.

The plot showing cloud cover at Anderson Drive suggests that although there is not much of a difference in NO_2 concentration at different cloud cover levels, there is a small increase in NO_2 as cloud cover increases. The next plot (depicting temperature) suggests the temperature at Anderson Drive has a negative effect on the NO_2 concentration i.e. as temperature increases, the NO_2 concentration decreases. This is up until approximately 15°C, where higher than 15°C the NO_2 concentration increases.

Showing relative humidity, the next plot suggests that NO_2 concentration increases as relative humidity increases. The relationship could be described as linear. The following plot (showing

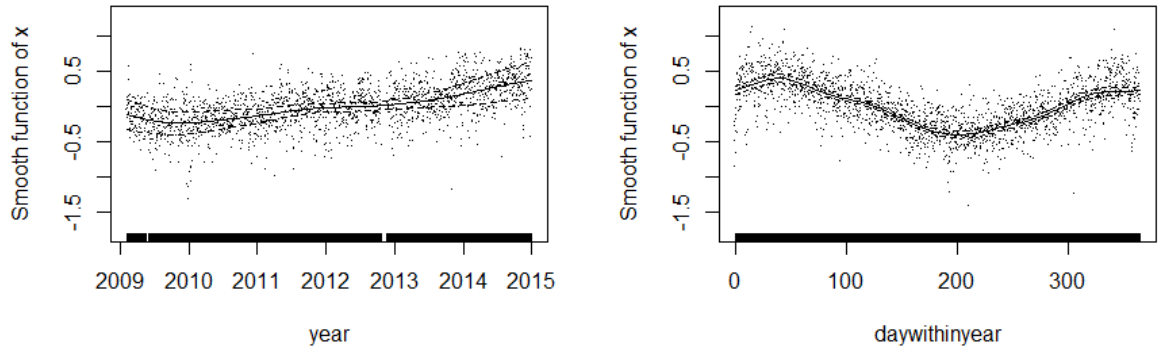
pressure at mean sea level) shows that there does not seem to be much of a change in NO_2 concentration as pressure at mean sea level changes, although the same plot does suggest that NO_2 concentration increases up until a point before levelling off as the pressure increases.

The final plot of figure 2.10.2 shows the total number of all motor vehicles, and it tends to suggest that as the total number of all motor vehicles increases so too does the NO_2 concentration. This could almost be described as a linear relationship given this final plot.

King Street

The next figure shows the plots for the generalized additive model created for King Street.

The first plot in Figure 2.10.3 shows that the NO_2 concentration increases from year to year. This is similar to the recordings taken at Errol Place. Also similar to Errol Place and different to Anderson Drive, as suggested by the next plot, is that the NO_2 concentration is lower in Summer months compared to Winter months.



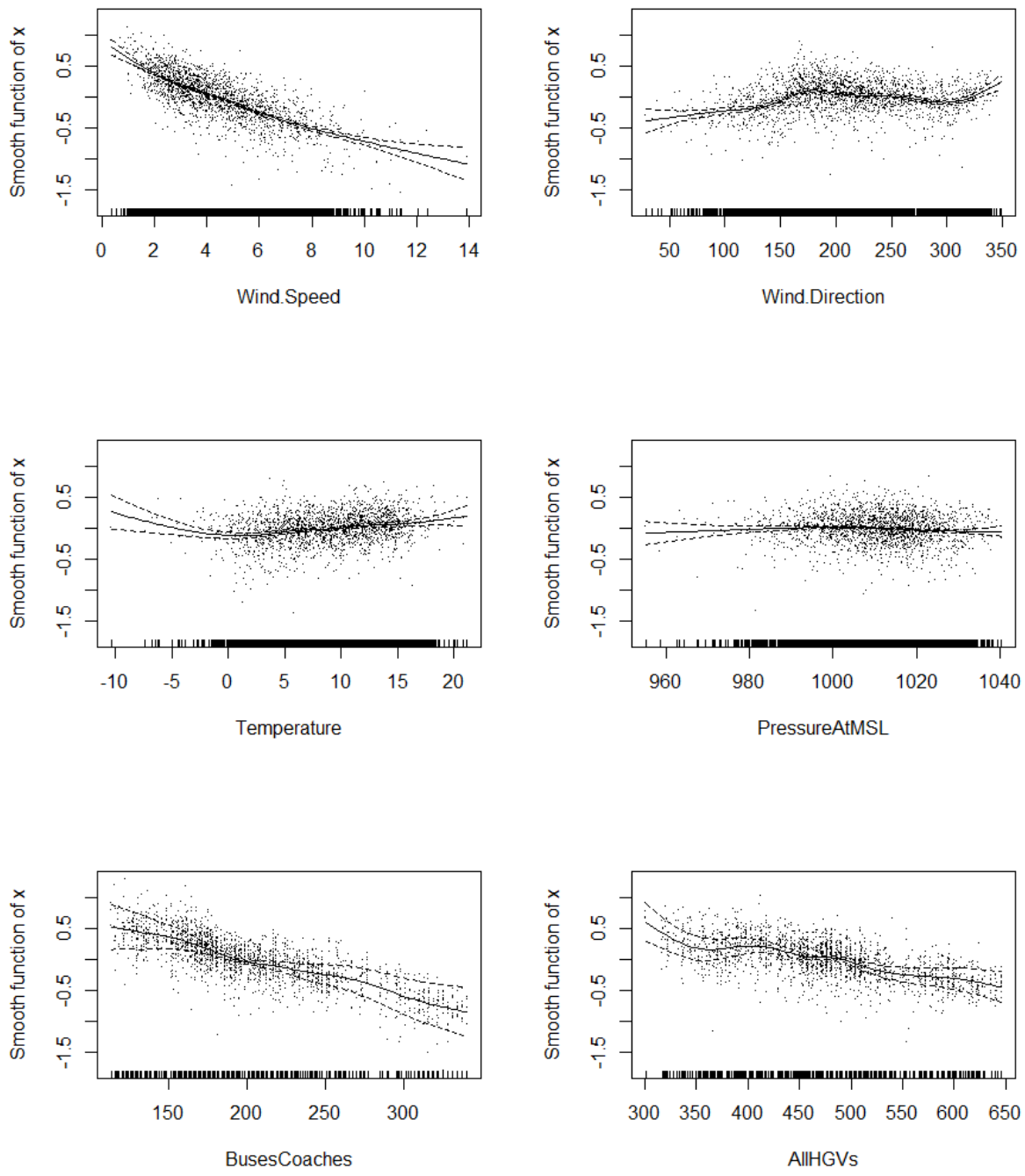


Figure 2.10.3: Plots of the fit of the explanatory variables in the model for King St

Similar to both Anderson Drive and Errol Place, the plot showing wind speed indicates that NO_2 concentration deceases at King St as the wind speed increases. The plot showing wind direction shows that the wind direction has a different effect on the NO_2 concentration when compared to the wind direction on NO_2 concentration at Anderson Drive or Errol Place – between North and

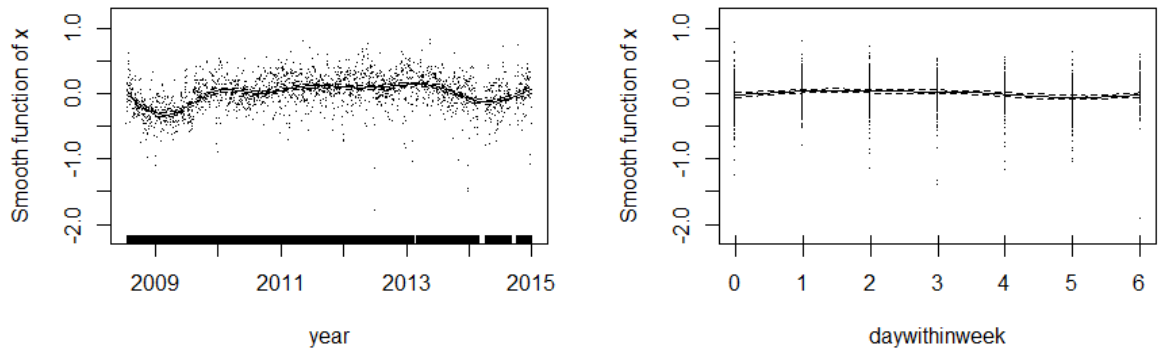
South (travelling around the compass in a North, Northeast, East, Southeast, South direction) the NO_2 concentration stays approximately at the same concentration. Then between South and North (travelling around the compass through West) the NO_2 concentration is a little higher, although still stays around the same concentration between South and North.

The plot showing temperature suggests as temperature increases, the NO_2 concentration stays at approximately the same concentration for King St. The plot which shows pressure at mean sea level for King St suggests that as pressure at mean sea level increases, there is very little change in the NO_2 concentration at King St.

Counter-intuitively the plot showing buses and coaches suggests that as the number of buses and coaches increases, the NO_2 concentration decreases. This is different from the buses and coaches plot at Errol Place, which suggest the converse – as the number of buses and coaches increases, the NO_2 concentration increases. Similar to the buses and coaches plot, the showing the number of HGVs suggests that as the number of HGVs increases, the concentration of NO_2 decreases. This is also counter-intuitive like the buses and coaches plot and is different to those plots for vehicle classes and NO_2 concentration at Anderson drive or Errol Place.

Union St

The next figure of plots shows the meteorological and traffic variables and their relationship with log NO_2 at Union St.



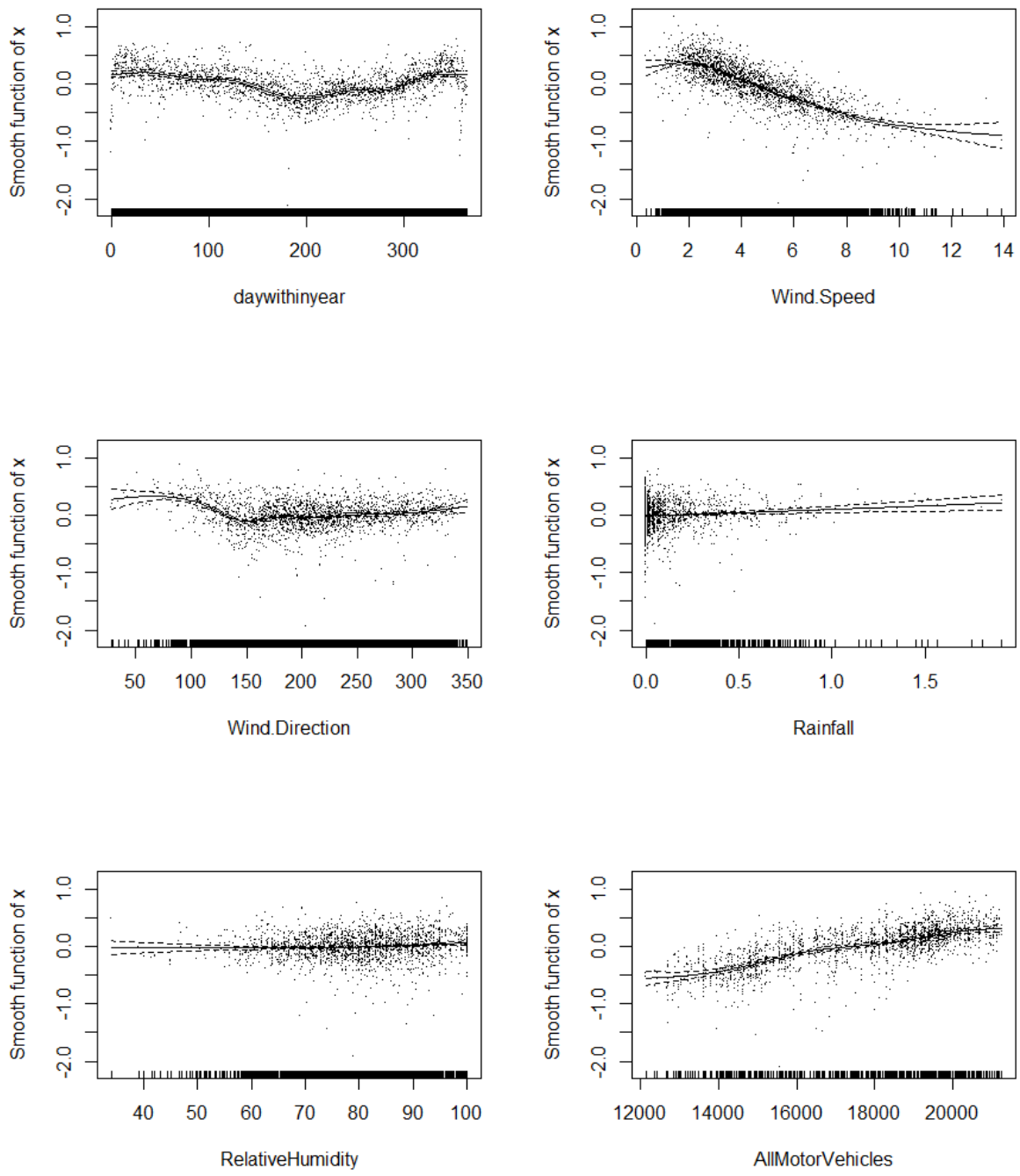


Figure 2.10.4: Plots of the fit of the explanatory variables in the model for Union St

At Wellington Road, from looking at the first plot in Figure 2.10.4, it can be seen that the NO_2 concentration is approximately the same between 2010 and 2013. Between 2009 and 2010 the NO_2 concentration takes on a quadratic shape, with similar behaviour between 2013 and 2015. According to the second plot in the same figure the NO_2 concentration does not change

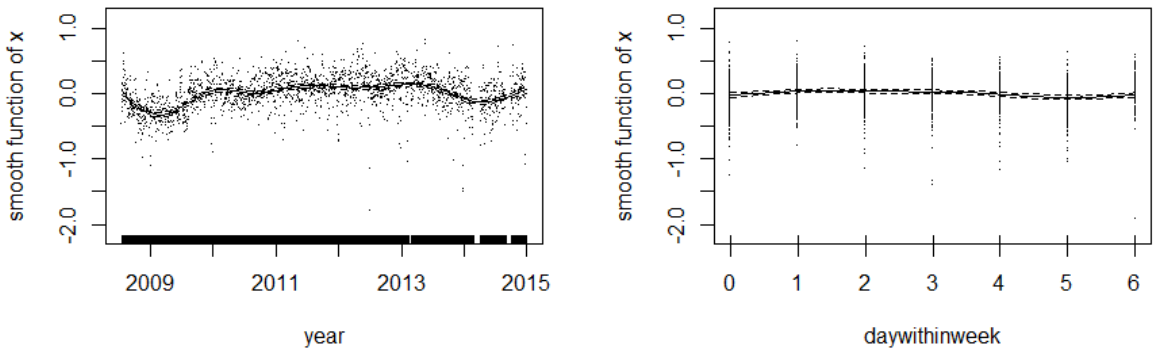
according to the day of the week for Wellington Road. This is similar to the day within week variable for Errol Place – not much change from day to day in the NO_2 concentration.

The day within year variable plot of Figure 2.10.4 shows that similar to Errol Place and King St, there is a lower concentration of NO_2 in the summer months while there is a higher concentration in the winter months. The next plot in the figure (wind speed) shows wind speed increasing has the effect of a lower NO_2 concentration at Wellington Road. This is consistent with the wind speeds effect on NO_2 concentration at all other sites.

The wind direction from East around to North doesn't seem to affect a change in the NO_2 concentration much according to the above plot. In saying that there is a decrease in NO_2 concentration as the wind direction goes from North to East. This is seen in the plot depicting wind direction. The plot in Figure 2.10.4 showing rainfall suggests that as rainfall increases so too does NO_2 concentration at Wellington road. This relationship can be described as almost linear. It should be noted that most observations of rainfall are recorded between 0.0mm and 0.5mm, which makes rainfall relatively difficult to model.

At Wellington Road, the plot showing all motor vehicles suggests that as the number of all motor vehicles increases, so too does the NO_2 concentration, which is in conjunction with other vehicle plots from Errol Place.

Wellington Road



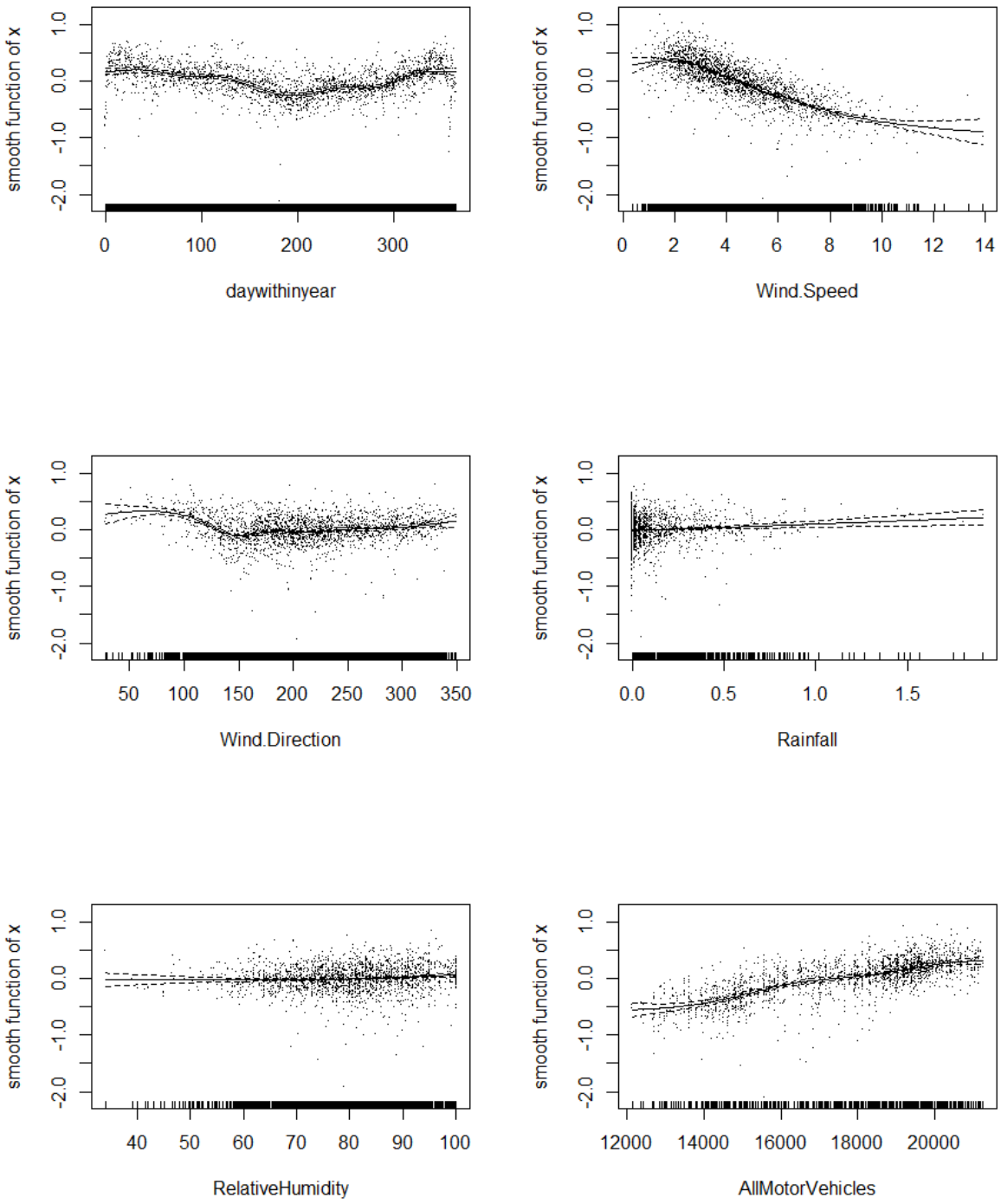


Figure 2.10.5: Plots of the fit of the explanatory variables in the model for Union St

At Union St the NO_2 concentration rises steadily from 2006 to 2011 and after 2011 it decreases up until 2015. This can be seen above in the first plot of Figure 2.10.5. Similar to other sites, the

NO_2 concentration does not change much from day to day at Union St as can be seen from the second plot of Figure 2.10.5.

Similar to other sites for the day within year covariate, the NO_2 concentration is lower in summer months compared to winter months for Union St. This is seen in the plot with the day within year covariate. Consistent with all other sites, the NO_2 concentration at Union St decreases almost linearly as wind speed increases as can be seen with the plot showing wind speed.

The plot in Figure 2.10.5 which shows wind direction suggests that from Northeast to West, the NO_2 concentration increases. The next plot (showing relative humidity) suggests that the NO_2 concentration changes only slightly as relative humidity increases. This is similar to other sites.

As with Anderson Drive, Errol Place and King St, NO_2 concentration changes only slightly as the Pressure at mean sea level changes. This can be seen in the plot showing Pressure for Union Street. From the plot showing the number of HGVs, it can be said that NO_2 concentration decreases as the number of HGVs increases before levelling off. This is different to other sites. As the number of all motor vehicles increases, the NO_2 concentration also increases for Union St. This is seen in the final plot of Figure 2.10.5 and this is similar to other sites throughout Aberdeen.

All of these models have been fitted using software R Studio, and using the R package "mgcv", and hence the R command "gam". Using the "mgcv" package means the smoothing parameters are estimated using the GCV (Generalised Cross Validation) criterion;

$$n\frac{D}{(n-DoF)^2}$$

Or an Unbiased Risk Estimator (UBRE) criterion;

$$\frac{D}{n}+2\,s\frac{DoF}{n}-s$$

where D is variance, n is the number of data, s is the scale parameter and DoF is the effective degrees of freedom of the model. There are other ways of estimating the smoothing parameter – namely using maximum likelihood or restricted maximum likelihood [73]. These methods are discussed in more detail in the next chapter.

It is also of interest to see how a GA model with an autoregressive error process implemented may have an effect on parameter estimates. This checks the assumptions of the gams, the assumptions being that the data and also the errors should be independent. Plots of the autocorrelation function of the GAM with an AR (1) process can be seen below. This was done for Union St and the methodology is reflective of the methodology for GAMs at other sites.

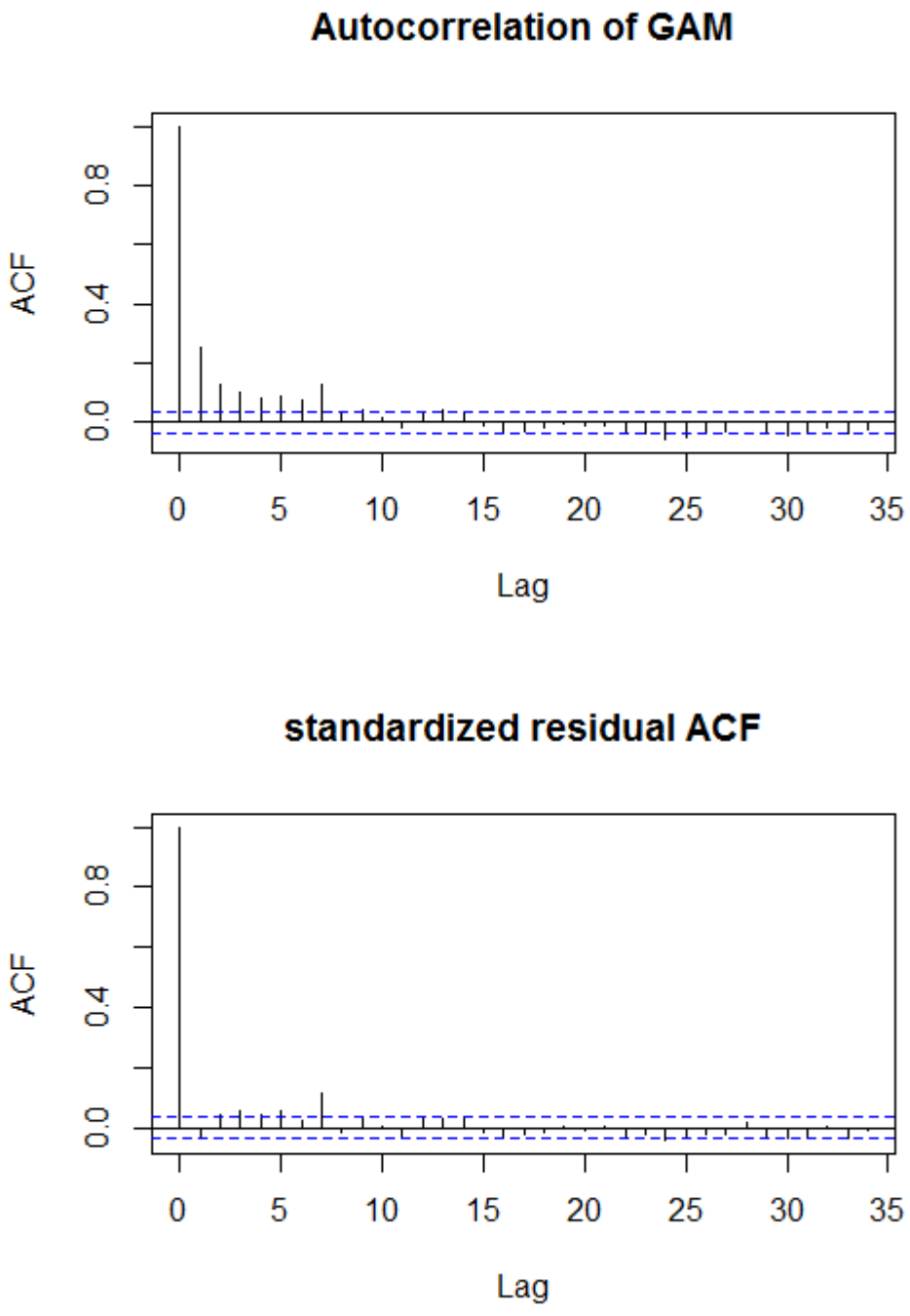


Figure 2.10.6: The AR(1) process for the GAM built at Union St and standardizd residual ACF

2.11 Conclusions for time series modelling using covariates

Carrying out linear and generalised additive modelling at all sites leads to the conclusion that a generalised additive model is more appropriate for all of the 5 site locations. This can be seen by comparing the R^2 adjusted values and the AIC values from tables 2.9.1 and 2.10.2. It is also conclusive from looking at summaries of the different models that the more accurate models contain explanatory variables related to time, meteorological factors, and traffic factors (again comparing tables 2.9.1 and 2.10.2). It is of interest that some of the models differ from site to site, that different explanatory variables are statistically significant at some sites, and are statistically insignificant at other sites. For example, the DWW variable is significant at Errol Place, Wellington Road and Union Street and insignificant at Anderson Drive and King Street. Buses and Coaches are significant at 2 out of the 5 sites, namely Errol Place and King Street, whilst Light Goods Vehicles and All HGVs are significant at King Street and Union Street, whereas they are not significant at all other sites. Rainfall, Temperature and Pressure also come under this category of being significant at some sites whilst not significant at others.

Comparing the R^2 Adjusted values of the generalised additive models with the corresponding values from the linear models show that the GAMs are better models in the sense that more variance is explained by the GAMs. It is also a conclusion that the GAMs are more appropriate than the linear models at each site due to the nature of the data – time series data of this nature cannot be assumed to consist of independent observations, which is one of the assumptions needed for the linear models to hold true.

From comparing the covariate day within year at each site with one another, there seems to be a difference between them – different shapes. These can be seen in Figures 2.10.1 – 2.10.5, these Figures also display the smooth function of each covariate included in the corresponding model to each site. These are accompanied by a rug plot on the same plot as the smooth function and 95% confidence intervals. These confidence intervals can be seen to be spanning outwards at either end of the curve for most covariates. This is due to a lack of data, before or after the respective start or end of the curve. Some of the plots show that the terms are almost linear, and that these should be included in the final model, only not as a smooth function term, but as a regular covariate i.e. a covariate which has a linear relationship with the response. Some of the variables in the model are treated with cyclic smoothing, meaning that the start and end of the

curves would meet if they were to be put end to end. This can be seen in any model which has day within year or day within week as a covariate.

It has been proven from this work that some variables which would be expected to have an adverse effect on NO_2 concentrations, namely different vehicle classes, do not have as much of an expected effect in comparison to weather factors, which are uncontrollable. Take for example, the models built for Wellington Road and Anderson Drive – these particular models only consist of the one vehicle class "All motor vehicles", whereas one would expect, intuitively, that other vehicle classes would help explain the concentration of NO_2. In comparison these models consist of a number of meteorological variables which explain the NO_2 concentrations (in conjunction with the other explanatory variables) reasonably well – this according to AIC and R^2 adjusted values. This is vital when considering a model, since although we can control for the number of vehicles travelling down a particular road on any given day, we cannot control for the speed the wind is travelling or how warm that particular day is. This is why other avenues must be explored to see what other factors are affecting the NO_2 concentrations in Aberdeen, namely spatial factors. The vehicle variables themselves are in fact model as discussed in a previous section. This means that the data may not be truly reflective of the relationship between vehicle classes and NO_2 concentration. Further work which can be done is research into methods of modelling data so that the traffic data can be modelled in an improved way, or perhaps even include uncertainty and see how the models will change.

What is next in the model building stage is to move in to the spatial dimension and model the data in this way – sole temporal analysis has been completed, and the next logical step is to see how sites depend on one another in relation to space, and in relation to their location.

Chapter 3: Spatial Modelling of Air Quality in Aberdeen

3.1 Introduction

The previous Chapter focuses on time series analysis of NO_2 at 5 sites across Aberdeen. It was shown that the NO_2 concentrations were not uniform across space, and in fact, varied from site to site. These monitoring sites are not located uniformly through space, but fall at different locations across the city. This Chapter, focussing on the spatial aspect of the data, takes into account the 5 sites mentioned previously, as well as 51 diffusion tube monitoring sites which are also located throughout space within Aberdeen City. This is the same diffusion tube data that are described in Chapter 1. The locations of the monitoring sites are included in this Chapter in a later section (section 3.3).

The preliminary investigation of the spatial aspects of the data is relevant to checking whether assumptions made by a potential model are relatively satisfied. Also of interest is how the NO_2 concentrations change over space – this is done using Eastings and Northings i.e. coordinates. Later in the Chapter, in addition to spatial locations, variables that are used to model the NO_2 concentrations at each location are included in the data set. This included traffic and meteorological variables, which are the same as the traffic and meteorological variables described in previous Chapters.

Chapter 3 starts with highlighting the main methods that are used to analyse the data in a spatial context. Potential spatial patterns of the annual mean NO_2 values are explored initially, and then a more formal investigation is undergone using a full spatial model. This is discussed in further detail later. In conclusion, the Chapter finishes with some final thoughts, discussion and further work on the spatial trends of NO_2 across Aberdeen City. The annual mean NO_2 is used as the available data are limited to an annual concentration for all of the sites investigated in this Chapter – although there are daily and hourly values of NO_2 available for the AURN sites, the data from the LAQM diffusion tube sites are at an annual concentration (and a monthly concentration – for the purposes of this paper, the annual mean is used, which also falls in line with the regulations which are phrased in terms of an annual mean). To be consistent, a log transformation of the NO_2 values has been taken.

The aims of this Chapter are to (spatially) map the NO_2, interpolate and predict future values of NO_2 concentrations; generate a map of the NO_2 concentrations across Aberdeen city while

locating hotspots and evidence of spatial patterns; and finally to investigate change of the NO_2 concentrations through space.

3.2 Geostatistical Modelling Methods

3.2.1 Spatial Process

Since NO_2 is recorded at a specific geographical location, and the fact NO_2 is everywhere, the NO_2 concentration can be described by a geostatistical process. The prefix *"geo"* appears to refer to statistics pertaining to the earth, and this was indeed its original meaning, although now geostatistics has taken on a much more universal role, one which is concerned with statistical theory and applications for processes with continuous spatial index [35].

Defining a spatial process helps explain geostatistical modelling. Take a stochastic process, X the spatial data can be thought of as being generated from this process [36] but instead of a time index, a spatial index is being used, which indicates locations.

If S is taken as the stochastic process, then $S(x)$ describes the concentration of NO_2 as a function of location, x, of said stochastic process. Taking a Gaussian model, and keeping it as simple as possible while still meeting the requirements of $S(x)$, the model can be seen below in equation 3.2.1.1 [36]. In its simplest form, a set of geostatistical data is denoted by $(x_i, y_i) : i = 1, \dots, n$ where x_i are spatial locations and y_i is the measured value associated with the location x_i. The assumptions underlying this model are as follows;

- $\{S(x) : x \in \mathbb{R}^2\}$ is a Gaussian process with mean μ, variance $\sigma^2 = Var\{S(x)\}$ and correlation function $\rho(u) = Corr\{S(x),\ S(x')\}$, where $u = \|x - x'\|$ and $\|\cdot\|$ denotes distance;
- Conditional on $\{S(x) : x \in \mathbb{R}^2\}$, the y_i are realisations of mutually independent random variables Y_i, normally distributed with conditional means $E[Y_i|S(\cdot)] = S(x_i)$ and conditional variances τ^2.

Equivalently, the model can be defined as

$$Y_i = S(x_i) + Z_i : i = 1, \dots, n \qquad (3.2.1.1)$$

where $\{S(x) : x \in \mathbb{R}^2\}$ is defined by the first assumption above, and the Z_i are mutually independent $N(0, \tau^2)$ random variables.

The correlation function $\rho(u)$ must be positive definite for a legitimate model to be defined [35]. This condition imposes non-obvious constraints to ensure that, for any integer m, of locations x_i, and real constants a_i, the linear combination $\sum_{i=1}^{m} a_i S(x_i)$ will have non-negative variance. The focus is primarily on a flexible, two-parameter class of correlation functions which, due to Matèrn [37], takes the form;

$$\rho(u;\ \phi, \kappa) = \{2^{\kappa-1}\Gamma(\kappa)\}^{-1} \left(\frac{u}{\phi}\right)^{\kappa} K_{\kappa}(\frac{u}{\phi}) \qquad (3.2.1.2)$$

Where $K_{\kappa}(\cdot)$ denotes the modified Bessel function of the second kind, of order κ. The parameters $\phi > 0$ determines the rate at which the correlation decays to zero with increasing u. The parameter $\kappa > 0$ is called the *order* of the Matèrn model, and determines the differentiability of the stochastic process $S(x)$ [35].

The notation used here for $\rho(u)$ presumes that $u \geq 0$. However, the correlation function of any stationary process must be symmetric in u, hence $\rho(-u) = \rho(u)$.

The stochastic variation in a physical quantity is not always well described by a Gaussian distribution. One of the simplest ways to extend the Gaussian model is to assume that the model holds after applying a transformation to the original data. For positive-valued response variable, a useful class of transformation is that of the Box-Cox family [38]. Although for the nature of the NO_2 data, a log transformation will suffice.

3.2.2 Stationary and Isotropy

In spatial analysis, by saying stationary here, it should be understood as the following: the distribution of the random process has certain properties which are the same everywhere, including the covariance, and it has no spatial trend or spatial periodicity [39]. Since S is a random process, it can be said to be strictly stationary if the joint distribution of $S(x_i)$ is the same as $S(x_i + h)$ for $x_1, \ldots, x_k$ i.e. every $S(x)$ in the spatial domain in question is identically distributed and the locations do not affect the distribution, only the distance between said locations. When this is the case, and only distance is of importance then the process is known as isotropic.

The covariance function of the $S(x)$ (assuming stationary and isotropic) can be described as the following;

$$C(\underline{h}) = cov\left(Y(\underline{s}), Y(\underline{s}+\underline{h})\right)$$

$$= E((Y(\underline{s}) - \mu)(Y(\underline{s}+\underline{h}) - \mu)) \qquad (3.2.2.1)$$

Naturally, it follows that if $Y(\underline{s}) = Y(\underline{s}+\underline{h})$then the above equation defines the variance (which is assumed to be finite and the same everywhere). The above covariance function stands true for the two positions $\underline{s}$ and $\underline{s}+\underline{h}$, where $\underline{h}$ is the distance between the two positions [40]. The process is described as weakly stationary if the above equation has μ as a constant which does not depend on $\underline{s}$ and the covariance function is a finite constant which depends on $\underline{h}$ but not on $\underline{s}$. This covariance function is a function of the lag [40].

The correlation function of a stationary process is defined as:

$$\rho(\underline{h}) = \frac{C(\underline{h})}{\sqrt{C(\underline{0})C(\underline{0})}} = \frac{C(\underline{h})}{C(\underline{0})} \qquad (3.2.2.2)$$

3.2.3 Variograms

Covariance functions are a usual statistical tool for quantifying and modelling the correlation between observations. In geostatistics, however, a slight variant called variograms are also commonly used [40]. The use of variograms is in fact more common than the use of covariance functions, in the case that one is trying to identify whether spatial correlation exists in the data, at least.

The semi – variogram of a geostatistical process $\{Z(s): s \in D\}$ is a function denoted by $\gamma_Z(s,t)$, and measures the variance of the difference in the process at two spatial locations s and t. It is defined as

$$\gamma_Z(s,t) = \frac{1}{2}Var[Z(s) - Z(t)] \qquad (3.2.3.1)$$

Traditionally $2\gamma_Z(s,t)$ is called the variogram and $\gamma_Z(s,t)$ is called the semi – variogram.

It should be noted that when the relative variance of the difference $Z(s) - Z(t)$ is small then $Z(s)$ and $Z(t)$ are similar in the sense that they are spatially correlated [41]. When the difference between the two is large, they are more likely to be independent or less similar.

And when Z is stationary as described above:

$$C(\underline{h}) = \lim_{\|\underline{u}\| \to \infty} \gamma(\underline{u}) - \gamma(\underline{h}) \qquad (3.2.3.2)$$

The variogram and the semi-variogram share the following descriptive parameters: the nugget (ϕ^2), is the difference between the origin line and the limiting value of the variogram as $t \to 0$; the sill, is the limiting value of the variogram as $t \to 0$; the partial sill (σ^2), which is equal to the sill minus the nugget; and the range (λ) which is the distance at which the variogram reaches the sill [38].

This is seen more clearly in the image of a generic variogram below;

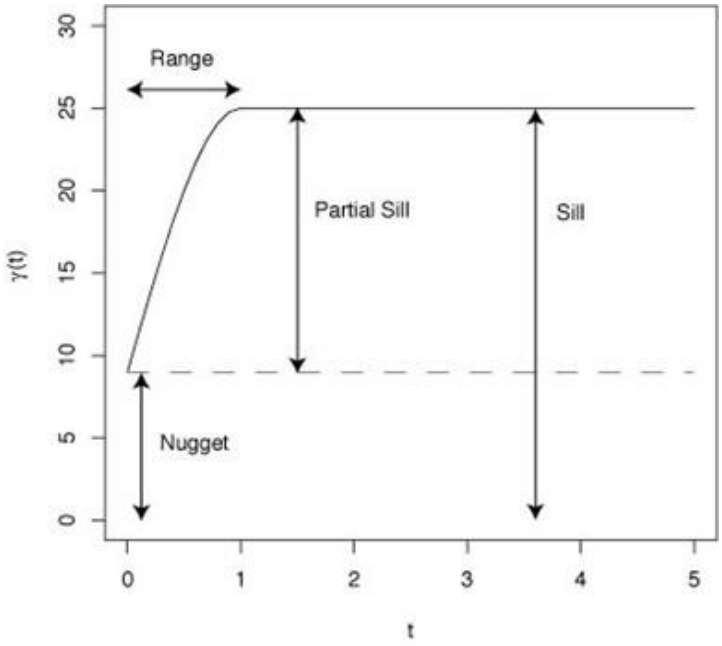


Figure 3.2.3.1: Generic Variogram [42]

An empirical variogram is a method which can be used to estimate a theoretical variogram. A binned empirical variogram can also be used. This is another method which is used to estimate a theoretical variogram, and is named "binned" as the process divides the distances into a number of intervals, so that we have;

$$I_l = (t_{l-1}, t_l], l = 1, \dots, L \qquad (3.2.3.3)$$

Letting $t_l^m = \frac{t_{l-1}+t_l}{2}$ denote the midpoint of the pairs of distances for each of the L intervals, then the binned empirical variogram is given by;

$$\hat{\gamma}(t_l^m) = \frac{1}{2j} N(t_l) \sum_{(s_i,s_j)\in N(t_l)} \left[y(s_i) - y(s_j)\right]^2 \qquad (3.2.3.4)$$

where $N(t_l) = \{(s_i, s_j): \|s_i - s_j\| \in I_l\}$.

When interpreting binned empirical variograms, caution should be used as the measures of uncertainty are relatively difficult to calculate [43]. It does occur that there are not enough pairs in the bins, particularly with the observations at a greater distance from one another, and consequently one should proceed with caution during inference.

Cressie [40] argues that by construction, the empirical variogram is robust to the presence of trends (since only the differences are used), and that the estimator is unbiased. Although, according to Banerjee et al. basing an estimator on differences is not a direct measure of dependence. [36]

Choosing a variogram model is of importance as not all models are useful for a variogram. It is a particularly special type of function as it must be negative semi-definite and for 2nd order stationary processes it must reach upper bounds. It must also hold true that such a function must monotonically increase with an increasing lag, while having a constant maximum or sill, as well as a positive intercept (i.e. nugget). The most common parametric model used is the exponential variogram $(\gamma(t))$ and covariance function $\left(C(t)\right)$. These are expressed below, taking $t = \|s_i - s_j\|$;

$$C(t) = \begin{cases} \sigma^2 \exp\left(-\frac{t}{2}\right) & if\ t\ \geq 0; \\ \phi^2 + \sigma^2 & if\ t = 0, \end{cases} \qquad (3.2.3.5)$$

and

$$\gamma(t) = \begin{cases} \phi^2 + \sigma^2 \left(1 - \exp\left(-\frac{t}{\lambda}\right)\right) & if\ t\ \geq 0; \\ 0 & if\ t = 0, \end{cases} \qquad (3.2.3.6)$$

This is only one example of a parametric model which can be used for covariances and variograms, which give an initial idea of how spatial distance and relationships between the parameters change as a function of distance. There are different types of variogram models, different from the exponential one. This is discussed solely because it is the most common model. Other models include the spherical variogram model and the Gaussian variogram model.

3.2.4 Assessing Isotropy

To use a variogram, isotropy (defined as describing a variogram which depends only on distance and not on direction) must be assumed; although this is not always applicable as it may not be reasonable to assume isotropy. A directional variogram can be used to test this assumption as this combines multiple differently angled variograms into a single, unified variogram. Isotropy can then be assumed if each of these variograms follow the same trend [38].

There are numerous methods in the available literature which assess whether or not isotropy is a reasonable assumption for geostatistical data. The simplest method consists of, but is not limited to, restricting the pairs that appear in the empirical variogram so that only dependence in certain directions are measured [41]. These displays are called directional variograms. The directional variogram will look the same irrespective of the direction analysed, if the process is isotropic.

3.2.5 Monte Carlo tests

There exists a simulation-based method for the assessment of evidence for support of different hypothesis, known as Monte Carlo tests [44]. The comparison of a test statistic with a number of statistics computed from the null hypothesis is what is essentially involved in this method.

One way of assessing spatial correlation is to plot the semi-variogram, and overlay on top the upper and lower limits for the set of semi-variograms that would have occurred under independence. These limits are computed using Monte Carlo tests and are also known as Monte Carlo envelopes. If the estimated semi-variogram from the data lies completely inside the envelope, then the data contain no substantial correlation. This can be seen in the example below – since no points lie out with the upper or lower limits, the data can be said to have no substantial correlation. Figure 3.2.5.1:

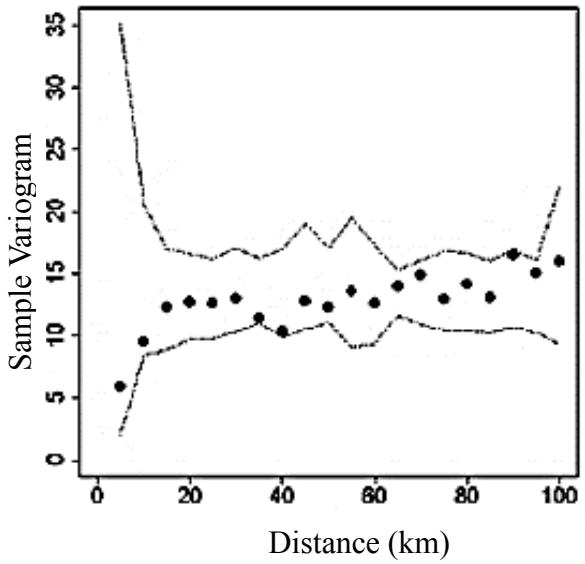


Figure 3.2.5.1: Example of data lying within Monte Carlo envelopes, or upper and lower limits. A semivariogram of this nature indicates that the data have no substantial correlation [45].

The Monte Carlo envelopes (or just envelops) are computed using the geoR package in R and the "variog.mc.env" command in R [74].

3.2.6 Multiple Covariates and regression models for the mean

Suppose there are *p* spatially varying covariates, such that

$$\{x_j(\underline{s}): \underline{s} \in D\}, j = 1, \dots, p. \qquad (3.2.6.1)$$

These covariates exist due to any number of reasons, some of these reasons are listed below;

- Other phenomena measured in space, such as traffic variables
- Functions of space, for example latitude and longitude (or Eastings and Northings) which are measured at a spatial location $\underline{s}$, and are equal to $x_2(\underline{s})$, $x_3(\underline{s})$ respectively.
- Non-linear functions of the spatial co-ordinates mentioned above.

The spatially varying mean can then be modelled by the regression:

$$\mu_z(s) = \sum_{j=1}^{p} \beta_j x_j(s) \qquad (3.2.6.2)$$

For each $s \in D$ where $x_1(s) = 1$. The parameters $\underline{\beta} = (\beta_1, \dots, \beta_p)$ can be estimated using least squares.

This estimation process should be seen as having limitations. The limitation highlighted here is that the data are assumed to be independent, which is unrealistic since a spatial correlation is expected. Spatial correlation is expected as observations made close to one another are more likely to be correlated with one another than observations made further away from one another.

3.2.7 Estimating Model Parameters – MLE and REML

Writing a regression model:

$$y(\underline{s}) = \underline{x}^T(\underline{s})\underline{\beta} + \underline{\varepsilon}(\underline{s}) \qquad (3.2.7.1)$$

and letting $(\underline{s}): \underline{s} \in D$ be a Gaussian geostatistical process with mean $\mu(\underline{s}) = \underline{x}^T(\underline{s})\underline{\beta}$, covariance $C_\theta(\underline{s}, \underline{t})$ and $\underline{\varepsilon}(\underline{s}) \sim N(0, \sigma^2)$. The likelihood of the data $\underline{y} = (y_1, \dots, y_n)^T$ at locations $x_i (i = 1, \dots, n)$ is explained by the following equation, given the mean parameters $\underline{\beta}$, covariance parameters $\underline{\theta}$ and where n equals the sample size and $\sum_{\underline{\theta}}$ is the covariance matrix of $y(\underline{s})$with (i, j) element $C_\theta(s_i, s_j)$:

$$L\left(\underline{\beta}, \underline{\theta}\right) = (2\pi)^{\left(-\frac{n}{2}\right)}(det\sum_{\underline{\theta}})^{-1/2} \exp\left(-\frac{1}{2}\left(\underline{y} - \underline{X\beta}\right)^T \Sigma_{\underline{\theta}}{}^{-1}\left(\underline{y} - \underline{X\beta}\right)\right) \qquad (3.2.7.2)$$

The log-likelihood is then calculated by taking the log of the expression $L\left(\underline{\beta}, \underline{\theta}\right)$:

$$l\left(\underline{\beta}, \underline{\theta}\right) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}(det\sum_{\underline{\theta}}) - \frac{1}{2}\left(\underline{y} - \underline{X\beta}\right)^T \Sigma_{\underline{\theta}}{}^{-1}\left(\underline{y} - \underline{X\beta}\right) \qquad (3.2.7.3)$$

Minimising $l\left(\underline{\beta}, \underline{\theta}\right)$ and calculating the derivative with respect to $\underline{\beta}$ leads to obtaining the GLS or Generalised Least Squares. Clearly, the MLE of $\underline{\beta}$ is dependent on the spatial parameters $\underline{\theta}$, and can be used in the equation for log-likelihood (seen above) and maximised with respect to $\underline{\theta}$. It should be noted that this method may introduce bias in θ, due to the estimate of $\underline{\beta}$. An alternative approach to this is to use the restricted maximum likelihood (REML), which minimises the bias when estimating the parameter $\underline{\theta}$ [35].

Using Restricted Maximum Likelihood (REML) as an approach is using a form of maximum likelihood estimation, as it also requires that y follows a multivariate normal distribution. This method is used to estimate the nugget, sill and the range which are denoted by the spatial model parameters in $\underline{\theta} = (\phi^2, \sigma^2, \lambda)^T$. One difference between using REML and MLE is that REML allows the user to ensure less biased estimates of $\underline{\theta}$, as it calculates the likelihood function from a transformed dataset, ensuring that the nuisance parameters have no effect on the estimates. When estimating model parameters, one model of interest is the Gaussian random fields model. The Gaussian random fields model is defined below;

$$Y(s) = \mu(s) + Z(s) + \varepsilon \qquad (3.2.7.4)$$

Using this model $E[Y] = X\beta$, the data can be transformed linearly to $Y^* = AY = X(X^TX)^TX^TY$, where Y^* does not depend on β [39]. Following the transformation, the model remains multivariate Gaussian. Y^* not depending on β means that the dimensions of y is reduced from n to $n - p$, with p denoting the rank of X. Maximising the likelihood for $\underline{\theta}$ based on Y^*, the REML estimates for $\underline{\theta}$ are computed. $\underline{\theta}$ is, as mentioned above (3.2.7), the covariance parameters.

3.2.8 Spatial prediction (Kriging)

The use of the word "Kriging" in spatial statistics has come to be synonymous with "optimal prediction" in space, using observations taken at known nearby locations. Linear and non linear observations are used for predicting, using kriging as a "method of interpolation for a random spatial process" [65]. Kriging was originally a linear predictor whereas in more recent developments in geostatistics methods of optimal nonlinear spatial prediction have become part of the kriging [66]. Practically implementing Kriging methods come in the form of estimating the variogram, which is discussed previously (section 3.2.3). From [66], Ordinary Kriging can be described as the following;

"Prediction based on (ordinary) kriging is equivalent to spatial blup. The predictor minimizes the mean-squared prediction error over all linear unbiased predictors, for a given covariance function $C(.,.)$. "

The spatial blup here is an abbreviation for the best linear unbiased predictor and the covariance function is of the kind discussed in section 3.2.2. Ordinary Kriging in a little more detail is seen

in [66] as the following. These sentences and equations are mathematical explanations of the spatial prediction which takes place later in the chapter. They explain unbiased predictors, and how to find the best predictor.

By restricting the class of linear predictors to the so-called homogeneous linear predictors;

$$\sum_{i=1}^{n} \lambda_i S(x_i) \qquad (3.2.8.1)$$

Further restriction of uniform unbiasedness yields the condition;

$$\sum_{i=1}^{n} \lambda_i = 1 \qquad (3.2.8.2)$$

Thus, one could look for the best linear unbiased predictor (blup), obtained by minimizing;

$$E(S(x_0) - \sum_{i=1}^{n} \lambda_i S(x_i))^2 \qquad (3.2.8.3)$$

Over $\lambda_1, \dots, \lambda_n$, subject to equation 3.2.8.2.

By the method of Lagrange multipliers, the optimal values are

$$\lambda' = (c + (1 - c'C^{-1}1)(1'C^{-1}1)^{-1}1)'C^{-1} \qquad (3.2.8.4)$$

where c and C are given as $c \equiv (C(x_0, x_1), \dots, C(x_0, x_n))'$ and $C \equiv \left(C\left(x_i, x_j\right)\right)$ respectively. Thus the optimal linear predictor of $S(x_0)$ is;

$$\hat{S}(x_0) = c'C^{-1}S + (1 - c'C^{-1}1)(1'C^{-1}1)^{-1}(1'C^{-1}S) \qquad (3.2.8.5)$$

The ordinary kriging predictor. The mean squared prediction error is;

$$E(S(x_0) - \hat{S}(x_0))^2 = C(x_0, x_0) - c'C^{-1}c + (1 - c'C^{-1}1)^2(1'C^{-1}1)^{-1} \qquad (3.2.8.6)$$

3.3 Spatial Trend Analysis of Annual Mean NO_2 Data

The spatial distribution of NO_2 data collected at AURN and diffusion tube locations throughout Aberdeen in 2014 is explored in this section. A preliminary idea of the minimum and maximum values as well as the spatial trend of NO_2 across the city is given in the table 3.3.1. This table gives a summary of the log transformed values of NO_2 at 51 locations in Aberdeen.

Min	1st Quantile	Median	Mean	3rd Quantile	Max	Std. Dev
2.351	3.292	3.503	3.503	3.820	4.059	0.392

Table 3.3.1: Summary Statistics of log NO_2 values throughout Aberdeen

The following histogram shows the distribution of the distances between the locations where NO_2 was recorded.

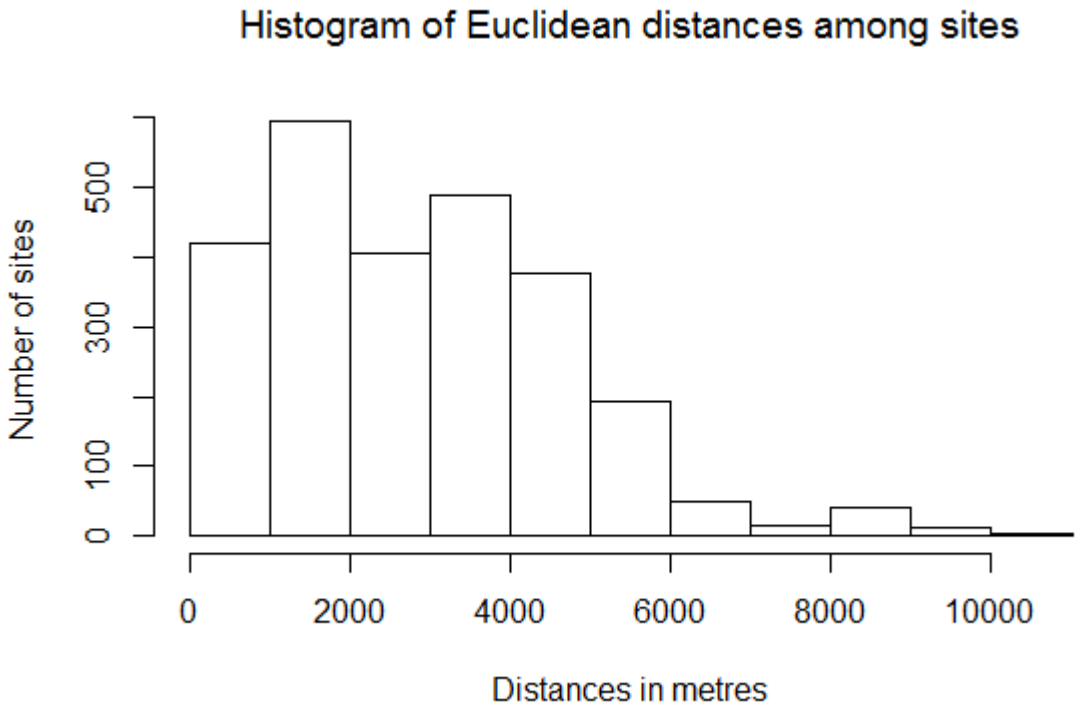


Figure 3.3.1: Histogram showing the frequency of the distances at which NO_2 concentrations are recorded

In this Chapter, the data which are looked at are from 2014, and the data are measured at a number of locations throughout Aberdeen. Again, as with the previous Chapter, NO_2 is taken as the response variable, with meteorological and traffic variables taken as the explanatory variables. In addition to the traffic and meteorological variables, there are Eastings and Northings which are used to describe the monitoring sites location. Since there are not traffic

counts available at the exact locations of the monitoring sites for NO_2, an approximation was made using two separate maps, with two separate data sets – one consisting of the NO_2 monitoring sites (included as AURN and diffusion tubes), and the other consisting of traffic counts, known as count points. The Eastings and Northings included in a model are the coordinates of the NO_2 monitoring sites, while the traffic counts are taken from the nearest, or most appropriate, as measured by Euclidean distance, count point to the relative monitoring site. This distance is a limitation to the analysis. The map with the count points which was used can be found at [46], and it was this map which was used in conjunction with [47] (using the Eastings and Northings of the diffusion tube and AURN sites) to match traffic counts with monitoring sites. It is expected that the observed points which are closer together will be more alike than the observed points which are farther away from one another – this a basic concept in geography [48].

The image below (image 3.3.1) shows a sub-section of a larger map which has NO_2 monitoring sites and count points for monitoring traffic. The image shows the limitations of the locations of the monitoring sites i.e. the NO_2 and traffic monitoring sites are in different locations meaning the NO_2 recorded doesn't match spatially exactly with the traffic recorded.

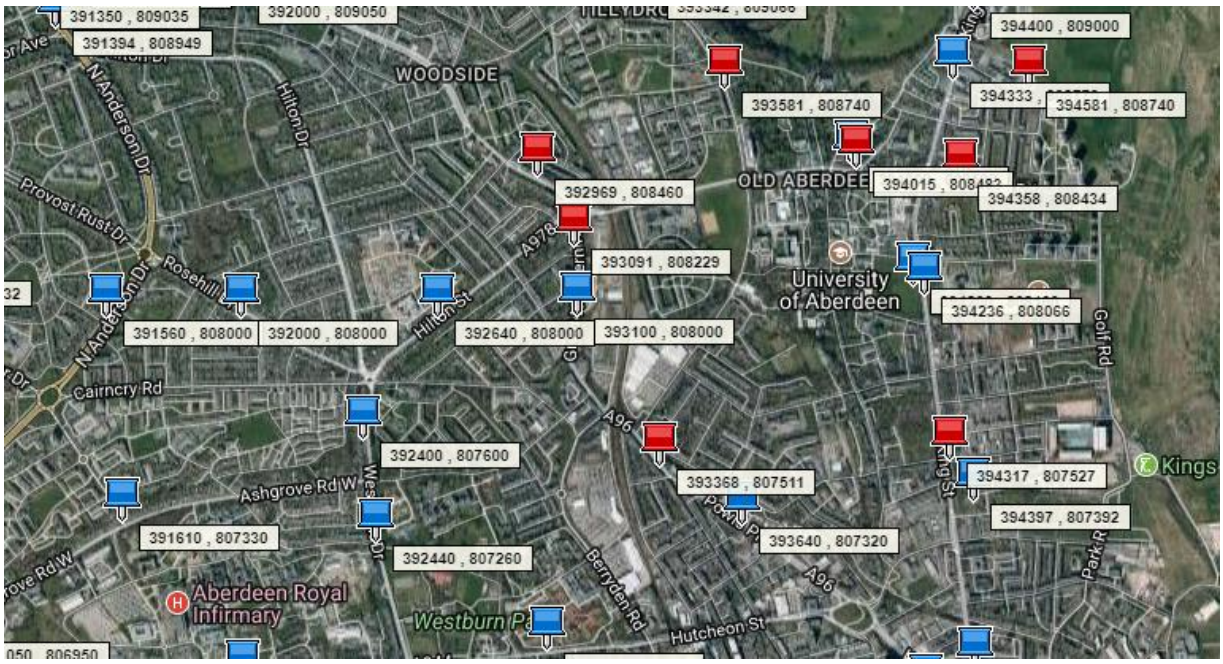


Image 3.3.1: Sub – section of full map which shows NO_2 monitoring sites (in red) and count points for traffic (in blue). This map shows that sites are located at relatively far distances in some cases and so is a limitation when modelling the data.

Unlike the previous Chapter, the data are recorded as an annual Figure, as space is of primary interest here, as opposed to time. This annual Figure is obtained as a mean NO_2 value for each location, with meteorological Figures being consistent from site to site, as the only location monitoring weather was Dyce airport – no other monitoring location for meteorological variables exist.

As will be shown later, the weather variables are in fact degenerate, as they are repeated values for each location and can be described as singularities. This is a slight problem in the spatial modelling phase, although the model can still be built, only without meteorological factors.

3.3.1 Exploratory spatial analysis

Using the "geoR" library in R [49], geostatistical analysis of the NO_2 concentrations in Aberdeen is possible. Useful questions to ask when conducting spatial analysis include the following; what is the average NO_2 concentration across Aberdeen? What is the spatial pattern in the NO_2 concentration across Aberdeen, that is which areas have high NO_2 concentrations and which have low NO_2 concentrations? What is the difference between the minimum and maximum NO_2 concentrations in Aberdeen? Looking at the spatial locations of the NO_2 monitoring sites, using the Eastings and Northings, we can see that they are not spatially distributed uniformly. The exploratory plots of the log NO_2 data with the Eastings and Northings is below 3.3.1.1;

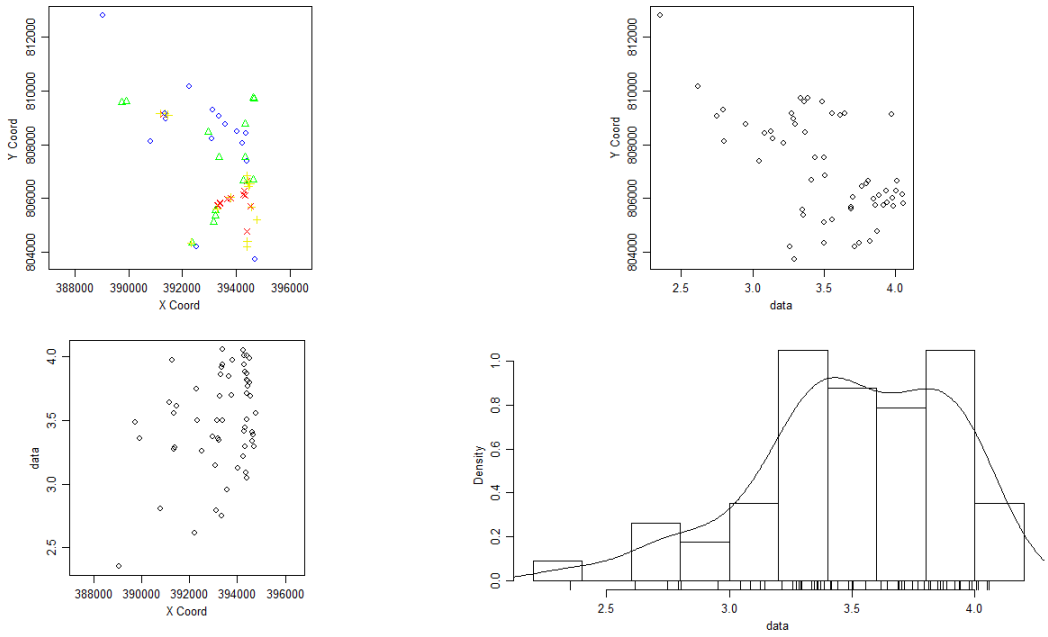


Figure 3.3.1.1: Initial exploratory plots of NO_2 concentrations*, their geographical locations, and a histogram showing the density of the distribution of* NO_2 concentrations.

Looking at the top left panel it can be seen that these monitoring sites have different colours. These colours represent the different concentrations of NO_2. The following colours represent NO_2 concentrations from highest to lowest concentrations – red, yellow, green, blue. Looking at the top right hand plot, there is a decreasing linear trend with the NO_2 and Y coordinate, with an increasing variance i.e. going from North to South in Aberdeen the NO_2 concentration decreases. Looking at the plot in the bottom left hand side, it shows that going from West to East in Aberdeen there is not a distinguishable change in the concentrations of NO_2, with the possibility of a slight increase in NO_2 further East, although there are more points to the East of the central point of all monitoring sites. When additional covariates are included in the model there needs to be an additional question asked of the model – what impact do these covariates have on NO_2 concentration in Aberdeen?

3.4 Spatial Trend Analysis of traffic data

It is of use to explore the traffic data in a spatial context to see how it relates to the NO_2 data from the AURN and diffusion tube monitoring sites, as well as to get an idea of the traffic data recorded at the count points in a spatial context. This can be seen from the histograms, and scatterplots, which are shown in Figures 3.4.1 - .2. Since the vehicle classes consist of more than just the total number of motor vehicles, two models and two methods of exploratory analysis

need to be implemented. One analysis takes place consisting of "All motor vehicles" at each of the air quality monitoring sites. The All Motor Vehicles class consists of 7052 vehicles. When other classes of vehicles are included in the analysis (for example Buses and Coaches, All HGVs, Light Goods Vehicles) the all motor vehicles class needs to be adjusted so that vehicles are not counted twice i.e. the 7052 units of the All Motor Vehicles class are broken down into other classes as well. This is shown in the example below (table 3.4.1);

Location	Buses and Coaches	LGVs	All HGVs	All Motor Vehicles
Errol Place	478	890	595	5089

Table 3.4.1: The number of vehicles for each class at Errol Place in 2014. This is the annual average daily flow, which is discussed in more detail later.

Focussing on the spatial aspect of the traffic data, an exploratory analysis shows the following results in Figure 3.4.1. These plots in Figure 3.4.1 show the total number of vehicles in relation to the locations of the AURN and diffusion tube sites. They also show (in the top left hand plot) where the most and least vehicles are recorded, going from high to low they are red, yellow, green and blue. From the top right and bottom left plots we can tell that, on a map, the locations with the highest number of vehicles would be found in the South-Eastern part of the map.

The histogram showing the density of the total number of motor vehicles is seen in Figure 3.4.1, including a curve depicting the density of the data. Another plot of interest is the log NO_2 values plotted against the total number of motor vehicles, regardless of class. One can conclude that there is a weak relationship. This is seen in Figure 3.4.2;

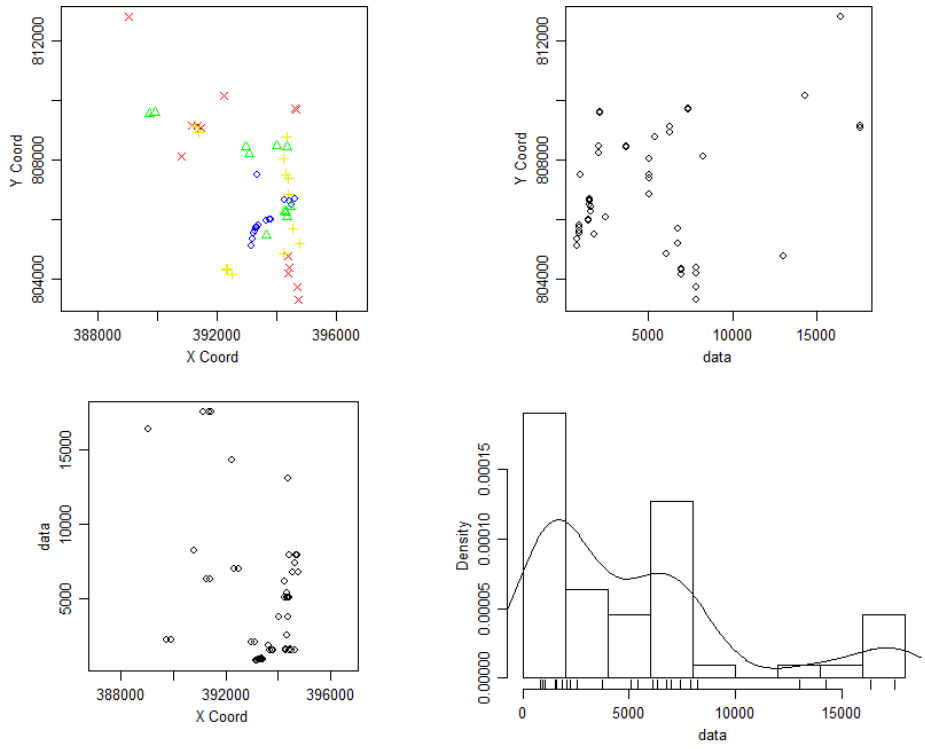


Figure 3.4.1: Summary plots of Motor Vehicles – the number and location of. Also included is a histogram showing the density of the motor vehicle data.

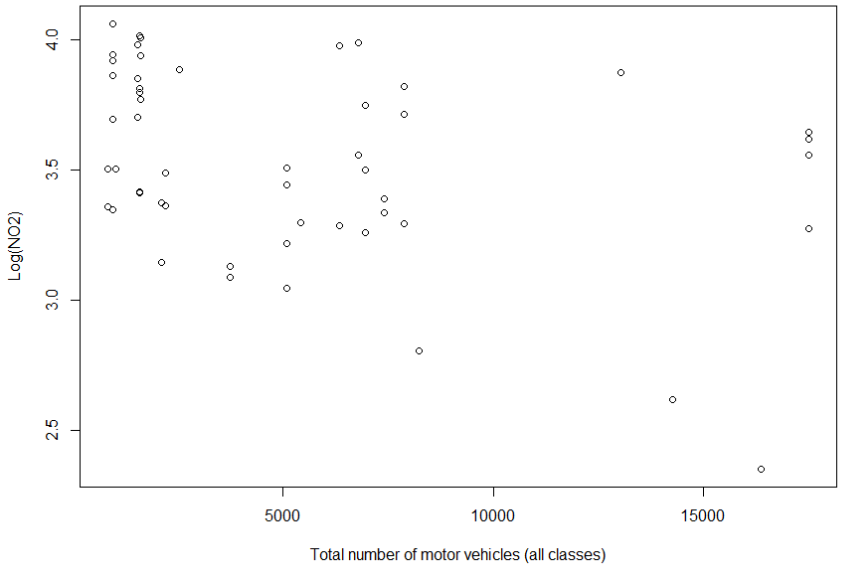


Figure 3.4.2: Total Number of motor vehicles, regardless of class vs the log values of NO_2. These values correspond to the year 2014 at all locations in Aberdeen for the AURN and diffusion tube sites.

The traffic values, for the different classes are measured as AADF, or annual average daily flow. This is the average in a full year of the number of vehicles passing a point on a road in both directions. This can be seen more clearly in the image below, which is taken from a webpage related to road traffic estimates on the UK government website [50].

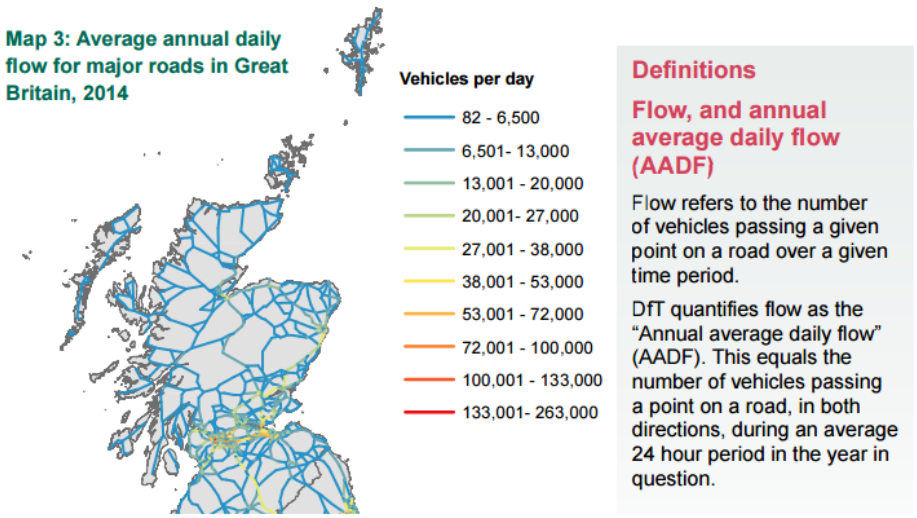


Figure 3.4.3: Map showing the average annual daily flow for major roads in Scotland, including definitions [50].

3.5 Spatial Trend Estimation of the NO_2 Data

A more formal exploration of the NO_2 data takes place in this section, in order to explain clearly the spatial distribution of NO_2. This model, initially, will be a simple one, consisting of only the response (log transformed NO_2 values); the intercept; the covariates (latitude and longitude represented as Eastings and Northings respectively). Later in the Chapter more covariates are introduced, these include meteorological and traffic count variables, which lead to a second model.

3.5.1 Initial Model

The model is explained in more detail below;

$$\log y(s_i) = \beta_0 + \beta_1 easting(s_i) + \beta_2 northing(s_i) + \varepsilon(s_i) \qquad (3.5.1.1)$$

where $y(s_i)$ corresponds to the value of NO_2 at each spatial location s, where $i = 1, \ldots, 56$; each β are regression parameters and $\varepsilon(s_i)$ represents the residuals which are assumed to be normally distributed such that $\varepsilon(s_i) \sim N(0, \sigma^2)$.

After fitting the linear model above the Easting term is removed as it is not statistically significant, the analysis of the summary statistics of the model are that the Northing term is statistically highly significant; approximately 34% of the variance in the model is explained as the adjusted R^2 term is 0.3393; and the model itself is statistically significant as the p-value is less than 0.05. These can be seen in the table 3.5.1.1;

	Estimate	**Standard Error**	**p-value**
Intercept	9.124e+01	3.114e+01	0.00496
Northing	-1.156e-04	2.631e-05	5.25e-05
Easting	1.412e-05	3.608e-05	0.69711
AIC value		37.1384	
R^2 Adjusted value		0.3393	

Table 3.5.1.1: estimates, standard errors and p-values of intercept Northing and Easting, with measures of goodness of fit

Taking the residuals from the initial model, that is the model with both the Northings and Eastings included, the following residual plots can be obtained. The plots include the residuals against the Eastings, the residuals against the Northings, as well as the Normal Q-Q plot. These highlight how well (or not so well, as the case seems to be) the model estimates the NO_2 concentrations at the recorded sites. The Easting term is not statistically significant although I decided to leave it in the model as it describes the location of each monitoring site.

It is clear from this Normal Q-Q plot (Figure 3.5.1.1) that the residuals do not, in fact, follow a normal distribution as there is deviation from the Q-Q line. Although the data follows the line almost entirely, there is some deviation in the tails. This may be the case because the data is not entirely normally distributed, there is some skewness in the tails. There is not much more that can be done to normalise the data, as a log transformation of the data has already taken place.

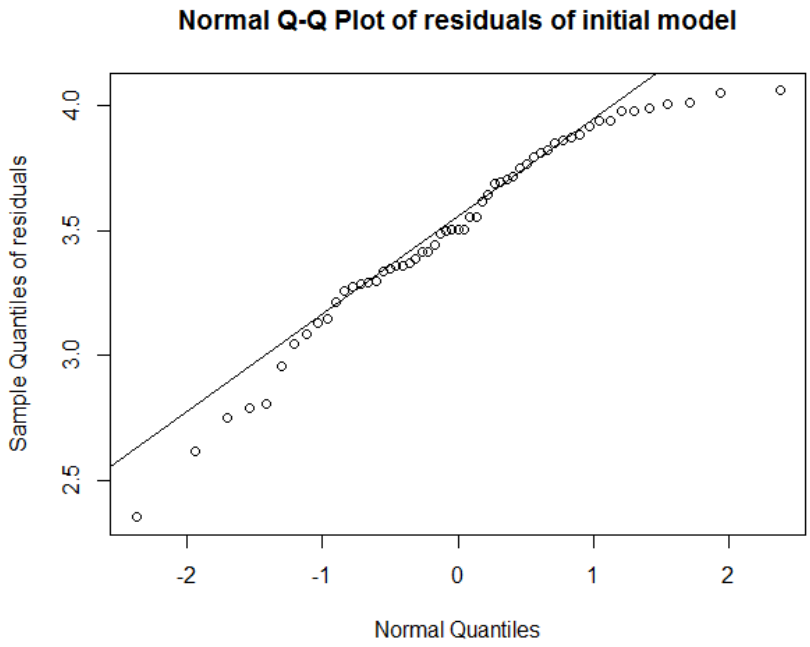


Figure 3.5.1.1: Normal Q-Q plot of the residuals of the initial model in Aberdeen

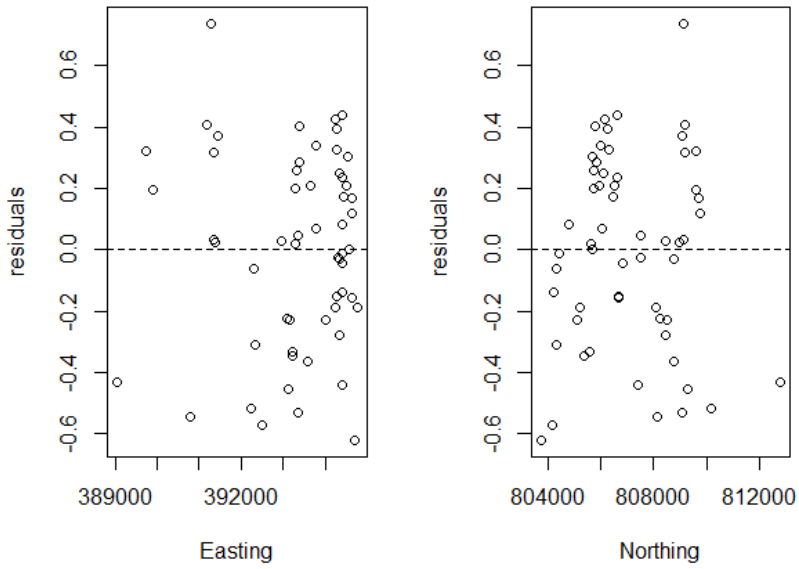


Figure 3.5.1.2: Residual plots of the initial model against the Eastings (left panel) and Northings (right panel)

The residual plots in Figure 3.5.1.2 show that there may be some bias in the data and heteroscedasticity may be present. Bias in the model means the independence assumption of the observations may be violated. The heteroscedasticity means that there is not constant variance.

3.5.1.1 Estimating empirical variogram for residuals

Estimating the semivariogram with a variogram cloud initially gives an indication of the spatial structure of the NO_2 as measured at AURN and diffusion tube sites, although it is quite noisy, and hence difficult to interpret, as can be seen below;

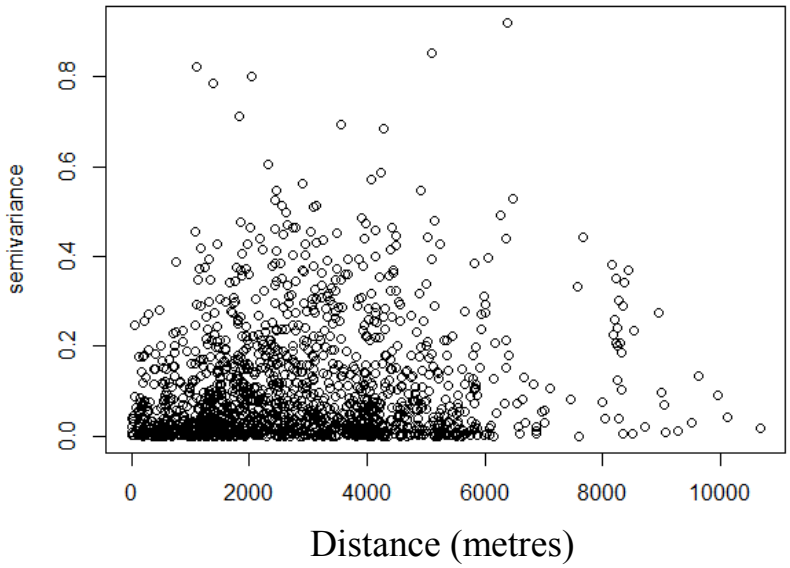


Figure 3.5.1.1.1: Variogram cloud for model containing only Easting and Northing

As expected, observations which are closer together have a smaller semivariance and points which are further to the right of the x-axis have a higher semivariance. This is more of the rule rather than the exception as there are quite a few points which are between the distance of 2000 and 9000 which have a relatively high semivariance and a mid – range distance.

Using the empirical variogram to gain a better idea of the spatial structure of the data, an estimate can be produced. Although, due to the limited amount of sites, how "good" this estimate is, is questionable. This leads to the use of a binned empirical variogram, which is considered to be more robust [35].

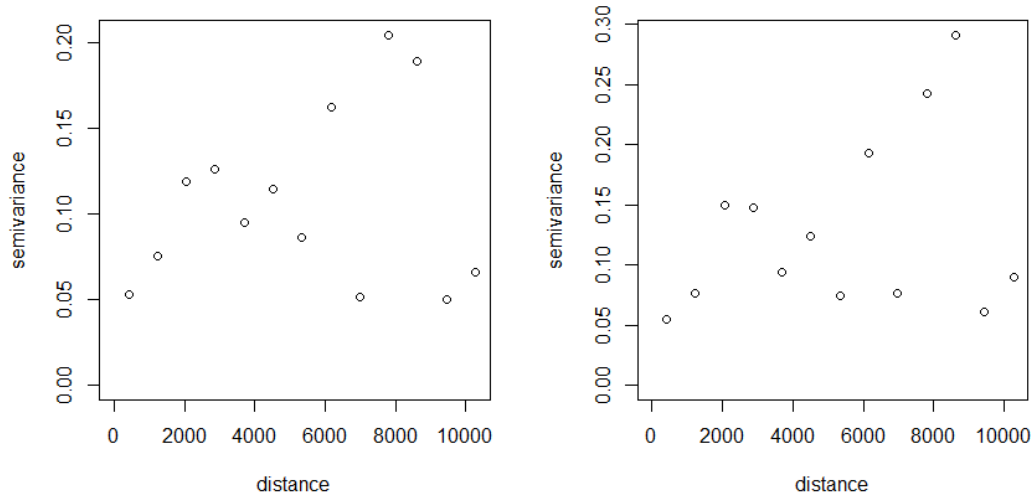


Figure 3.5.1.1.2: Empirical variogram (left panel) and more robust binned empirical variogram (right panel) for model containing Eastings and Northings only

From looking at both variograms above, the presence of trend is obvious, as neither seem to stabilise anywhere on the plot, and definitely not anywhere close to the sample variance. This leads on to assessing the presence of isotropy. The semivariance seems to be oscillating in both variograms, with increasing variance as distance increases. Further work could be to use the median instead of the mean in each of the bins, as the median is a more robust estimator than the mean. Furthermore the bin size could be varied so that a stronger mean may be obtained i.e. noise in the model may be reduced.

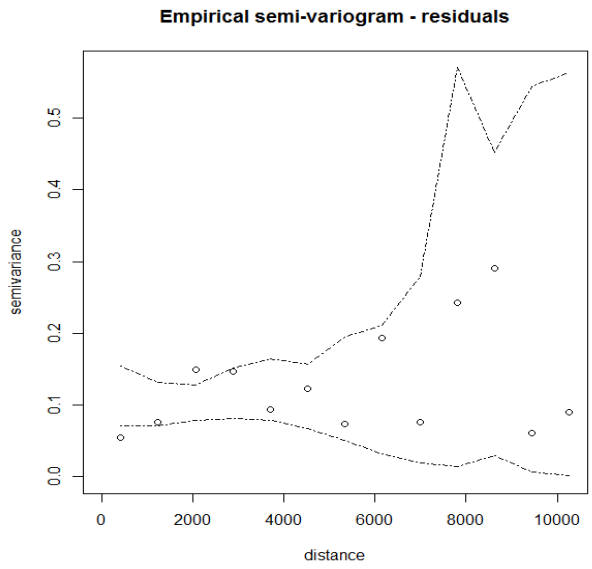


Figure 3.5.1.1.3: Empirical variogram with Monte Carlo envelopes for model containing Eastings and Northings only as covariates

The previous Figure (3.5.1.1.3) has values lying outside of the Monte Carlo envelopes. This suggests that there is spatial correlation between observations, which is explained in section 3.2.5. There could also be the case that the points lying outside of the Monte Carlo envelopes are lying outside of it by chance and there is only very limited spatial correlation, if any at all.

3.5.1.2 Estimating model parameters

This section starts by estimating the variogram model parameters for an exponential model. Selecting an exponential model (or any model) influences the prediction of the unknown values. This is more the case if the curve when near the origin is relatively steep. A steeper curve near the origin means the closest neighbours have a greater influence on the value of the prediction. Estimating the model parameters initially by using weighted least squares results in the following (Figure 3.5.1.2.1), and showing the fitted variogram over the binned estimator (Figure 3.5.1.2.2);

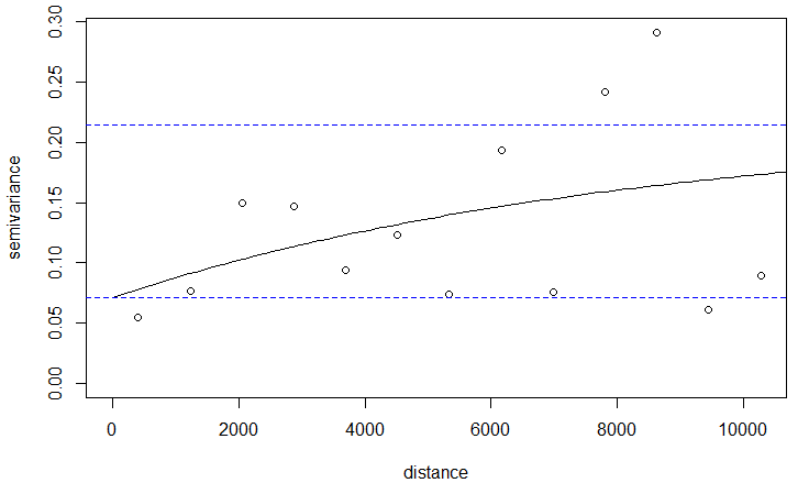


Figure 3.5.1.2.1: Fitted variogram over robust binned estimator with nugget (bottom line) and sill (top line) included for model containing Easting and Northing coordinates exclusively

The semivariogram above shows that there is a progressive decrease in spatial autocorrelation i.e. an increase in semivariance, almost continually with a slight arc shape appearing. This shape, and model, can be described as exponential.

Estimating the covariance parameters using maximum likelihood, while assuming the residuals are Gaussian, followed by estimating the covariance parameters using restricted maximum

likelihood, the following plot (Figure 3.5.1.2.2) is obtained. This shows the fitted maximum likelihood and the restricted maximum likelihood based variograms over the robust estimator;

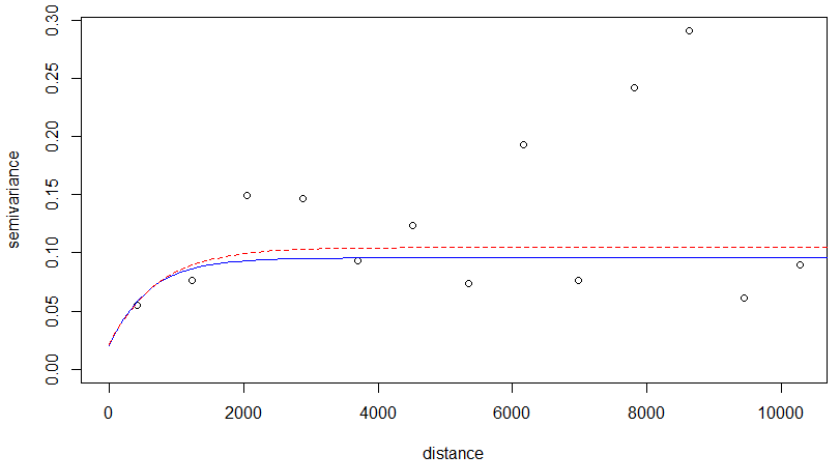


Figure 3.5.1.2.2: Fitted ML (blue) – and REML (red) – based variograms over the robust binned estimator

The actual parameter estimates for the exponential model are summarised in the table below;

Model	Nugget	Partial Sill	Range
MLE	0.0204	0.0754	596
REML	0.021	0.0837	726.5

Table 3.5.1.2.1: Covariance parameters for the exponential model, for both the MLE and REML methods.

3.5.1.3 Spatial prediction (Kriging)

In this section predictions of the NO_2 field are shown using an exponential variogram. Firstly, the ordinary kriging predictor is shown, (as discussed in section 3.2.8) followed by its standard error. This is done first for the model containing only Easting and Northing as explanatory variables, and in a later section kriging for the model containing all explanatory variables. The

kriging procedure is often described as optimal [51] as it produces optimal predictions when the covariance structure is known. For parameter estimates for model, and measures of goodness of fit see table 3.5.1.1.

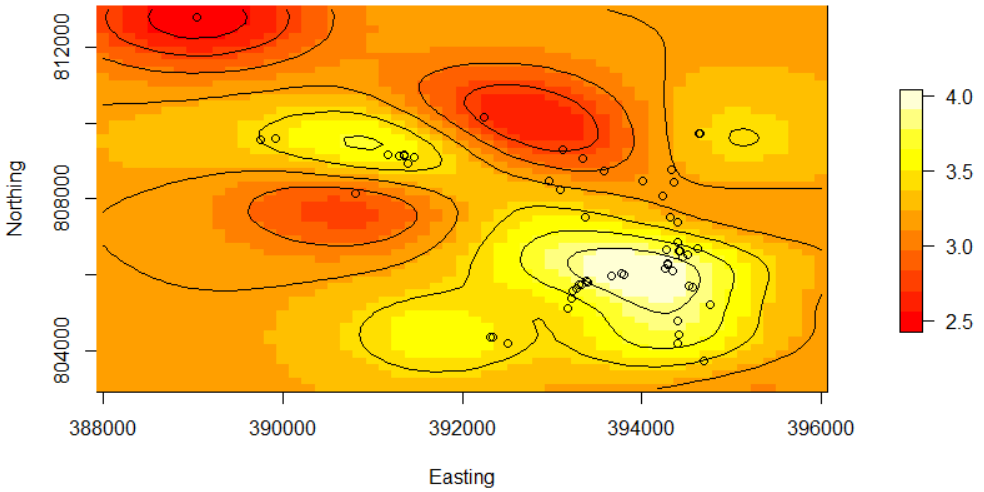


Figure 3.5.1.3.1: Predicted field of NO_2 values for simple model

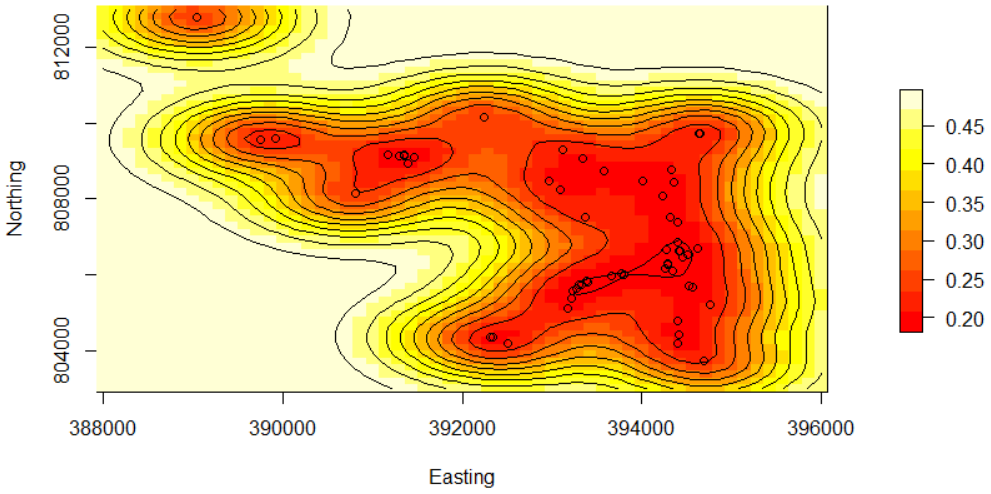


Figure 3.5.1.3.2: Standard errors of predictions for simple model

Figure 3.5.1.3.1 shows that the log NO_2 concentration is predicted to be highest closest to the city centre and in the North West of the map, further away from the city centre. The concentrations of NO_2 may also be predicted to be higher in the city centre because there is more traffic or more congested traffic. Also, it should be noted that there are more recording instruments in the city centre and so the recordings may be more accurate compared to the recordings made outside of the city centre.

It is clear from the field of standard errors of predictions that there is more uncertainty where there are fewer sites and more certainty where there are more sites, which is to be expected, intuitively. Although the expected outcome of locations closer to one another being more similar in NO_2 concentrations is tested here as it seems there are groupings of predicted field values which are the same as other, not necessarily neighbouring, groups. This can be seen in Figure 3.3.2.1.8, where the most North-West and South-East groups are similar but are at opposite ends of the City. These are the areas with the lowest predicted NO_2 concentration.

3.5.2: Full linear model, including all covariates

Including the meteorological variables and traffic variables in the model it is shown that the meteorological variables are redundant as they are singularities, i.e. the variables become degenerate. This is due to the same value being repeated at each site, for any given meteorological variable. This is an unavoidable issue for the data at hand as the weather recorded at Dyce airport is the only meteorological data available for Aberdeen in 2014. So, even though weather is probably important, it has to be left out of this model. This is discussed further in the final Chapter. The spatial model including all of the variables can be seen below;

$$\begin{aligned}\log y(s_i) = {} & \beta_0 + \beta_1 easting(s_i) + \beta_2 northing(s_i) + \beta_3(Buses\ and\ Coaches)(s_i) \\ & + \beta_4(Light\ Goods\ Vehicles)(s_i) + \beta_5(All\ HGVs)(s_i) \\ & + \beta_6(All\ Motor\ Vehicles)(s_i) + \varepsilon(s_i)\end{aligned}$$

Equation 3.5.2.1

Standard errors, p-values and estimates can be seen for the variables in the table below;

Variable	Estimate	Standard Error	p-value
Intercept	7.498e+01	3.331e+01	0.02901
Easting	1.143e-05	4.173e-05	0.78539
Northing	-9.401e-05	3.127e-05	0.00419
Buses and Coaches	2.302e-04	3.550e-04	0.51977
Light goods vehicles	-4.241e-04	2.190e-04	0.05872
All HGVs	2.908e-04	3.284e-04	0.38022
All motor vehicles	1.994e-05	3.305e-05	0.54918

Table 3.5.2.1: The estimates, standard errors and p-values of all variables for the model

From the p-values it can be seen that the statistically significant variables are limited to the Northing and Light goods vehicles. This full model, has the following diagnostics, and can be seen to be a better fit to the data than the model with only the Northings and Eastings as covariates – this can be seen given the normal Q-Q plot and the plot of the residuals. The residuals plot of the model show random scatter around the dashed line (zero line). The normal Q-Q plot (Figure 3.5.2.2) shows the residuals following a straight line, mostly, as the points deviate from the line a little in the middle, and at the tails. The adjusted R^2 value for this model is 0.29, so the model does not explain much of the variation in log NO_2.

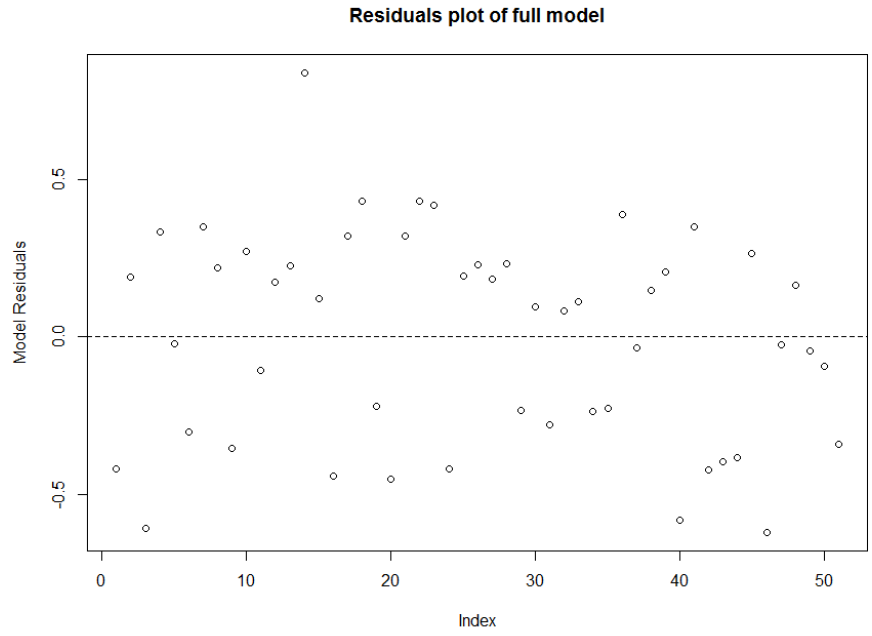


Figure 3.5.2.1: Residuals plot from the full model, showing a distribution which seems to be following a pattern which is not random.

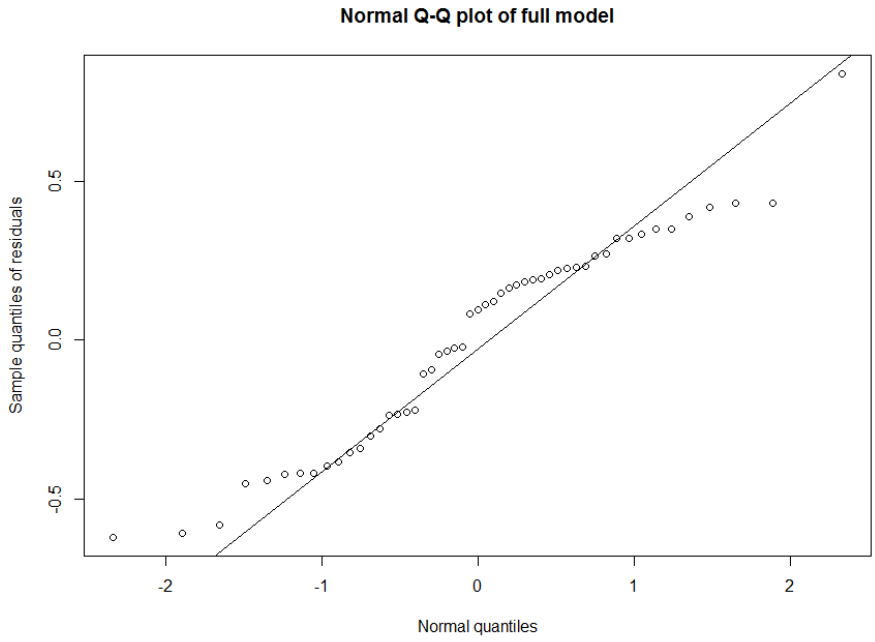


Figure 3.5.2.2: Normal Q-Q plot of full model. Follows the normal line mostly but deviates at the tails.

3.5.2.1: Estimating empirical variogram for residuals

The semivariogram cloud for the residuals from the full model of all of the sites can be seen below (Figure 3.5.2.1.1);

Empirical variogram and binned empirical variogram for the model containing all of the covariates available can be seen below (Figure 3.5.2.1.2). This is different to the model containing only Eastings and Northings as there is a smaller maximum semivariance for this model. This can also be seen from the empirical variogram with Monte Carlo envelopes below (Figure 3.5.2.1.3).

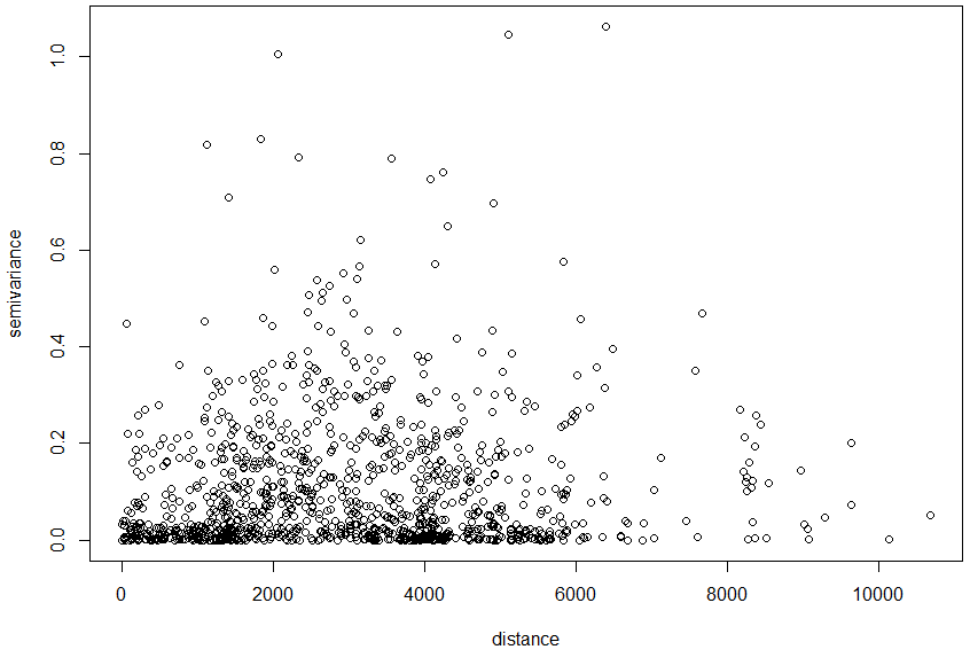


Figure 3.5.2.1.1: Variogram cloud for full model containing all covariates

The following Figure (Figure 3.5.2.1.2) shows an empirical variogram and the more robust binned empirical variogram for the full model. As can be seen from these panels, the two methods used for estimating the empirical variogram compute similar sample variograms.

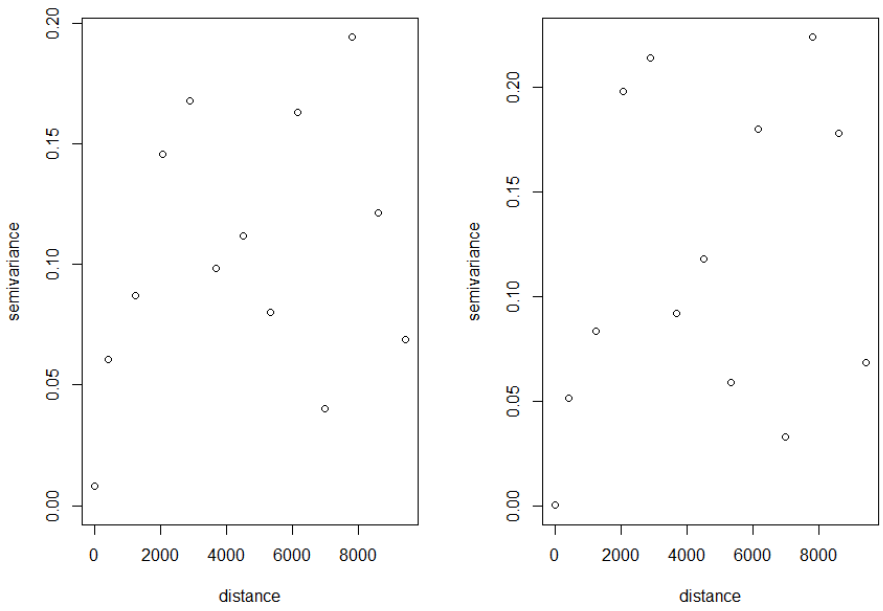


Figure 3.5.2.1.2: Empirical variogram (left panel) and more robust binned empirical variogram (right panel) for model containing all covariates available

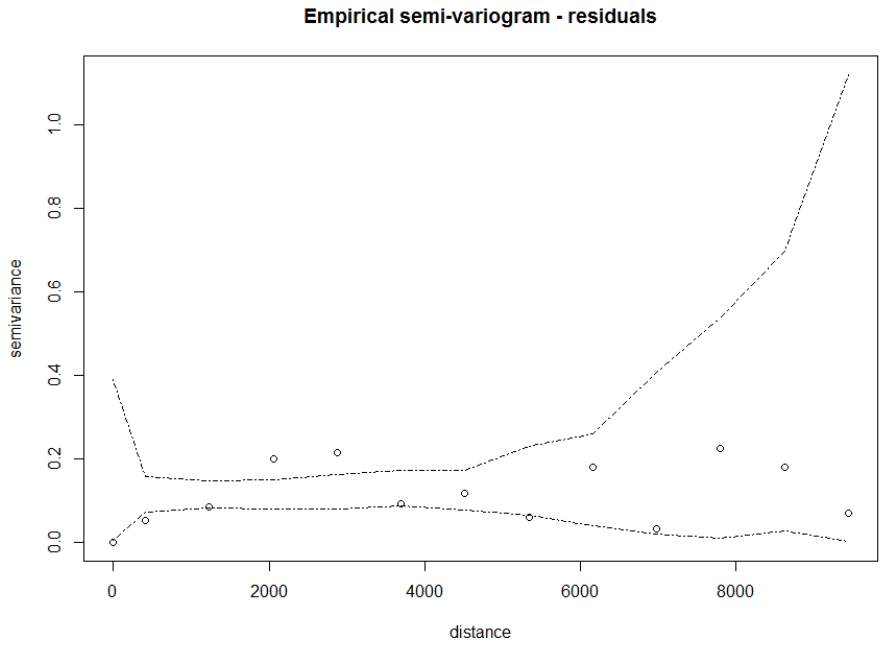


Figure 3.5.2.1.3: Empirical variogram with Monte Carlo envelopes for model containing all covariates available

3.5.2.2: Estimating model parameters

Since the empirical variogram with Monte Carlo envelopes shows evidence of some spatial dependence, it could be argued that a simple variogram model could be chosen. The model parameters are estimated for the model containing all covariates. This can be seen in the following plots, which show fitted variogram (Figure 3.5.2.2.1) which is exponential and the fitted maximum likelihood and restricted maximum likelihood based variograms (Figure 3.5.2.2.2). The semivariogram model is very uncertain given the data available.

The maximum likelihood and the restricted maximum likelihood are seen below (Figure 3.5.2.2.2) and these show the estimated values of the parameter at different distances. The maximum likelihood is represented by the blue solid line while the restricted maximum likelihood is represented by the dashed red line. Between 0 and 2000 the curve takes on a logarithmic shape and after of which it levels off somewhere around the 0.1 mark for semivariance.

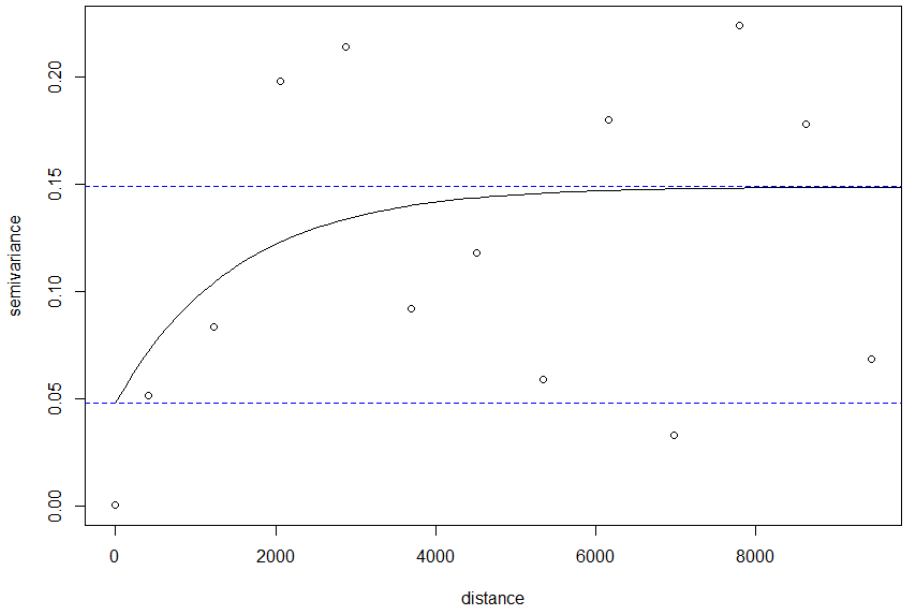


Figure 3.5.2.2.1: Fitted variogram over robust binned estimator with nugget (bottom line) and sill (top line) included for full model containing all covariates

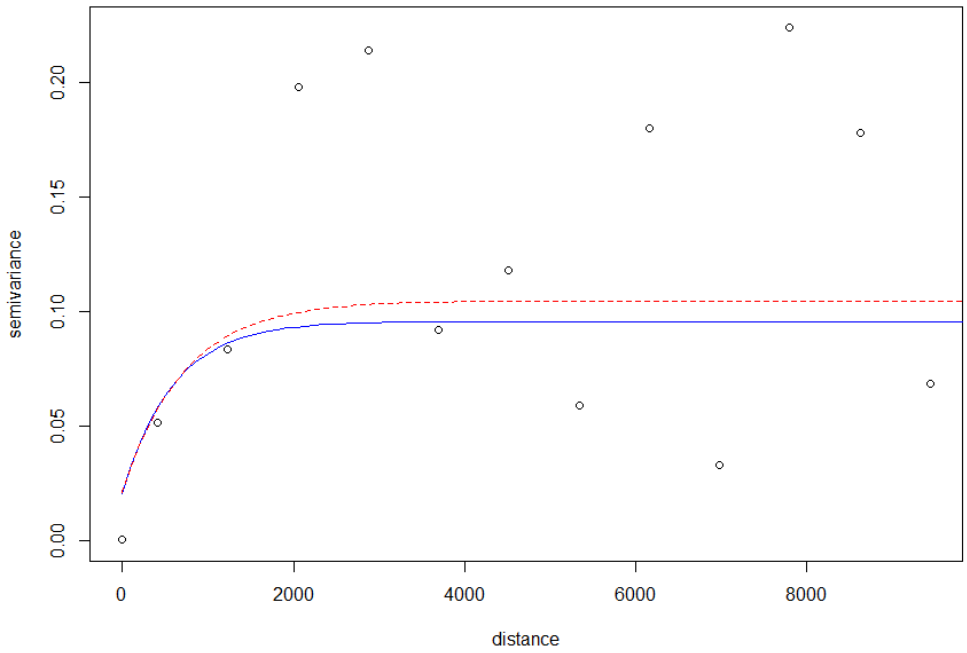


Figure 3.5.2.2.2: Fitted ML (blue) – and REML (red) – based variograms over the robust binned estimator for the full model, containing all covariates

3.6: Replacing explanatory variables with emission factors for 2014

Now emission factors are going to be introduced, since it is the emissions which are of concern, more so than the total number of vehicles. Emissions, which may be precursors of air pollutants, have an adverse effect both on the environment and human health. According to [67] "The principal air-quality pollutant emissions from petrol, diesel, and alternative-fuel engines are carbon monoxide, oxides of nitrogen, un-burnt hydrocarbons and particulate matter". The emission focused on here is principally oxides of nitrogen, more specifically NO_2. Emissions are also described in [68] as "gases or particles which are put into the air by various sources". One of these various sources is vehicles, which is what is analysed in this section – vehicle emissions. In this section, rather than vehicle counts emission factors are used, which are available from the Department for Environment, Food and Rural Affairs website. These figures for the emission factors are the same unit of volume and are available for different years (including 2014) and for different vehicle classes. These emission factors are obtained from the naei website [52]. The emission factors used in the models in the next section (section 3.7) are created by taking the respective emission factor for each vehicle classes and dividing it by the

emission factor for buses and coaches. The emission factor for buses and coaches is used as a base line. These figures are then multiplied by their respective vehicle count. The reason for doing this is so that all of the emissions are on the same scale and can be used in a model. These values are shown in table 3.6.1;

Vehicle class	**Buses and Coaches**	**LGVs**	**All HGVs**
Emission factor	0.118	0.424	0.115

Table 3.6.1: Emission factors for different vehicle classes. These are the fractions of Nitrogen Oxides emitted by vehicles as NO_2

The model used here is a linear one, with the emissions factors as well as Northings and Eastings used as covariates, with an intercept term, parameter coefficients and an error term.

The model that these emission factors are used in are similar to the previous linear model with vehicle classes included in it, although this time there are different estimates, standard-errors and p-values which are summarised in the table below (table 3.6.2).

Variable	Estimate	Standard Error	p-value
Intercept	7.498e01	3.331e+01	0.02901
Northing	-9.401e-05	3.127e-05	0.00419
Easting	1.143e-05	4.173e-05	0.78539
Buses and Coaches	1.180e-01	3.550e-04	<2e-16
LGVs	4.240e-01	2.190e-04	<2e-16
All HGVs	1.150e-01	3.284e-04	<2e-16
All motor vehicles	1.994e-05	3.305e-05	0.54918

Table 3.6.2: Estimates, standard errors and p-values of linear model using emission factors

Table 3.6.2 shows that provided all other covariates are kept constant, a unit increases in buses and coaches emissions will increase the log NO_2 value by $0.118\mu gm^{-3}$. If all other covariates

are kept constant, and LGVs increase by a unit then log NO_2 increases by $0.424\mu gm^{-3}$ and similarly a unit increase in HGVs will increase log NO_2 by $0.115\mu gm^{-3}$.

3.7: Using Linear models and GAMs for the NO_2 emissions data, without the spatial information

Similar to the section 2.10, generalised additive models are used in this section for the spatial NO_2 data. This is due to the reasonable assumption that some of the relationships between log NO_2 and the explanatory variables of the models in section 3.3 may not be linear. This is also reasonable due to the low adjusted R^2 values of previous (linear) models. The model diagnostics can be seen below for a model containing only emission factors for HGVs and an average for all motor vehicles.

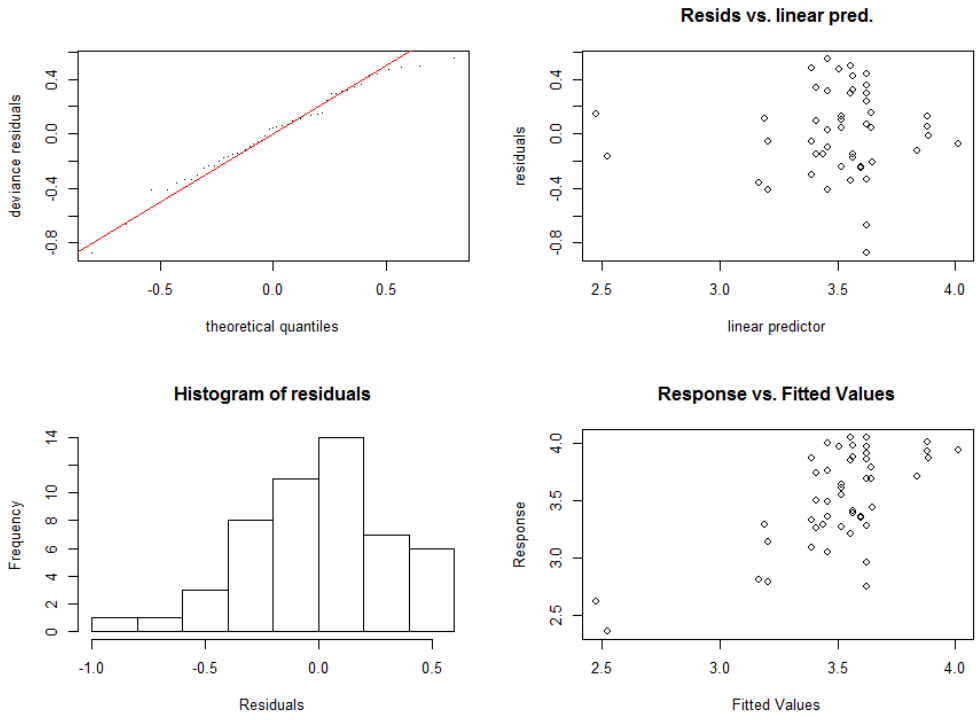


Figure 3.7.1: Diagnostic plots of GA model containing HGV emissions LGV emissions and Buses and Coaches emissions as explanatory variables

The residual plots display heteroscedasticity. This does mean that the model is inherently poor, although it could still be improved. This improvement is usually through a transformation of the response, which has already been done, or by including other variables. The inclusion of other variables is reasonable as we know already that there are missing meteorological variables due

to the lack of data available. It should be noted that the sample size is quite small. The Generalised additive model for the NO_2 and traffic data is as follows;

$$\log NO_2 = \mu + \beta_1(Buses\ emissions) + \beta_2(LGV\ emissions) + s(HGV\ emissions) + \varepsilon$$

Equation 3.7.1

Where μ is an intercept term and the ε term is an identical and independent error term, assumed to be from the normal distribution.

Looking at the relationships of these covariates with the response variable, taking, say the LGV emissions and plotting against the log NO_2 values, the following plot (Figure 3.7.2) is obtained;

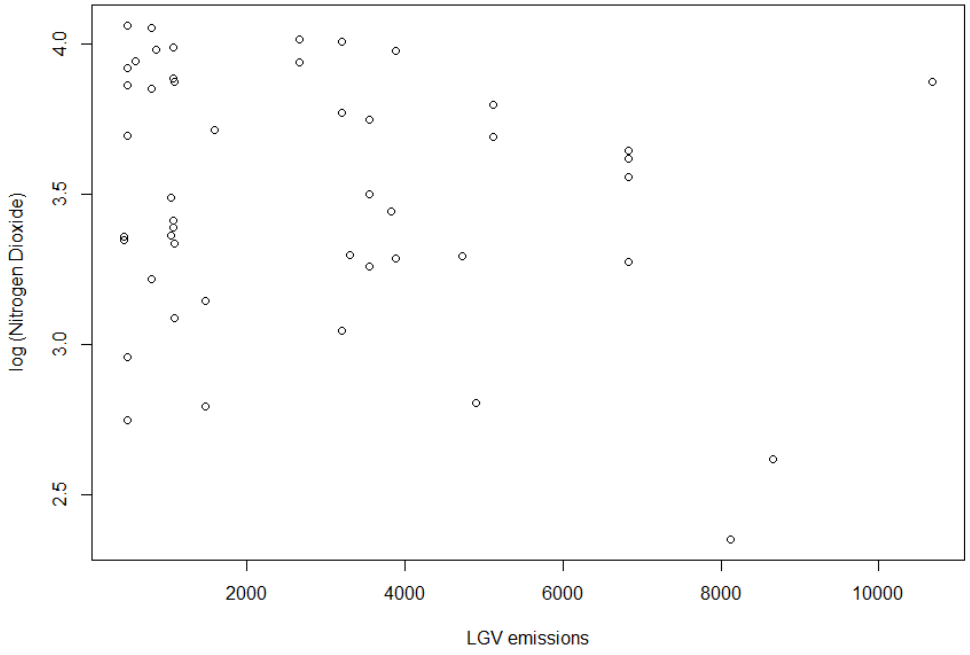


Figure 3.7.2: Log NO_2 vs LGV emissions for 2014 at locations in Aberdeen

A generalised additive model is created, with linear terms included in it. These linear terms are the buses emissions and the LGV emissions. The other term of HGV emissions is taken as a smooth term in the model, and the plot of this covariate is seen in Figure 3.7.3.

From this plot (Figure 3.7.3), one can tell that as HGV emissions increase, so too does the log NO_2 concentration. From 0 to 500 there is a cyclical behaviour of the NO_2 concentration and after 500 the NO_2 concentration steadily increases. The cyclical behaviour could be due to the

relatively low number of observations recorded. With the relationship between NO_2 and HGV emission. a GAM model is fit with Northing and Easting terms and found to be statistically not significant.

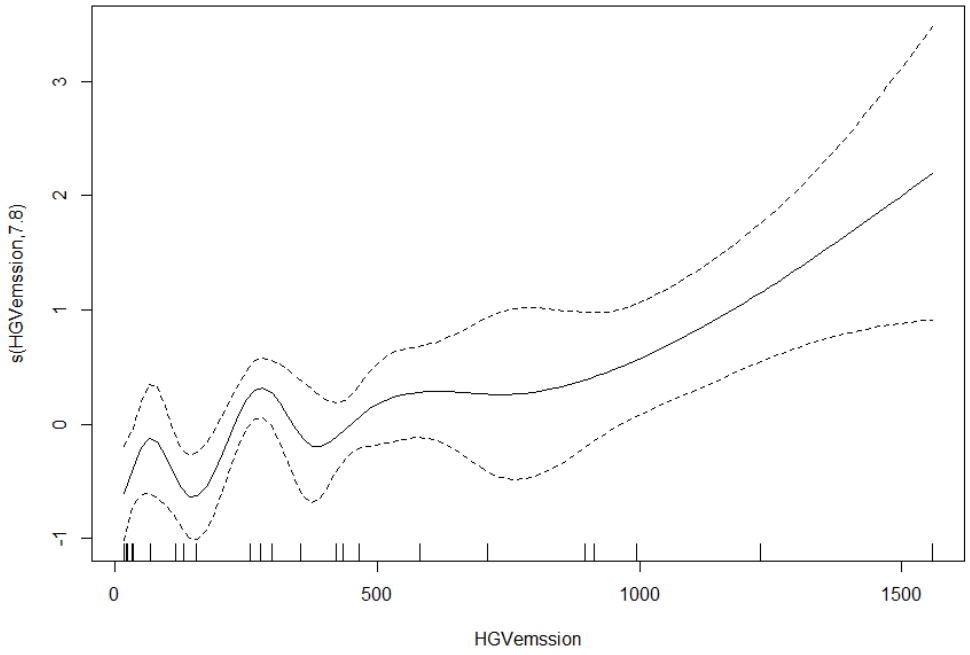


Figure 3.7.3: Plot of HGV emissions

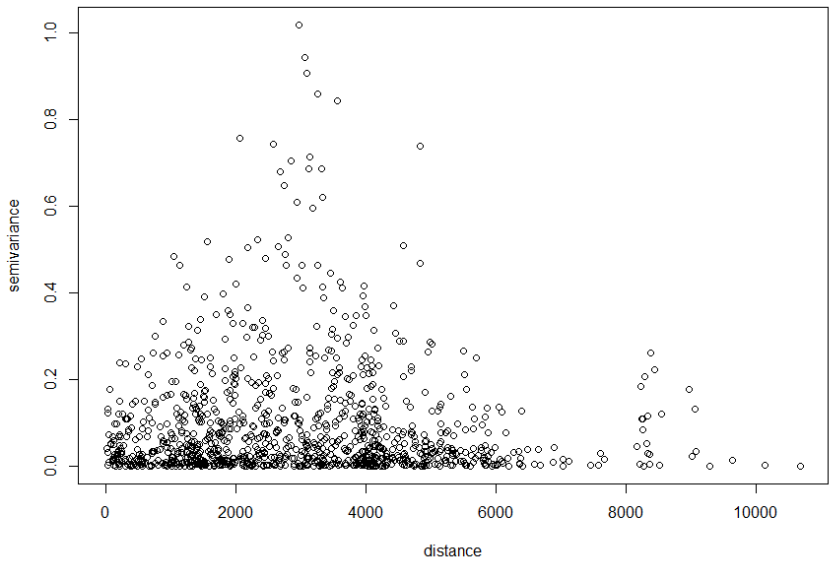


Figure 3.7.4: Variogram cloud of locations of monitoring stations

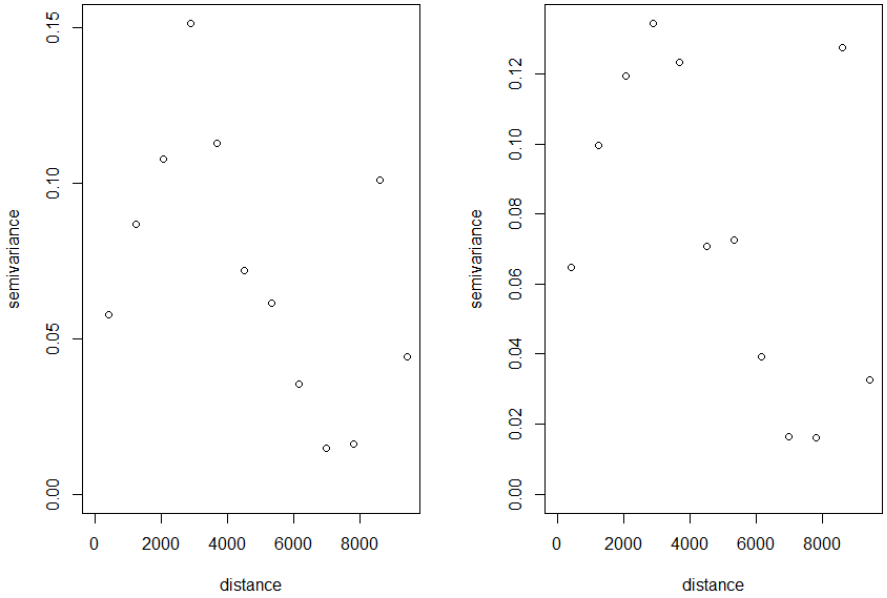


Figure 3.7.5: Robust estimator and binned estimator of the data

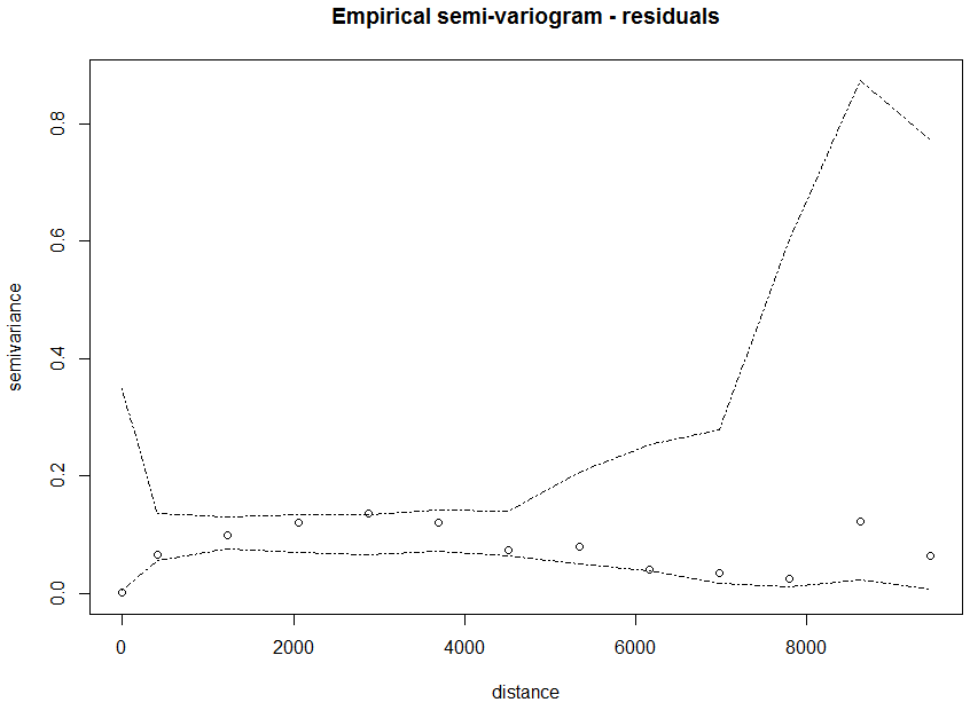


Figure 3.7.6: Semivariogram using binning and a robust estimator.

Using this plot to estimate parameters, the following results are obtained;

Model	Nugget	Partial Sill	Range
MLE	0.046	0.0459	0.4606
REML	0.047	0.0469	0.4606

Table 3.7.1: Parameter estimates obtained from MLE and REML methods

Although the generalised additive model is seen to be a better modelling approach than that of the linear model, it is still worthwhile taking a look at the linear modelling approach as was done in section 3.5. the genealised additive model is a better modelling approach as it has a higher R^2 adjusted value and a lower AIC than the linear model for the same data. The R^2 adjusted is 0.34 (.07 higher than the linear model) and the AIC is 27.2 (also lower than the linear model).

3.8 Conclusion and further work

In conclusion to this Chapter, spatial patterns have been modelled in the data recorded at both the AURN sites and the diffusion tube for the year 2014. The effects of the traffic covariates have been modelled also. It has been shown that a linear regression model works for the data i.e. that there are statistically significant variables when log NO_2 is modelled linearly with different vehicle classes as covariates. It has also been shown that the model parameters for the variogram can be estimated by restricted maximum likelihood and that any concentrations of NO_2 in the future are predicted with a higher degree of accuracy where there are a number of monitoring stations in a small area, compared to a larger area with fewer monitoring stations (see Figure 3.5.1.3.1). Further work which could be done on the spatial modelling of the Aberdeen air quality and traffic data is to compare the spatial data from year to year, instead of a single year analysis, and then make a spatial temporal model, as analysing one year does have certain limitations. Also, the spatial analysis would benefit from more monitoring stations so that a higher degree of accuracy i.e. a lower standard error would be obtained.

The traffic effects on the data show that as HGV emissions increase, so too does the NO_2 concentrations. It can also be noted that although statistically significant in the model both buses emissions and LGV emissions have a relatively small effect on NO_2 concentrations (-0.0008155 and -0.0002343 respectively) when all other covariates are held constant and a unit increase in the variable of interest is incorporated.

Generalised additive modelling is better than linear modelling for the spatial data as the model obtained from generalised additive has a higher adjusted R^2 value as well as a lower AIC value than the model made from linear modelling for the same data. This is because relationships between the response and covariates, and between covariates themselves, are more complex than linear.

The linear regression modelling approach had Easting and Northing used in it, and was shown to be a good fit to the data, with spatial autocorrelation modelled exponentially in a kriging framework. Once this was done I introduced additional covariates and the model improved. It was found the R^2 adjusted value increased as more variance in the model was explained. The small sample size may be the reason for the weak indication of spatial covariance.

For reference, the model used in the next Chapter, the linear model that is, is the same as the one looked at in section 3.5.2 although for easier inverse regression, less covariates are used. The covariates used in the linear model discussed in the next Chapter are only HGV emissions

and LGV emissions, the emissions of the vehicle classes which are discussed in detail in section 3.6.

Using the values and the model described in section 3.6, with the exception of some of the variables, it is of interest to see what concentrations the emissions must be at if a certain level of air quality - a “good” level of air quality - is to be attained i.e. what range of values can the traffic emission take to give a certain confidence of meeting an air quality target? This is discussed in the next Chapter, inverse regression.

In this next Chapter, which covers inverse regression, two models are looked at; a linear and a generalised additive model. These models contain the covariates in table 3.5.2, excluding all motor vehicles and the Easting and Northing covariates. These models are discussed in more detail in the next section, with plots showing the relationships between the response and the covariates.

Chapter 4: Inverse Regression

4.1 Introduction

In this Chapter I will examine how the models I have developed in Chapters 2 and 3 might be used to help in decision making concerning management of vehicles to ensure (with a given confidence) a certain NO_2 concentration. This Chapter will look at certain methods that are available to calculate the explanatory variables given we know the NO_2 concentration, or that we want the NO_2 concentration to be. Some of these include graphical procedures [53]. This type of statistical method is known as statistical calibration or inverse regression. For example, if we know the NO_2 concentration at a given place and a given time, we should be able to say how many Motor Vehicles are at this place in time, or how many Light Goods vehicles and so on. In other words, if it is possible to regress y on x, then it is possible to regress x on y [54]. Although there are a number of variables available for our data, there are examples in the literature with fewer covariates and which are advisable method(s) to use [53].

In context to this Chapter, NO_2 is used to determine LGV emission and later the NO_2 concentration is used to determine HGV emission, given an LGV emission, or an LGV emission given an HGV emission. The first is done by using a 2D plot and a line, the latter by using a 3D plot and a plane. These are seen in Figures 4.3.1 and 4.3.2 respectively.

For managing air pollution, a local authority has limited tools – primarily they can manage traffic, for example, by banning certain types of vehicles or restricting the days that certain areas can be entered. This strategy of restricting the days that certain areas can be entered was implemented after the Beijing Olympics [69].

4.2 Inverse Regression and Calibration

4.2.1 Calibration

Calibration in statistics can mean a reverse process to regression, instead of having a future response variable being predicted from explanatory variables which are already known, a known observation of the response variables is used to predict a corresponding explanatory variable [55].

The calibration problem in regression is the use of known data on the observed relationship between a response variable and an explanatory variable to make estimates of other values of

the explanatory variable from new observations of the response variable. This is also known as inverse regression.

An example of inverse regression is that of dating objects, using observable evidence such as tree rings in dendrochronology. The observation estimated the age of the object, rather than the converse, and the aim is to use the method for estimating calendar ages based on new measured ages [53]. A thorough overview of the calibration problem, which also includes Bayesian approaches and multivariate calibration, can be found in a review by Osborne. [55]

The objective in this thesis is to identify the conditions in the explanatory variables that would guarantee conditions in the response, for example the annual average does not exceed $40\mu gm^{-3}$, which is equal to a log NO_2 value of $3.69\mu gm^{-3}$. This value of $40\mu gm^{-3}$ is used as it is the NO_2 concentration limit of the EU, the UK and Scotland, as seen in table 1.1.1.

4.2.2 Inverse Linear Regression

Inverse regression looks at an already established relationship between a response Y and a covariate x using a set of training data $(x_1, Y_1), \dots, (x_n, Y_n)$. This relationship is then used to calculate the covariate value x_0 corresponding to an observed response Y_0. A more extensive summary is provided in the theory given by Brown [56]. This theory can be very shortly summarised as approximating graphical methods to estimate some unknown covariate given a relationship between a known set of responses and covariates has been established.

This Chapter focusses on the inverse linear regression of a linear model with initially one explanatory variable, followed by a linear model with two explanatory variables. This theory could be continued on to p – explanatory variables. When there is only one explanatory variable one can find the unknown covariate by using a graph – from reading the desired value of the response across to a diagonal line (which describes the linear relationship between the response and covariate) and down to the covariate value on the x axis. This can be seen in 4.3 and Figure 4.3.1.

As more covariates are introduced, this inverse linear regression by approximating graphical methods becomes more difficult as it is more difficult to plot the relationships of more explanatory variables with a response. This can be seen as one needs a 2-dimensional plot if one is modelling a response against some covariate, and needs a 3-dimensional plot if one is modelling a response against two covariates, and so one would need a $p + 1$ dimensional plot

if one was modelling a response against some p covariates. Section 4.3 shows up to a 3-dimensional plot.

4.2.3 Nonlinear calibration

Since many of the relationships observed between logarithmic transformed NO_2 and the explanatory variables are nonlinear, an inversion interval described in [56] provides an approximate $100(1-\alpha)\%$ confidence interval for x_0. This is also seen in Schwenke et al. [58].

Consider the following regression model;

$$Y_i = f(x_i;\ \theta) + \ \epsilon_i \qquad \textit{(Equation 4.2.3.1)}$$

where $i = 1, \dots, n$ and f may or may not be linear in θ. The goal is to solve x_0 i.e. the estimate, given an observation y_0, then the point estimate $\hat{x}_0$ can be found by solving $y_0 = f(x;\ \hat{\theta})$ for x. This solution should be unique in theory, as long as we have an f which is monotonic in the region of interest [59]. Since we are trying to find a value of an explanatory variable at a given point in time and space, we can use the above equation (equation 4.2.3.1) to compute $\hat{x}_0$ while using the relevant software to solve

$$y_0 - f\left(x; \widehat{\boldsymbol{\theta}}\right) = 0 \qquad \textit{(Equation 4.2.3.2)}$$

numerically for x [60]. This can be used for a linear result or a non-linear one.

A summary of this method of using the above equations is that these are methods for analysing data. The models themselves can be multinomial non – linear models and the software implemented is "nls2" (implemented in R Studio). This package also includes the tools which allow for confidence regions to be calculated for function of parameters or calibration intervals.

4.2.4 Inversion Interval

An exact $100(1-\alpha)\%$ confidence interval for x_0 can be given by;

$$\hat{x}_0 + \frac{(\hat{x}_0 - \bar{x})g \pm \left(\frac{t\hat{\sigma}}{\hat{\beta}_1}\right)\sqrt{\frac{(\hat{x}_0 - \bar{x})^2}{S_{xx}} + (1-g)\left(\frac{1}{m} + \frac{1}{n}\right)}}{1-g}$$

(Equation 4.2.4.1)

where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$, $g = \frac{t^2\hat{\sigma}^2}{\beta_1^2 S_{xx}} and\ t = t_{\frac{\alpha}{2},n+m-3}$ is the upper $1-\alpha$ percentile of a Student's t-distribution with $n+m-3$ degrees of freedom. This inversion interval is also seen in [57]. Having this confidence interval allows for us to say with a $100(1-\alpha)\%$ probability of being correct that the value of an explanatory variable lies within some range of values [58].

The literature uses graphical methods for illustrating and evaluating the process of inverse regression [58] [61]. The graphical methods cover inverse regression for the case of a bivariate response, which is the case with the NO_2 data and the traffic data collected at the count points, as there are multiple vehicle classes.

4.3 Data output and plots

In this section, two models are considered. These are two models which are discussed in the previous Chapter, one a linear model, the other a general additive model. These models have one exact explanatory variable and one approximate explanatory variable respectively. The models are those which use the emission factors as opposed to the actual vehicle counts. These are explored in section 3.5 and 3.6. There are two different plots which can be used to interpret the relationship between log NO_2 and LGV and HGV emissions. These are the following plots (4.2.1 and 4.2.2) and are similar to those seen in section 3.6. Within this section there are two subsections, one concerning a single explanatory variable and the other concerning two explanatory variables.

4.3.1 The single explanatory variable

When using inverse regression I chose to model log NO_2 against LGV emission. I chose this covariate as it was the most statistically significant covariate in the spatial model created in section 3.4. This simple case of inverse regression with one explanatory variable can be seen in Figure 4.3.1 and it will be developed into a case of two explanatory variables which can be seen in Figure 4.3.2.

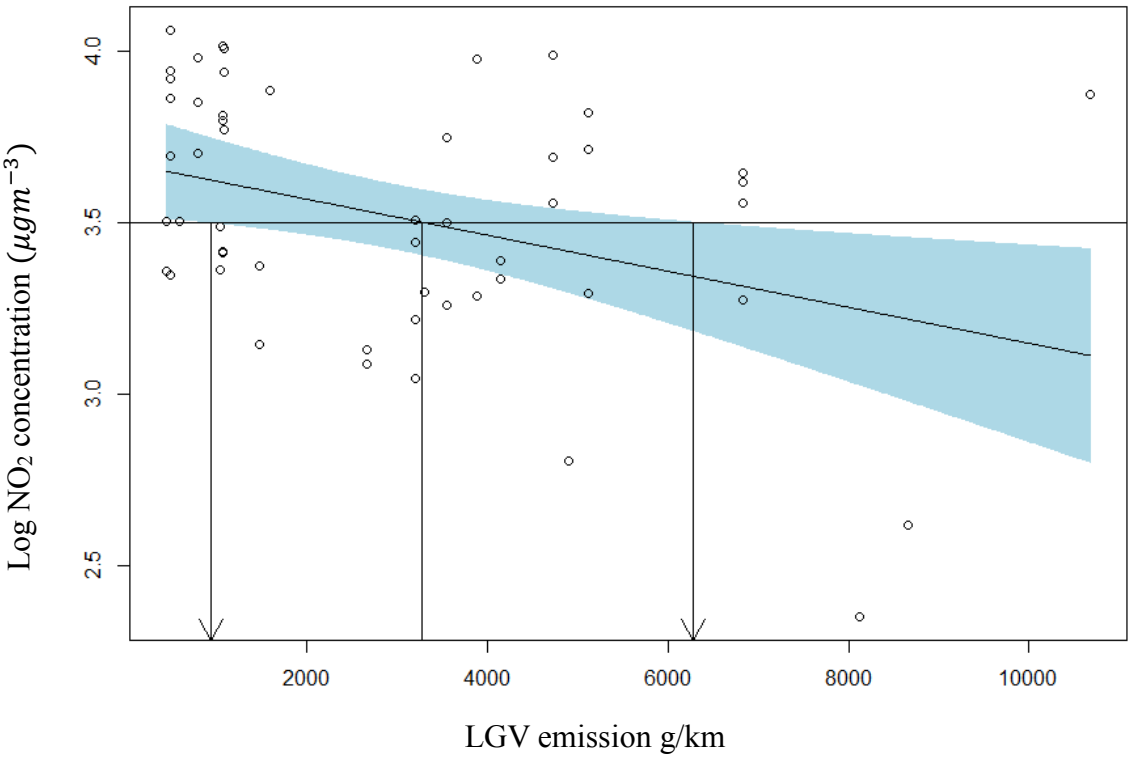


Figure 4.3.1 Linear plot showing the confidence interval for LGV emissions

This plot shows log NO_2 against LGV emissions, and suggests a general trend that as LGV emissions increase, log NO_2 decreases. It should be kept in mind that other variables might be needed. Also this is intuitively incorrect according to much literature and the general consensus of the scientific community – as the converse is accepted to be true – as emissions increase, so does log NO_2. The plot also shows confidence intervals which provide a guide to room for error in the relationship between the two variables. The horizontal line at $3.5\mu gm^{-3}$ represents the mean log NO_2 value. The horizontal line intersects the upper and lower bounds for the confidence interval and the mean.

From looking at Figure 4.3.1 it is likely that when LGV emissions are less than or equal to 1500 g/km or more than or equal to 6200 g/km that the log NO_2 concentration will exceed the limit. Less than 6200 g/km and there is a chance that the limit is exceeded.

4.3.2 Two explanatory variables

Creating a linear model which consists of log NO_2 as the response and LGV and HGV emissions as the explanatory variables, leads to the creation of another plot – this one 3D. These two covariates are chosen as they are the most statistically significant covariates of the model described in section 3.6 and 3.7, hence they have the most statistically significant effect on the NO_2 concentration. The 3D plot (Figure 4.3.2) shows how log NO_2 at each of the 55 sites is related to these emission factors for these particular vehicle classes. Figure 4.3.2 can be seen below;

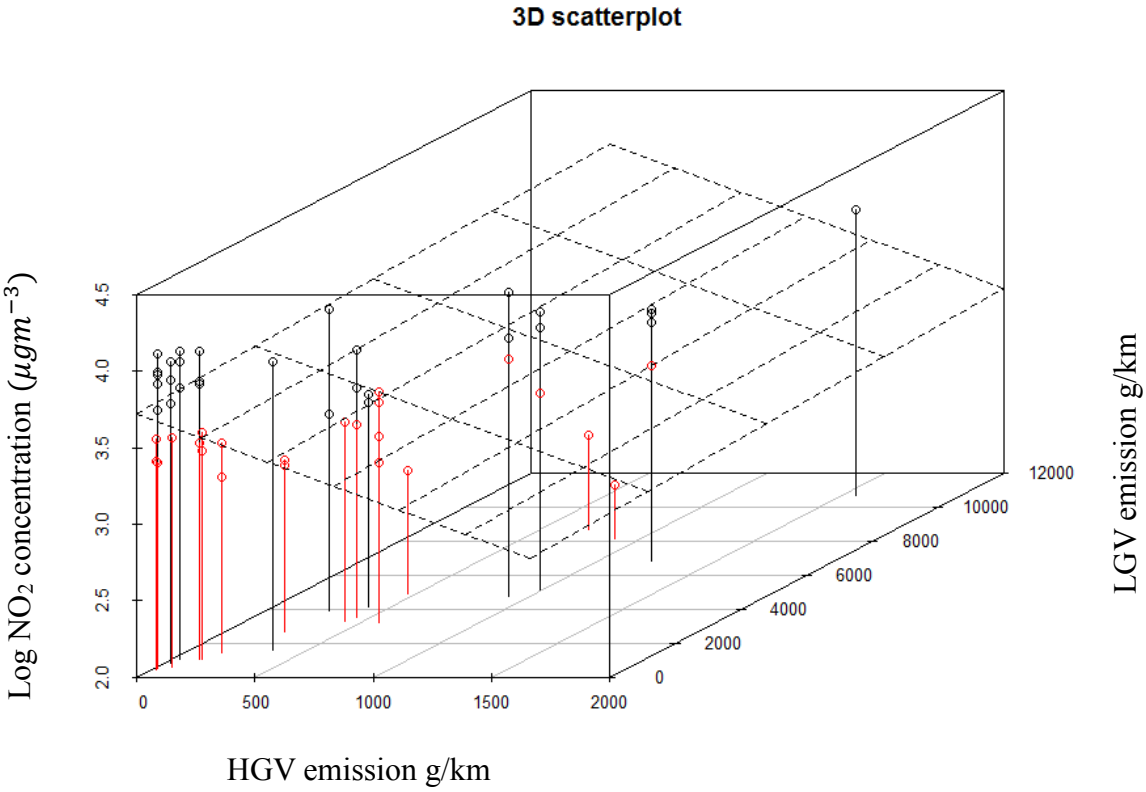


Figure 4.3.2: 3D plot showing the plane of linear model and the response values

4.3.2 shows a plane which represents the following linear model;

$$y = \beta_{0+}\beta_1 x_{1+}\beta_2 x_2 + \varepsilon \qquad \textit{(Equation 4.3.1)}$$

In this model y is log NO_2 value, β_0 is the intercept, β_1, β_2 are coefficient terms, x_1, x_2 are LGV and HGV emissions respectively, and ε is an error term.

The points on the plot are the y values of the model, with the black points being the values which lie above the plane, and the red points being the points which lie below the plane. The plot shows that some of the points are below the plane and some above, which means some recorded observations are higher than the models estimate and some observations are lower than the models estimate, as we would expect, given particular emission factor values.

In the bivariate case it is of interest to evaluate for each possible pair (x_1, x_2) if the NO_2 limit is exceeded or not. Figure 4.2.3 shows that as LGV emissions are between 2000 and 6500 g/km, and HGV emissions are between 290 and 1550 g/km then the limit will be exceeded.

It is now useful to try and explain how, empirically, it is possible to derive the conditions on two explanatory variables. The following plots (Figures 4.3.3 – 4.3.5) can be described as indicator plots. They show that some values of log NO_2 are above the threshold of $3.69 \mu g m^{-3}$ and some are below this threshold, which is $\log(40) = 3.69\ \mu g m^{-3}$. The latter plots (Figures 4.3.4, 4.3.5) depict plus and minus the confidence band of the linear model which is created and described above. In other words, these are three plots showing for which values of the LGV emissions and HGV emissions, the log NO_2 values which are above (in black) the log NO_2 value of $3.69 \mu g m^{-3}$ and which are below (in white) the log NO_2 value of $3.69 \mu g m^{-3}$. The first plot is completed using the modelled log NO_2 values, while the second and third plots are the log NO_2 values with $+2\sigma$ and -2σ, respectively.

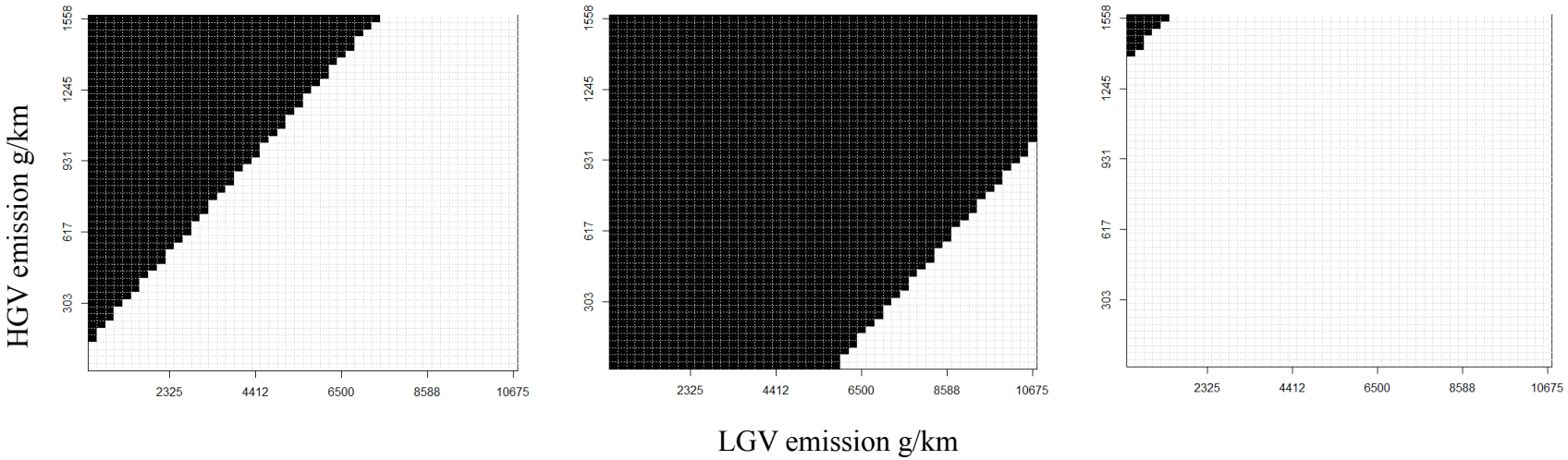


Figure 4.3.3 – 4.3.5: (From L to R) Observed log NO_2 values; log NO_2 values $+2\sigma$; log NO_2 values -2σ. White values are below a certain value and black values are above a certain value. The value is $3.69 \mu g m^{-3}$.

A plot showing values which are above and below the threshold of $3.69\mu gm^{-3}$, minus 2 standard deviations, shows that most values for different combinations of HGV and LGV emissions are below the value of $3.69\mu gm^{-3}$. These plots (Figures 4.3.3 – 4.3.5) indicate that the log NO_2 value will be below $3.69\mu gm^{-3}$ at all times, (provided other variables are kept constant), if LGV emissions are above $3577\mu gm^{-3}$ and HGV emissions are kept below $1495\mu gm^{-3}$. This is simply counter intuitive.

The approach used for the general additive model is an approximate one, modelled on the linear case. Namely, we consider the confidence band for the smooth functions, and then project where our NO_2 limit crosses the upper and lower band. These plots below (Figures 4.3.6 – 4.3.8) shows the more flexible nature of the NO_2 surface that has been fit. An observation could be that as the values of emissions increases then the log NO_2 value is more likely to be above than below $3.69\mu gm^{-3}$

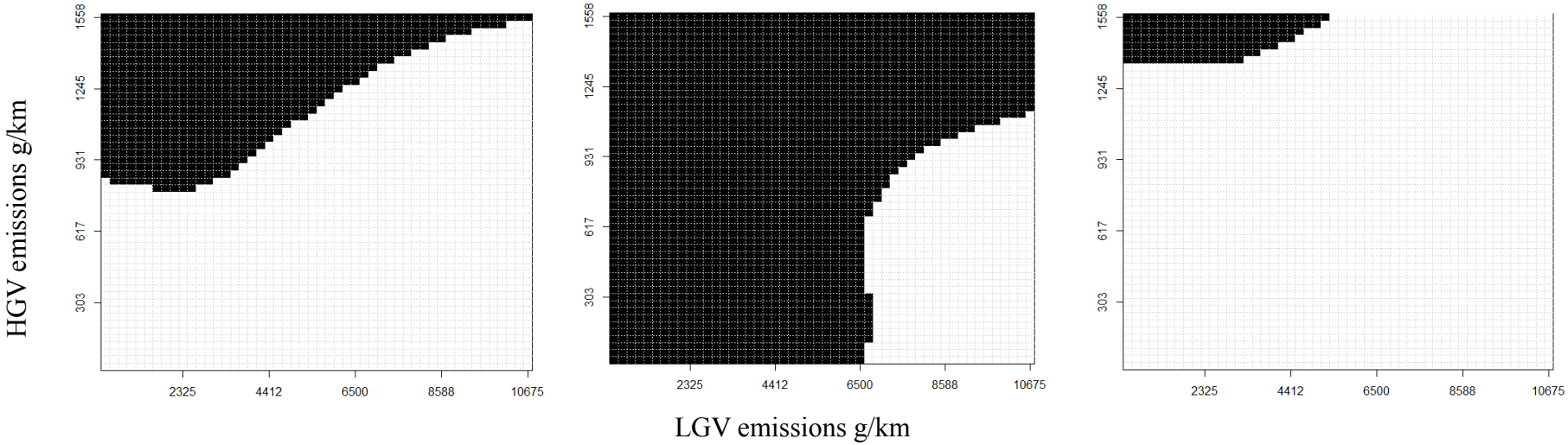


Figure 4.3.6 – 4.3.8: (From L to R). Observed log NO_2 values; log NO_2 values $+2\sigma$; log NO_2 values -2σ. White values are below a certain value and black values are above a certain value. The value is $3.69 \mu g m^{-3}$. This is for a generalised additive model.

Interpreting these 3 previous plots for the GA model; Figure 4.3.5 suggest that when HGV emissions are low, LGV emissions can take any value up to $10675\ gkm^{-1}$ and $\log(NO_2)$ will stay below $3.69\mu gm^{-3}$. This is the case, according to the Figure, for HGV emission values up to approximately $850gkm^{-1}$. HGV emissions higher than this, with a relatively low LGV emission level (about $3300gkm^{-1}$) will result in $\log(NO_2)$ values being above the value of $3.69\mu gm^{-3}$. Notice how that when LGV emissions increase that there needs to be a higher level of HGV emissions to maintain a $\log(NO_2)$ value. This is counter-intuitive. This behaviour may arise due to any number of reasons; perhaps the data wasn't recorded appropriately, or maybe the data could be modelled better, or it may be the case that there are other variables which are not being accounted for. The other plots of plus or minus 2 standard deviations show a similar pattern only that they have more and less (respectively) combinations of HGV emissions and LGV emissions which indicate a log NO_2 value above $3.69\mu gm^{-3}$.

The GAM model has a smooth surface in terms of LGV and HGV emissions meaning the regions are no longer bounded by straight lines but rather by curves. There may also be the case that LGV and HGV emissions are correlated since one would expect that where there are more LGVs there may also be more HGVs and vice versa. Although I have mentioned that the results are counter intuitive in nature regarding the relationship between NO_2 and LGV emissions, a relationship is still discovered. It has also been mentioned that it is reasonable to assume that there are missing confounding variables, including meteorology which are important. That said, the method is valuable (even if the actual results are not) as it allows us to explore under what conditions in the covariates we would or would not exceed the limit value.

4.4 Conclusions for Inverse Regression

Inverse regression is a useful tool for finding out the conditions which need to be met if a particular log NO_2 value is to not be exceeded. It was shown that the more complicated a model i.e. the more covariates in the model, the more difficult it is to plot that model and get accurate results. From the four models made, two with only one explanatory variables and two with two explanatory variables, it was clear that there are some missing confounding factors as the relationship between the response and the covariate(s) was not as expected. These models are similar to those discussed in Chapters 2 and 3 and can be seen in section 3.6.

A simple linear model with only LGV emissions showed, when put under inverse regression, that when LGV emissions are less than 6200 units, there was a chance that the log NO_2 limit exceeds the limit in place. Adding other covariates (namely HGV emissions) makes the model more complicated although yields more informative results. The inverse regression of this model yields the result that the limit will be exceeded when LGV emissions are between 2000 and 6500 g/km, and HGV emissions are between 290 and 1550 g/km.

The reason why the conclusions are limited and the results unexpected could be due to the small number of observations that are being modelled. More observations would allow for more accurate results. Another reason could be that the data that is recorded is not reliable enough i.e. the data could be cross referenced with other data recorded to ensure that the information which is recorded is accurate.

Chapter 5: Conclusions and further work

5.1 Conclusions

In this thesis, I have examined the utility of readily available air quality, traffic and meteorological data to inform the public and policy related understanding of drivers of air quality in Aberdeen. I have developed a series of linear and smooth regression models to model NO_2 temporally over 5 years at the AURN sites, and spatially (for the annual mean) using an additional 51 diffusion tube sites. I have incorporated traffic data in the form of traffic counts (derived from the department of transport data) and also using emission factors.

These models have shown some counter-intuitive results arising from artefacts – particularly in the traffic data. The meteorological data while only being available from the single site at the airport, will introduce confounding effects.

5.1.1 Air Pollution and Health affects

Air pollution is known to have negative effects on the environment, welfare and human health. More recently it has been shown to also have a negative effect on well – being [70], so one could say air pollution has negative effects on both the physiological and psychological states. The World Health Organisation estimates that every year 7million people worldwide die prematurely because of air pollution [3], in London the annual death toll is at 9500 deaths [71], and in Scotland it is 2500.

5.1.2 Work being done in UK cities to improve air quality

In cities throughout the UK, through DEFRA and the devolved administrations of Scotland, Northern Ireland and Wales, work is being done to mitigate the negative effects of air pollution. Some of this work consists of annual reports which summarise the measurements from national monitoring networks, more funding being used for the monitoring of air quality, and by introducing "Clean Air Zones". These Clean Air Zones ensure that more modern vehicles which have relatively clean engines are the only vehicles allowed into certain parts of the cities in which the zones are implemented, whereas vehicles which are suspected of contributing to a greater deal of air pollution are not allowed in without incurring a charge.

Other measures in place to reduce air pollution include using cleaner fuels or cutting edge technology in vehicles such as taxis and buses. These types of vehicle would be the kind allowed

entry into London's ULEZs or Ultra-Low Emission Zones, which are to be implemented by 2020. In Scotland a similar plan is in place to reduce the amount of traffic and hence, air pollution, by introducing low emission zones. The first low emission zone is to be introduced in Scotland by 2020.

There are many local campaign groups in Scotland who campaign for cleaner air. These are similar to those seen working with Friends of the Earth. There are EPAs (or Environmental Protection Agencys) set up in England, Scotland, Northern Ireland and Wales who monitor and collect data on air quality such as NO_2.

5.1.3 The data

Some of these data are seen in this thesis in the form of NO_2, traffic counts, vehicle emissions, and meteorological observations. The NO_2 data are available within the desired time frame from 2006 – 2015, and during this period there is relatively little missing data and no notable outliers. The data followed a seasonal pattern with a higher concentration of NO_2 in the winter compared to the summer, as well as a higher concentration of NO_2 during weekdays compared to the weekend. At all sites there was seen to be a higher concentration of NO_2 between 6am and 6pm compared to the concentration of NO_2 between 6pm and 6am. These data are collected by AURN sites and diffusion tubes located throughout Aberdeen, and although these sites and tubes are not at uniform locations throughout the city they do give a good spatial representation of the NO_2 concentration during 2014. NO_2 is seen to be gradually increasing over time, and also it is clear that there is a higher concentration of NO_2 where there is more traffic. This is reflective of a higher NO_2 concentration in the city of Aberdeen compared to on the outskirts of the city.

The data for traffic are relatively more difficult to work with, as it needs to be disaggregated before it can be used in a temporal or spatial context. These data show issues which it was not possible to fully resolve but reflect data made available on official sites. There are 4 vehicle classes found to be of interest and these are HGVs, LGVs, all motor vehicles and buses and coaches. The data in all of the vehicle classes makes a sudden change in 2012 and this can be put down to a different counting method.

The data for meteorological variables are all collected at Dyce airport thus limiting their usage to only temporal modelling i.e. they are not useful in spatial modelling. The covariates included in the meteorological variables are wind speed and direction, cloud cover, rainfall, temperature

relative humidity and pressure. Over time wind speed had a strong seasonal pattern, with higher wind speeds in winter compared to Summer, whereas wind direction did not seem to follow a seasonal pattern. Cloud cover was mostly quite high with a lot of variation while rainfall was consistently between 0.0 mm hour^{-1} and 1.0 mm hour^{-1} with the exception of a few observations. Temperature had the strongest seasonal affect which is present in Figure 2.4.2 with higher temperatures in the summer months while there are lower temperatures in the winter months, as to be expected. Pressure follows a similar seasonal pattern although there is more variation in the observations compared to temperature. Relative humidity does not seem to have any strong seasonal affect although there may be one present. Plots of the meteorological data against the NO_2 concentrations at Anderson Drive showed how NO_2 changed corresponding to different meteorological variables. As wind speed increased it was clear that NO_2 decreased, while there was a quadratic relationship between wind direction and NO_2 concentration. Cloud cover and rainfall had no obvious relationship with NO_2 concentration while temperature had a negative relationship with NO_2 at Anderson drive. Relative humidity and pressure both had a positive relationship with NO_2 so as they both increased so too did NO_2 concentration, for Anderson Drive.

5.1.4 Modelling the data

Fitting a linear model, it is found that not all of the variables are statistically significant for all sites. This is seen as some variables are removed for the linear models such as day within week and day within year for the linear model at Anderson Drive, or Cloud cover and pressure are removed from the model built for Union Street while other variables are left in such as wind speed, humidity and buses and coaches at Union Street. It is found from these linear models that a unit increase in the total number of motor vehicles travelling on Union St results in a $0.0091 \mu g m^{-3}$ increase in log NO_2 concentration provided every other variable is held constant.

Fitting generalised additive models showed non-linear relationships between NO_2 and some covariates. For example, at Union St NO_2 increased from 2006 to 2011 and thereafter decreased up until 2015. The day within year showed that the NO_2 concentration was lower in Summer months compared to Winter months at Union St. A higher wind speed resulted in a lower log NO_2 concentration – this was shown to be the case at all sites. It was shown at Wellington Road that more motor vehicles resulted in a higher concentration of NO_2. There are some surprises in the generalised modelling of an increase in the number of vehicles resulting in a lower NO_2

concentration, although this could be due to confounding factors. It could also be because of the monitoring of both metrics being in different locations. There is also shown to be a positive relationship between the number of buses and coaches and NO_2 concentration at Errol Place in Aberdeen.

From modelling the data spatially it was found that a linear regression model is suited to the data i.e. modelling log NO_2 linearly with different vehicle classes as covariates shows that there are statistically significant variables. Concentrations of NO_2 can be predicted with a higher degree of accuracy where there are more monitoring stations, compared to where there are less monitoring stations. Including traffic covariates in this model shows that as HGV emissions increase so too does the NO_2 concentration. A unit increase in buses emissions and LGV emissions have a relatively small effect on NO_2 concentration when all other covariates are held constant – these concentrations are -0.0008155 and -0.0002343 respectively.

Incorporating general additive modelling to the spatial data results in a relatively high R^2 adjusted value when compared to the linear model, as well as a lower AIC value. The linear model approach was still a good one as it was shown to be a good fit to the data, and adding in more covariates showed that the R^2 adjusted value only increased as more variance in the model was explained.

The air quality in Aberdeen is generally good, although there are some sites which have exceedances of the annual average limit of NO_2 which is $40\mu gm^{-3}$. This annual average is from the European Commission and there are sites in Aberdeen which have an NO_2 concentration as high as $59\mu gm^{-3}$ at some times. These sites include Union St and Anderson Drive.

5.1.5 Inverse Regression

Inverse regression is a tool which might allow a model to be used for management. It can be used under certain conditions that may allow for a reasonable probability that standards would not be exceeded. Although in this thesis using inverse regression provided results of a counter-intuitive nature, where it described the relationship between NO_2 and the LGV emissions covariate, it is still a tool which can be used to explore under different conditions covariates would need be if the NO_2 concentration was to exceed, or not exceed as may be the case, a limit.

One limit to inverse regression is that it becomes more complicated as more covariates are used in the model. This can lead to important confounding variables being left out of a model and some results being produced which are not as valuable as the method of inverse regression is itself.

5.1.6 Chapter analysis and limitations

Chapter 1 gives a short summary on air pollution. It has also introduced the data; for NO_2, meteorological variables and traffic variables. The NO_2 is introduced at the different sites for both AURN sites and diffusion tube sites. Finally, this Chapter discusses the aims of the thesis. The main overall aim was to build a model (or models) which show the behaviour of NO_2 over time, and then over space. Limitations of the data for the traffic variables - these are limited and had to be disaggregated.

The model was built and discussed in Chapter 2. Both a linear model and a generalised additive model were built for each of the 5 AURN sites. It was found that at each of these sites the generalised additive model proved to be the better model when compared to the linear model corresponding to the same site. The generalised additive model proved to be a better model in each case as more variance was captured as well as having a lower AIC value compared to the corresponding linear model. These methods are discussed in [62] and have links with extensive information. From plots in section 2.7 one can conclude that there is no overall trend but there are strong seasonality and day of the week effects. Limitations to the temporal modelling were that only 5 sites could be modelled temporally, given the nature of the diffusion tube data. Only the AURN sites could be used. This is a limitation of the work carried out in the thesis as opposed to the limitation of the methodology. The methodology itself can be carried out at other locations and during other years if this was of interest.

Chapter 3 looked at the data in the spatial dimension, focussing on the year 2014 for approximately 50 locations. Two models were made in this Chapter; one with only spatial variables such as Easting and Northing, and a second with the same variables used that were used in Chapter 2, namely traffic variables consisting of different vehicle classes and meteorological variables also. In this Chapter, it was quickly shown that weather variables were not useful as there was only one site where the data could be collected, namely Dyce Airport, thus resulting in there being no data to model any spatial variation from NO_2 monitoring site to NO_2 monitoring site. Other variables like the traffic variables were useful when it came to

building a linear model, and the conclusion is that the traffic variables LGV emissions and HGV emissions are both statistically significant, albeit that LGV emissions were proven to be unintuitively negatively correlated with NO_2 concentrations. One would expect that as LGV emissions increased, so too would NO_2 concentrations, although the work done in this thesis suggests otherwise. This is not to say that the work here is incorrect, or indeed that it is correct, it is there to be disproven. It is possible to do further work on the sites by collecting more data, establishing weather monitoring stations at different locations, and carrying out other statistical methods for modelling on the same data. The results from Chapter 3 confirmed that areas which were closer in space were more like one another, for example that the areas which had the highest NO_2 concentration was in the city centre, whereas rural areas had a lower concentration of NO_2. The results presented corresponded only to the year 2014 as this was used as an illustrative example of the analysis which could be undertaken.

Chapter 4 focused on the inverse regression methods used to show, given a particular concentration of NO_2 concentration, what the number of a particular vehicle class had to be equal to, for a specific place during a specific year. For example during 2014 at Anderson Drive, given we want the log NO_2 value to be below $3.69\mu gm^{-3}$, the level of HGV emissions must be below $200gkm^{-1}$ approximately, with other variables kept constant. This is for the model containing only an intercept, an HGV term, and an LGV term as discussed in sections 3.5, 3.6 and 4.3. The basic idea that NO_2 can be predicted given other covariates are known is presented in this Chapter. It is limited as the results are specific to Anderson Drive i.e. if inverse regression were carried out at more sites the results would be more interesting. Inverse regression in this Chapter was also limited o analysing only two covariates.

5.2 Further work

Moving forward with this work in mind, further work could be done to include a similar analysis to the likes that has been done, only this time including other pollutants (such as PM_{10} and $PM_{2.5}$) also. This is in the sense of spatially and temporally analysing pollutants in Aberdeen. This, in theory, could be carried out in other cities in Scotland (as seen in the work by Allison [63]), and indeed carried out in other cities internationally.

An ideal set up would be to have a number of count points, weather monitors, and pollutant monitors (for monitoring traffic, meteorological and pollutant variables respectfully) located throughout the city, at equal distances from one another so that a grid of monitors may be

established, say each square of the grid being $1km^2$. This would allow, in theory, for high resolution monitoring of NO_2 and other pollutants, and as a direct result high resolution modelling for the relationships between pollutants and other parameters. The same modelling methods employed in this thesis could be used for the modelling of data in this ideal set up. This set up would also overcome the key limitation that the measurements of NO_2 and traffic were not made in the same places. As can be seen from previous figures some of the locations are quite far away from one another.

This problem of varying distances between locations measuring pollution and traffic could be responsible for, or at least a major contributor, to the non – intuitive relationships discovered between NO_2 and some vehicle emission classes. With measuring locations not being in the exact same location, and with air pollution changing measurably over the distance of a few metres, the observed data are not as accurate as one would like. The resulting data is slightly misclassified. This problem of misclassification could perhaps be overcome by simply moving the diffusion tubes to the count point locations. Another way of overcoming this problem of distance between monitoring locations is by using land use regression as is done in [77]. Other variables (as well as emission factors) are used. For example, traffic intensity. This paper also highlights the problems with traffic data as it is not available throughout cities, usually the availability is restricted to major roads only. To overcome this problem, some land use regression studies have explored successfully the use of the length of specific road types when traffic intensity data is lacking [78]. It would be wise to have all monitoring locations, whether for NO_2 or traffic on the same map, for clearer comparison.

Further work which could be done on the spatial modelling of the Aberdeen air quality and traffic data is to compare the spatial data from year to year, instead of a single year analysis, and then make a spatial temporal model, as analysing one year does have certain limitations. Also, with the few numbers of monitoring stations, the spatial analysis would benefit from more stations so that a higher degree of accuracy i.e. a lower standard error would be obtained.

A more realistic set up is to continue with inverse regression using the models described in sections 3.4 – 3.6. The difference being between the work already done, and the further work to be done is that there are a number of parameters which have yet to be used in inverse regression i.e. there are a number of combinations which need to be set so that one can find how many of a certain vehicle type need to be going down a particular road at a particular time so that log

NO_2 can be below a certain concentration. This is the same use of the tool that is described in Chapter 4, only this time for other variables.

Another logical next step which could be done is to build a spatio-temporal model so that air pollution in Aberdeen is modelled across space and time simultaneously. This would be similar to the work done by Lindström et al. [64]. This is a natural progression from building separately both a spatial model and a temporal model.

List of References

[1] BBC (2014). Scotland's 'most polluted streets' identified. http://www.bbc.co.uk/news/uk-scotland-25895007

[2] Wikipedia (December 2016). Nitrogen Dioxide. https://en.wikipedia.org/wiki/Nitrogen_dioxide

[3] BBC (Jan 2016). Scotland's 'most polluted' streets named. http://www.bbc.co.uk/news/uk-scotland-35333076

[4] World Health Organisation, Media Centre. (March 2014). 7 million premature deaths annually linked to air pollution. http://www.who.int/mediacentre/news/releases/2014/air-pollution/en/

[5] International Energy Agency, World Energy Outlook special report. (2016). Energy and Air Pollution. https://www.iea.org/publications/freepublications/publication/weo-2016-special-report-energy-and-air-pollution.html

[6] Air Quality Expert Group (2004). Nitrogen Dioxide in the United Kingdom. https://uk-air.defra.gov.uk/assets/documents/reports/aqeg/nd-summary.pdf

[7] Vitousek, P., Aber, J., Howarth, R., Likens, G., Matson, P., Schindler, D., Schlesinger, W., Tilman, D. (1997). Human Alteration of the Global Nitrogen Cycle: Sources and Consequences. *Ecological Applications, 7(3),* 737 – 750.

[8] Kampa, M., Castanas, E. (Jan 2008). Human Health effects of air pollution. *Environmental Pollution 151(2),* 362 – 367.

[9] The World Health Organisation (2005). *Air Quality Guidelines, Global Update 2005, 89 – 101,* 333 – 373.

[10] Bobbink, R., Hornung, M., Roelofs, J. G. M. (October 1998). The effects of air-borne nitrogen pollutants on species diversity in natural and semi – natural vegetation. *Journal of Ecology, 86,* 717 – 738.

[11] OECD (June 2009). The Economic Consequences of outdoor Air Pollution. https://www.oecd.org/env/the-economic-consequences-of-outdoor-air-pollution-9789264257474-en.htm

[12] Met Office (2012). http://www.metoffice.gov.uk/education/teens/case-studies/great-smog.

[13] Bell, M., Davis, D., Fletcher, T. (2004). A retrospective assessment of mortality from the London smog episode of 1952: the role of influenza and pollution. *Environmental Health Perspective 112,* 6-8.

[14] Dockery, D., III, C. P., Xu, X., Spengler, J., Ware, J., Fay, M., Jr, B. F., and Speizer, F. (1993). An association between air pollution and mortality in six U.S. cities. *The New England Journal of Medicine 329,* 1753 – 1759.

[15] Pope III, C. A., Thun, M. J., Namboodiri, M. M., Dockery, D. W., Evans, J. S., Speizer, F. E., and Jr, C. W. H. (1995). Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults. *American Journal of Respiratory and Critical Care Medicine 3,* 669 – 674.

[16] Dominici, F., Peng, R., Bell, M., Pham, L., McDermott, A., Zeger, S., and Samet, J. (2006). Fine Particulate Air Pollution and Hospital Admission for Cardiovascular and Respiratory Diseases. *The Journal of American Medical Association 295,* 1127 – 1134.

[17] European Environment Agency (2016). European Environment Agency. http://www.eea.europa.eu/.

[18] EIONET (2016). European Environment Information and Observation Network. http://www.eionet.europa.eu/.

[19] Defra (2016). Department for Environment, Food and Rural affairs. https://www.gov.uk/government/organisations/department-for-environment-food-rural-affairs/.

[20] Scottish air quality (2016). Air Quality in Scotland. www.scottishairquality.co.uk/.

[21] Ye, X., Wolff, R., Ye, W., Vaneckova, P., Pan, X., and Tong, S. (2012). Ambient temperature and morbidity: a review of epidemiological evidence. *Environmental Health Perspectives 120,* 19 – 28.

[22] European Parliament Council (2002, July). *Sixth Environmental Action Programme*.

[23] The Supreme Court (2013, May). R (on the application of ClientEarth) (Appellant) The Secretary of State for the Environment, Food and Rural Affairs (Respondent). http://www.supremecourt.gov.uk/decided-cases/docs/UKSC_2012_0179_Judgement.pdf.

[24] National air quality objectives (2016). Defra documents on UK air quality. https://uk-air.defra.gov.uk/assets/documents/National_air_quality_objectives.pdf/

[25] Local Air Quality Management (2016). LAQM of Defra on diffusion tubes. http://laqm.defra.gov.uk/diffusion-tubes/diffusion-tubes.html

[26] Whiteman, C. D., et al. (2014). Relationship between air pollution and meteorological variables in Utah's Salt Lake Valley. *Atmospheric Environment 94*, 742-753.

[27] Hargreaves, P. R., et al. (January 2000). Local and seasonal variations in atmospheric nitrogen dioxide levels at Rothamsted, UK, and relationships with meteorological conditions. *Atmospheric Environment Vol. 34, Issue 6.* 843 – 853.

[28] Department for Transport (2016). DoT Average annual daily flow temporal traffic distributions. https://www.gov.uk/government/statistical-data-sets/tra03-motor-vehicle-flow

[29] Chatfield, C. (2004). The analysis of time series an introduction. *Texts in statistical science*, 6th edition

[30] Box, G., Jenkins, G., and Reinsel, G. (2008). Forecastion and Control. *Time Series Analysis. John Wiley and Sons.*

[31] Kupper, L. (1972). Fourier Series and Spherical Harmonics Regression. *Journal of the Royal Statistics Society 42*, 121 – 130.

[32] Hastie, T. & Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science Vol. 1, Issue 3*, 297 – 318.

[33] Hastie, T. & Tibshirani, R. (1990). *Generalized Additive Models*. Monographs on Statistics and applied probability Vol. 43, Chapman and Hall.

[34] Wood, N., S. (2006). *Generalized Additive Models.* Texts in Statistical Science. Chapman and Hall, 121 -140.

[35] Cressie, N. (1993). *Statistics for Spatial Data*. Wiley Series for Probability and Mathematical Statistics, New York.

[36] Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2004). *Hierarchical modelling and analysis for spatial data*. Chapman and Hall, Boca Raton, Florida.

[37] Matèrn, B. (1960). Spatial Variation, *Technical report*, Statens Skogsforsningsinstitut,

Stockholm.

[38] Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations (with discussion), *Journal of the Royal Statistical Society, Series B* 26: 211 – 252.

[39] Diggle, P. J. and Ribeiro Jr., P. J. (2007). Model-based geostatistics. *Springer.*

[40] Cressie, N. and Hawkins, D. M. (1980). Robust estimation of the variogram, *Mathematical Geology* 12: 115 – 125.

[41] Diggle, P.J. (2003). *Statistical Analysis of Spatial Point Patterns*, Academic Press.

[42] Scott, M. (2007). Spatial point patterns and geostatistics, *an introduction* http://slideplayer.com/slide/776353/.

[43] Stein, M. L. (1999). Interpolation of spatial data: *some theory for kriging.* Springer, New York.

[44] Lawson, A. B. (2006). Statistical Methods in Spatial Epidemiology. John Wiley & Sons.

[45] Example of variogram with data which has no substantial correlation, https://www.researchgate.net/figure/284091710_fig4_Fig-4-Sample-variogram-and-envelope-of-Monte-Carlo-variogram-simulations-for-N-canopy.

[46] Interactive map of count points for traffic in Aberdeen city: http://www.dft.gov.uk/traffic-counts/cp.php?la=Aberdeen+City

[47] Interactive map for finding diffusion tube locations in Scotland using Eastings and Northings: http://www.gridreferencefinder.com/

[48] Oliver, M. A. (1990). Kriging: A Method of Interpolation for Geographical Information Systems. *International Journal of Geographic Information Systems 4*: 313–332.

[49] Diggle, P. J., Christensen, O. F. and Ribeiro Jr., P. J. (2003). Geostatistical software – geoR and geoRglm, *Proceedings of the distributed statistical computing conference*.

[50] Definitions of flow and annual average daily flow, government webpage: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/487689/annual-road-traffic-estimates-2014.pdf

[51] Matheron, G. (1965). *The regionalised variables and their estimation,* Paris.

[52] Emission factors for different vehicle classes: http://naei.defra.gov.uk/data/ef-transport

[53] Jones, G., Lyons, P. (2009). Approximate Graphical Methods for Inverse Regression, *Journal of Data Science 7: 61 -72.*

[54] Li, K. C. (1988). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association 86:* Issue 414.

[55] Osborne, C. (1991). Calibration: A review. *International Statistical Review*, 59(3): 309 – 336.

[56] Greenwell, B. M. and Schubert Kabban, C. M. (2014). Inverstr: An R Package for Inverse Estimation. *The R Journal Vol. 6/1*

[57] Brown, P. J. (1993). Measurement, Regression and Calibration. *Oxford University Press.*

[58] Schwenke, J. R. and Milliken, G. A. (1991). On the calibration problem extended to nonlinear models. *Biometrics, 47(2):* 563 – 574.

[59] Seber, G. and Wild, C. (2003). Nonlinear regression. *Wiley Series in Probability and Statistics. John Wiley & Sons.*

[60] Huet, S. (2004). Statistical Tools for nonlinear regression: A practical guide with S-PLUS and R Examples. *Springer Series in Statistics, Springer.*

[61] Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 16:* 175 – 185.

[62] 7 types of regression techniques: *"7 types of regression techniques you should know!"*:

https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/

[63] Allison, K. J. (2014). Statistical Methods for constructing an air pollution indicator for Glasgow,*University of Glasgow*

[64] Lindström, J., Szpiro, A. A., Sampson, P. D. (2004). A flexible spatio-temporal model for air pollution with spatio and spatio temporal covariates. *Environmental and Ecological Statistics, 21(3)*: 411 – 433.

[65] Ord, J. K. (1983). Kriging entry in encyclopedia of statistical sciences. *Wiley, New York, Vol. 4:* 411 – 413.

[66] Cressie, N. (1990). The Origins of Kriging. *Mathematical Geology, Vol. 22, No. 3*.

[67] Vehicle certificate agency: *"Cars and air pollution":* http://www.dft.gov.uk/vca/fcb/cars-and-air-pollution.asp

[68] Air Pollution Emissions overview: *"United States Environmental Protection Agency":* https://www3.epa.gov/airquality/emissns.html

[69] The Guardian: *"Beijing keeps Olympic restrictions on cars after air quality improves":* https://www.theguardian.com/environment/2009/apr/06/beijing-pollution-carbon-cars

[70] The Guardian: *"Air pollution as bad for health as partner's death, researchers say*": https://www.theguardian.com/environment/2017/apr/17/air-pollution-as-bad-for-wellbeing-as-partners-death-say-researchers

[71] Walton, H., Dajnak, D., Beevers, P., Williams, M., Watkiss, S., Hunt, A., (2015). Understanding the Health Impacts of Air Pollution. *King's College London*

[72] UK Grid Reference Finder: www.gridreferencefinder.com

[73] "gam" help page: www.cran.r-project.org/web/packages/gam/gam.pdf

[74] "geoR" help page: www.cran.r-project.org/web/packages/geoR/geoR.pdf

[75] Cape, J. N. (2009) The Use of Passive Diffusion Tubes for Measuring Concentrations of Nitrogen Dioxide in Air. *Critical Reviews in Analytical Chemistry, 39, 289 – 310.*

[76] Lee, A., Szpiro, A., Kim, S. Y. & Sheppard, L., (2015). Impact of preferential sampling on exposure prediction and health effect inference in the context of air pollution epidemiology. *Environmetrics 26, 255 – 267.*

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

[78] Henderson, S., Beckerman, B., Jerrett, M., & Brauer, M., (2007). Application of land use regression to estimate long-term concentrations of traffic related nitrogen oxides and fine particulate matter. *Environ. Sci. Technol., 41, 2422 – 2428.*