Statistical Methods for Constructing an Air Pollution Indicator for Glasgow

Katie Jane Allison

A Dissertation Submitted to the

University of Glasgow

for the degree of

Master of Science

School of Mathematics & Statistics

February 2014

Abstract

Air pollution can have both a short term and long term detrimental effect on health. This thesis aims to provide an air quality indicator to be used as a simple and informative tool to track air pollution levels which can be used by both the public and governing bodies.

Chapter 1 discusses the background and motivation of the study. The chapter then moves on to outlining the aims and overall structure of the thesis and provides a description of the data used.

Chapter 2 explores the daily mean monitoring site PM_{10} data for Glasgow across the years 2010 to 2012. This chapter explores trends and seasonality in the PM_{10} data using exploratory measures and time series analysis.

Chapter 3 explores the gridded modelled annual mean PM_{10} map data across the years 2010 to 2012. The spatial aspects of PM_{10} are first explored using numerical and graphical summaries. A more robust approach is used to then produce a geostatistical model to explain the trend of PM_{10} across Glasgow.

Chapter 4 then focuses on producing naive indicators building upon the modelling and exploratory analysis conducted in Chapters 2 and 3. This forms the basis of a spatio-temporal model. This results in a final air quality indicator estimate with uncertainty which accounts for spatial and temporal dependence for Glasgow.

Chapter 5 ends the thesis with a discussion of the final indicator and the conclusions with consideration given to improvements which could be made and additional analysis for the future.

Acknowledgements

I would like to take this opportunity to thank my supervisors Marian Scott and Peter Craigmile for their invaluable guidance and support throughout this project. I would like to say how grateful I am to the ISD for funding my research.

I must say a massive thank you to my Anna Price, Elizabeth Irwin, Kirsten Fairlie and Rachel Holmes for making life in the Boyd Orr extra fun and full of laughs.

Last but not least, the biggest thank you goes to my mum, dad, brother Jack, my boyfriend Charlie and all of my friends who will be happy to never hear the word masters ever again.

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 reported in it, except where otherwise stated.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation and Air Pollution Background

An indicator is a simple statistic that can summarise the level of air pollution. Air pollution, as a whole, is complex and made up of a large number of pollutants which makes it difficult to track the current state. Indicators provide an easy and accessible way to assess the current state of air pollution and provides a platform to compare air pollution levels at different time points or spatial locations. Due to their simplicity, indicators are accessible to the general public as well as policy makers and governmental bodies. An air pollution indicator could be used to set standards and affect policies. Indicators can use a selection, weighting and aggregation process - each of which has no set rules nor is there an order in which to process these steps, of which both can have an impact on the final result. The selection process involves selecting which pollutants to include in the indicator. The selection could be due to availability and quality of data. A pollutant could be selected which is seen as more important in describing the overall trend. If a number of pollutants are selected then a decision has to be made about how to weight each pollutant - equally or with more weight on a certain pollutant. There are a range of ways to aggregate pollutants with different measurement units.

This brings us onto the motivation of this study. The BBC recently released an article which discusses Scotland's most polluted streets (BBC, 2014). This shows that the subject of air pollution in Scotland's cities is a high profile subject matter. The BBC article discusses the various health risks associated with high levels of air pollution, and a table within the article details the streets with the highest level of Nitrogen Dioxide (NO_2) and Particulate Matter which measures 10 micrometers in diameter or less (PM_{10}). Glasgow's Hope Street tops the list of highest NO_2 levels while Aberdeen's Market Street topped the list of PM_{10} levels. Air pollution has a detrimental effect on human health and the environment (Defra, 2013b). The earth's atmosphere is made up of a layer of gases which surround the earth. Air pollution can take the form of natural or man-made solid particles, liquid droplets, or gases. An airborne substance that has an adverse affect on human health and the environment can be described as air pollution. Pollutants can be described as primary or secondary; primary pollutants are produced directly from a process whereas secondary pollutants are formed in the air when other primary pollutants react. A number of primary pollutants that contribute to air pollution include: carbon monoxide, nitrogen oxides, sulphur oxides, particulate matter (PM), volatile organic compounds, radioactive pollutants and secondary pollutants are mainly formed from reactions involving sulfur dioxide and mono-nitrogen oxides (Scottish Air Quality, 2012a).

Air pollution can have both a short term and long term effect on health. Those with lung or heart conditions can experience a short term increase in symptoms when they face increased exposure to air pollution. Asthmatics, who suffer from a common form of lung condition, may notice an increased need to use a prescribed inhaler. The general population may experience a dry throat and sore eyes when subjected to very high levels of air pollution in a relatively short period of time. Long term or elevated long term effects of air pollution can lead to serious conditions which are detrimental to the health of an individual. These conditions mainly effect the respiratory and inflammatory systems but have also been shown to lead to cancer and heart disease (Scottish Air Quality, 2012b). Each pollutant can affect the human body differently. Nitrogen dioxide, sulphur dioxide and ozone can irritate the lungs and increase the symptoms of lung disease for those suffering. Particles can be inhaled deep into the lungs where they can then cause a worsening of heart and lung disease. Carbon Monoxide can lead to a reduction in oxygen reaching the heart in those suffering with heart disease.

In Britain, the negative effects of air pollution were not taken seriously until The Great Smog (or The Big Smoke) in 1952 (Met Office, 2012). A vast cloud of smoke descended over London for four days making it almost impossible to see only a few feet causing the transport system to come to a halt with reportedly more than 4,000 casualties, although some sources claim that the death toll was more likely around 12,000 (Bell *et al.*, 2004). These deaths were the result of a combination of a mixture of pollutants and adverse weather conditions. Usually the smoke from coal burning would rise into the atmosphere and disperse, however an anticyclone blocked this. An anticyclone, described by the The Oxford English Dictionary (2012) as a large-scale circulation of winds which centre around a region of high atmospheric pressure, resulted in the smoke being forced downwards causing a thick smog. London had previously experienced similar events but none were as significant as this in terms public awareness of the health effects of pollution and the resulting research and regulation. The UK government reacted to the catastrophic London smog and as a result the Clean Air Acts of 1956 and 1968 were passed (Met Office, 2012).

Sixty years on from the great smog and air pollution awareness and action is at the forefront of policy and research across the world. It is widely accepted in the scientific community that an increase in and long term exposure to air pollution can have a negative effect on health. One notable study by Dockery *et al.* (1993) focused on the implications of long exposure to air pollution by conducting a cohort study. This study followed up 8111 adults across 6 U.S cities over a period of 14 to 16 years and found that after controlling for smoking habits and other risk factors that there was a statistically significant association between air pollution and mortality and that air pollution was positively associated with lung cancer deaths and cardiopulmonary disease. Another cohort study focused on air pollution effects by Pope III *et al.* (1995) which used ambient air pollution data form 151 U.S. metropolitan areas in 1980. This study tracked over 500,000 adult residents and recorded their morbidity rates in 1989 and the research found that particulate air pollution was associated with cardiopulmonary and lung cancer mortality. The study by Dominici *et al.* (2006) looks at short-term exposure to air pollution by looking at time-series data for hospital admission rates and ambient air pollution levels, as well as temperature data between 1999 and 2002 with the conclusion that short-term exposure increases the risk of hospital admission for cardiovascular and respiratory diseases.

The increased level of awareness has led to the measurement of air pollution in countries across the world. The European Environment Agency (EEA) (European Environment Agency, 2012) in partnership with the European Environment Information and Observation Network (EIONET, 2013) monitor air pollution levels across European countries. The Eionet and the co-operating countries supports the collection and organisation of data. This enables the EEA to provide information to government bodies and institutions as well as the general public with a view to evaluating the data to understand the surrounding environment and to possibly affect policy. This ensures that governing bodies and decision makers as well as the general public are given access to relevant data and are well informed about environmental affairs.

The collection and analysis of information on environmental data across the years has led to the regulation of air pollutants. The European Union has regulations set out (discussed in section 1.1.1) which its member countries must adhere to. If a country does not meet the targets they could be subject to a fine. In addition to this, the Scottish Government have outlined a more strict set of air quality guidelines and targets to which it strives to achieve across the country. The Department for Environment, Food and Rural Affairs (Defra) and the government run Scottish Air Quality are the regulators and monitoring bodies in the UK and Scotland, respectively.

Particulate matter is one of the most regulated and therefore regularly monitored pollutants across Europe. Particulate matter, also known as PM_{10}, are particles which measures 10 micrometers or less. These particles are small enough that they are likely to be inhaled into the human body which can result in significant damage to internal organs. Particulate matter consists of a mixture of solid and liquid particles and various processes such as power plants and fossil fuel burning can produce PM_{10}. Naturally, PM_{10} can occur from volcanoes, vegetation and domestic fires. Road transport, coal burning and construction are the major sources of PM_{10}, all of which you would expect to observe in a large city such as Glasgow. PM_{10} is the pollutant chosen to produce an air pollution indicator for this thesis.

The existence of a relationship between air pollution and meteorological data has been clear for a number of years. Ambient temperature is the most commonly included covariate in air pollution studies and the effect of temperature in morbidity rates is becoming an increasingly important issue (Ye *et al.*, 2012). As previously mentioned, the combination of air pollution and adverse weather effects were the cause of the Great London Smog. This suggests that temperature and related weather effects, such as humidity, could be a confounding factor of air pollution.

1.1.1 Existing Air Pollution Standards

Currently air pollution standards are set by different bodies. The European Union has set up a large body of legislation which provides objectives for a number of different pollutants which are set to establish health based standards across Europe. The long term objective of the EU is to achieve levels of air quality that do not result in unacceptable impacts and risks to human health and the environment (European Parliament Council, 2002). If countries in the EU fail to meet the European standards they can be subject to large fines. Recently the United Kingdom supreme court ruled that the UK government had failed in their efforts to meet European air pollution limits (The Supreme Court, 2013).

Defra published the Air quality Strategy for England, Scotland, Wales and Northern Ireland (Defra, 2007) which outlined air quality objectives and strategies to improve air quality in the UK long term. The devolved administrations of Scotland, Wales and Northern Ireland set their own air quality targets whilst the Defra publication combines the targets for all parts of the UK. Table 1.1 is taken from the Defra air quality strategy publication (Defra, 2007) and outlines the air quality objectives for PM_{10} for the UK and the Scotland specific objectives. The table details both the UK and the Scotland specific targets, set by the devolved government, and the corresponding objective with the date in which the objective must be met. The UK annual mean objective states that PM_{10} should not exceed $40\mu gm^{-3}$ nor should the 24 hour mean exceed $50\mu gm^{-3}$ more than 35 times a year, these targets should have been implemented by the 31st December 2004 for the UK. The Scottish annual mean objective, however, states that the PM_{10} annual mean of $18\mu gm^{-3}$ should not be exceeded nor should the 24 hour mean exceed $50\mu gm^{-3}$ any more than 7 times a year, this objective should have been achieved and maintained by the 31st December 2010. While the Scottish objective is much stricter than the EU and UK objectives, they are all set using different time scales.

Table 1.1: National air quality objectives and European Directive limit and target values for the protection of human health

Pollutant	Applies to	Objective	Concentration measured as	Date to be achieved by and maintained thereafter
PM_{10}	UK	$50\mu gm^{-3}$ not to be exceeded more than 35 times a year	24 hour mean	31 Dec 2004
PM_{10}	UK	$40\mu gm^{-3}$	annual mean	31 Dec 2004
PM_{10}	Scotland	$50\mu gm^{-3}$ not to be exceeded more than 7 times a year	24 hour mean	31 Dec 2010
PM_{10}	Scotland	$18\mu gm^{-3}$	annual mean	31 Dec 2010

1.2 Discussion of Existing Indicators and Indexes

An environmental indicator or index is a simple statistic which provides an idea of the state of one part of the wider environment. These indicators are used by the government, non-government organisations, and research centres to establish the state of the environment. It provides these organisations with information on whether targets are being met and provides the general public with easy and simple information. Indicators can be an effective way to condense a large amount of data into a simple numerical summary. However, as there is no set way of producing an indicator this can lead to confusion and transparency issues. There are a number of environmental indicators and indexes available which have been constructed using various methods. The construction of indicators and indexes can affect their interpretability and robustness and therefore it is key that the steps in their construction are well thought out and transparent so as to keep the reader fully informed. The way in which an indicator is constructed can differ in the selection process, weighting, and aggregation. When constructing an indicator with multiple pollutants or factors that are believed to not be equal in relation to the subject of the indicator a weighting process is used. The factors are assigned a weight according to how important each factor that make up the indicator is believed to be. For example, household income could have a larger weighting than the percentage of hospital admissions in relation to constructing an indicator of deprivation. There is no set way to calculate this weight but it is usually assigned with the input of an expert on the topic. An aggregation process is used when there are multiple factors which need to be combined to produce an indicator. For example, five pollutants could be combined using aggregation to produce an air pollution indicator.

A composite indicator is constructed by compiling single indicators into one single index. In Tarantola and Saltelli (2008), the authors discuss the use of composite indicators for policy and decision making and put forward their own suggestions to improve the development of composite indicators. The authors provide the reader with a bad and good example of an indicator. The bad indicator was poorly weighted which leaves scope for misinterpretation. The authors state that this could be avoided if the indicator composition is made fully transparent, which they claim is almost never the case in mainstream media. The good indicator is based upon reliable and high quality data which is then weighted according to 19 different sources of subjective information. The publication proceeds to discuss robustness and sensitivity analysis and their key role in developing a composite indicator. The need for robustness and sensitivity analysis comes from the subjective building of composite indicators. There is no set way to build a composite indicator. There are many decisions throughout the process which are subjective, such as the weighting of indicators and the treatment of missing values. An article by Cherchye *et al.* (2007) also focuses on the design issues involved in constructing an indicator which can leave the index open to misinterpretation by the media and general public. These papers are clear that an indicator should be transparent and understandable to ensure that they are not open to miss-interpretation.

A widely used index in Scotland, known as the Scottish Index of Multiple Deprivation (SIMD), is outlined in the 2009 technical report (Office of the Chief Statistician, 2009). This index combines 38 indicators across 7 domains: income, employment, health, education, skills and training, housing, geographic access and crime. The index is made up of 7 domains which have been weighted based on the domains' importance in measuring deprivation and the robustness of the data. These weighting are published along with the index to ensure complete transparency. This index, however, does not take an environmental factor into consideration which suggests that a stand alone environmental indicator one which could be incorporated into the already existing SIMD could be an important next step in defining deprivation. The paper by Richardson *et al.* (2010) researches the spatial inequality of socioeconomic deprivation. The paper states that it is likely that the environment has a part in this spatial inequality. The paper moves on to develop two measures of health related multiple physical environmental deprivation for small areas. The two summary measures are named: the multiple environmental deprivation index (MEDIx) and classification (MEDCLASS). Four stages are carried out in developing the deprivation index including identifying UK specific environmental issues, acquiring the relevant data, checking associations between environmental dimensions and then finally constructing the summary measures. To construct the summary measures different environmental dimensions were recognised to be either beneficial or detrimental to human health. The index is then produced by looking at the distribution of values for each environmental index across the UK by constructing quintiles and those areas that are in the highest quintile are given a score or +1 if the dimension is thought to be detrimental and -1 for beneficial dimensions. The scores then range from -2 to +3 for areas in the UK. These scores are then classified using a two step clustering process. This indicator is constructed to provide an insight into the environmental effect of widening disparities in health in the UK. This discussion has identified some of the issues in choosing what dimensions to include in indicators or indexes. The final index or indicator is heavily dependent on which dimensions are included. The air pollution indicator, discussed in this thesis, would likely have a different conclusion depending on which pollutant is included which must be considered when interpreting the indicator.

Moving onto air pollution, an article by Lee *et al.* (2011) outlines a method for producing air quality indicators which results in an indicator for Greater London for August 2006. Three common issues are addressed in this article: which pollutants should be included, how these pollutants should be combined and in which order should space and pollutants be aggregated. A further two issues, which the authors claim have not been addressed in the literature previously, were firstly, how to produce an uncertainty measure and secondly how to address the issue of spatial representativeness of the data. In the first stage, the pollution data $Y_{t,j} = (Y_{t1j}, \ldots, Y_{tnj})$ is aggregated over space to estimate the average concentration across the study region which is denoted as R. Y_{tij} is the automatic monitoring site data and $j = 1, \ldots, p$ denotes the pollutant number, t denotes the time point and monitoring site location is denoted as $i = 1, \ldots, n$. The spatially-aggregated estimate is calculated using

$$\widehat{S}_{t,j} = \frac{1}{n} \sum_{i=1}^{n} Y_{tij}. \tag{1.1}$$

The second stage is to aggregate over pollutants $j = 1, \ldots, p$ as the estimates $\widehat{S_{t1}}, \ldots, \widehat{S_{tp}}$ are required to be combined. To overcome the issue of dominance from one pollutant to different orders of magnitude, the pollutants are re-scaled to get S_{tj}. The indicator is then constructed using

$$\widehat{AQI}_t = \frac{1}{p} \sum_{j=1}^{p} \widehat{S}^{*}_{tj}, \tag{1.2}$$

where $S^{*}_{tj} = \widehat{S}_{tj}/C_j$ and C_j is a pollutant-specific standardised value.

Lastly, the accuracy of the air pollution indicator is explored by looking at the amount of variation that could lead to errors and uncertainty estimates, how spatially correlated each pollution is, the number of monitors for each pollution, and the spatial locations of the monitors. Each of these factors could have an effect on the bias and uncertainty of an indicator. Two approaches were proposed for stage one, to aggregate the pollutants. The first approach takes each pollutant and represents them using a Bayesian geostatistical model assuming that the monitoring stations are independent. This model, where $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ denotes the observed daily average concentration of a pollutant at each of the n monitoring sites, is described in Equation (1.3). Let S_i be the natural logarithm of the true population pollutant value at location x_i. Then the set of the true values $\underline{S}$ is denoted by a linear regression model with covariates Z and regression parameters δ. Spatial variation is controlled by σ^2 and the level of the nugget effect is denoted by $v^2\sigma^2$ and the spatial correlation matrix $\Sigma[\phi]$ is specified by the Matern class of functions with the range parameter ϕ and fixed smoothness parameter k.

$$\begin{aligned}
ln(\underline{Y_i}) &\sim N(S_i, v^2\sigma^2) \text{ for } i = 1, ..., n;\\
\underline{S} = (S_1, \ldots, S_n) &\sim N(Z\delta, \sigma^2\Sigma[\phi]);\\
\underline{\delta} &\sim (\mu_\delta, \Sigma_\delta);\\
v^2 &\sim \text{beta}(a, b);\\
f(\sigma) &\propto 1;\\
f(\phi) &\propto \frac{1}{\phi} I[\phi \epsilon \phi_1^*, \ldots, \phi_r^*].
\end{aligned} \tag{1.3}$$

The second approach is an extension of the model used in the first approach which has been modified to allow for preferential sampling. Preferential sampling occurs when the value of the process being modelled (air pollution in this case) plays a role in where the process is monitored. In this case pollution monitors are typically located where concentrations are thought to be highest, so the worst case scenario can be observed. Therefore Diggle *et al.* (2010) extended the geostatistical model by allowing for this dependence between the locations at which the process was observed and the values of the process. Thus essentially, they additionally model the locations of the monitors as random quantities with a point process, rather than assuming they are fixed. After a thorough assessment of the approaches using simulated data and data for Greater London the authors conclude that both approaches perform well in terms of bias and root mean square error (RMSE). The first approach in which the model assumes independence between stations displays almost no bias and very low RSME for both types of data. The second approach, which allows for preferential sampling favors the data which is preferentially sampled but gives low bias and RSME for each case. Both of these approaches compare well against the existing method of using simple numerical summaries of the data. This paper gives a clear outline of the construction of an air quality indicator. Despite the more complex nature of the Greater London indicator a number of issues raised are similar in nature to issues faced in constructing the indicator for Glasgow including selecting pollutants and constructing a geostatistical model.

A general class of air quality indicators is proposed in Bruno and Cocchi (2002) which focuses on comparing situations in time and space, in particular when there are multiple monitoring stations in the one area. The paper works through an example where the data are collected according to the three dimensions: time, space and the type of pollutant. Firstly the aggregation process begins with the aggregation over time. The function in $Y_{qij} = q(Y_{ijt})$ is applied to the hourly monitoring data where Y_{itj} denotes the primary data where $i = 1, \ldots, I$ indexes the sites, $j = 1, \ldots, J$ indexes the pollutants and $t = 1, \ldots, T$ indexes the time occurrences of the observations. This function q produces an $I \times J$ matrix where each row contains the time synthesis of each pollutant at each ith site. The second step is to standardise for pollutants which can be done using a simple or complex method. The more complex method uses the health consequences of each pollutant. This is done by classifying the pollutants according to the different health risks, $c = 1, \ldots, C$. The standardising transformation in $f_R(Y) = \frac{b_{c+1} - b_c}{a_{(c+1)j} - a_{cj}}(Y - a_{cj}) + b_c$ is then used where a_{cj} represents the threshold that define the air quality classes for each pollutant and b_c denotes the standardised thresholds.

The order of the next two steps in then explained to be extremely important. There are two possible options: aggregating among the monitoring sites and then among pollutants or aggregating among the pollutants and then among monitoring sites. These two aggregation options are then discussed together to highlight the similarities and differences that arise by using a different aggregation order. Although, the Glasgow based air pollution indicator does not require an aggregation process over different pollutants, if this indicator was to expand to include other pollutants the author highlights some important aggregation issues.

1.3 Aims

There are three main aims in this thesis. The first aim is to explore statistical methods in order to model and summarise the distribution of PM_{10} levels. In order to investigate how PM_{10} levels are distributed across time and space a suitable analysis of two main datasets (PM_{10} monitoring site data and annual mean PM_{10} model data) is carried out. This analysis provides a starting point for building a model which combines both the time series and spatial aspect of the selected pollutant. The second aim is to produce a spatio-temporal model which accounts for the similarities and dissimilarities between PM_{10} across time and space. Lastly, the major aim for this thesis is to use what has been studied in the previous two aims to produce an air pollution indicator based on PM_{10} for Glasgow. This indicator can then be used as an easy and convenient way to assess Glasgow's PM_{10} levels as a whole.

1.4 Overview of Thesis

Two main datasets are discussed and analysed in this thesis. The first being the PM_{10} monitoring site data which contains the average level of PM_{10} each day across 11 different monitoring station sites across Glasgow and the second is the previously modelled annual mean PM_{10} for a 1 × 1 km map across Glasgow. Both of these data sites are analysed for only 3 years due to the availability and quality of the data.

Chapter 2 provides the reader with a detailed explanation of the trends and patterns of PM_{10} monitoring site data in Glasgow. This chapter then progresses on to find a suitable model which explains PM_{10} at each of the sites. The model incorporates an accompanying meteorological data set which provides daily averages for temperature and humidity amongst others. This model is not designed to be the best fitting model but a suitable model which can be used to provide an insight into the similarities and dissimilarities of PM_{10} across space and time. Once a suitable model has been decided upon conclusions and inferences can be made about changing levels of PM_{10} across the three years and the differences between the monitoring sites. This is essential in understanding the PM_{10} levels across the years and the monitoring site locations and is one step towards finding an overall description of PM_{10} for the whole of Glasgow.

Chapter 3 explores the previously modelled annual mean PM_{10} map data across the years 2010-2012. The spatial aspects of PM_{10} are first explored using numerical and graphical summaries. A more formal approach is used to then produce a geostatistical model to explain the trend of PM_{10} across Glasgow.

Chapter 4 then focuses on producing naive indicators using each of the data sets and the modelling and exploratory analysis conducted in Chapters 2 and 3. The advantages and disadvantages of these indicators are the basis for a combined spatio-temporal model which accounts for both the spatial and temporal aspects of the data. This modelling process results in a final indicator estimate for Glasgow which provides inferences and conclusions about the distribution of air pollution across Glasgow.

Chapter 5 ends the thesis with a discussion of the final indicator and the conclusions with consideration given to improvements which could be made and additional analysis for the future.

1.5 Data Description

This section gives a brief description of the data used in this thesis. The origin of each of the data sets, the variables in each data set and the measurement process are each explained in this section. Both the air quality and the weather data were extracted from publicly available online sources. The nature of the data meant that it had to be cleaned and manipulated to ensure it was fit for purpose. This included converting files to different formats, removing incomplete or redundant data and also reformatting data, such as dates.

1.5.1 PM_{10} Monitoring Site Data

The Air Quality data used were obtained from the Scottish Air Quality website (Scottish Air Quality, 2012a). This website, run by the Scottish Government, ensures that the data measured by the monitoring site is easily accessible and up-to-date. A comprehensive system of data verification and ratification was put into place by the Scottish Air Quality department to ensure that real-time data could be provided. There are various methods for monitoring air quality with automatic monitoring sites being one of the most accurate as it limits human error and can provide high temporal resolution data. Along with real time data simple statistics including daily maximum, minimum and daily mean PM_{10} values are available. There are over 80 automatic monitoring stations in Scotland which measure a variety of pollutants including PM_{10}, $\text{PM}_{2.5}$, Nitrogen Dioxide (NO_2), Ozone and Sulphur Dioxide (SO_2). Some of these sites have been running since the mid 1980s and there is available data which goes back to 1986. The concentrations for each pollutant are measured in $\mu g m^3$.

Daily mean concentrations of PM_{10} are available for 11 automatic monitoring stations around Glasgow, as shown in Table 1.1 and Figure 1.2. These locations are not equally spaced throughout Glasgow and there is no suggestion that these are a representative sample of Glasgow as a whole. Monitoring sites are classified according to the environment in which they are situated. This is an important aspect to fully understanding the data. The Scottish Air Quality website has 10 different monitoring site classifications, 4 of which appear in the Glasgow sites shown in Table 2.1. The most common in this data is the *roadside* classification, sites of this classification are between one meter of the kerbside of a busy road and the pavement which will usually be within five meters of the road. These sites are measuring high values due to the local traffic and are used to evaluate vehicle emission objectives and schemes set up to reduce traffic. The site classification *urban* traditionally has monitoring sites located in built-up urban areas where there are big open squares and very little or no traffic. These measure vehicle emissions, commercial and space heating and are used to identify long-term urban trends. *Urban central* is very similar to *urban* in that they are there to measure similar sources of emissions but are specifically at locations within city centres where there are pedestrian or shopping areas. *Rural* stations, unlike the other classification are situated in open countryside locations, as far as possible from roads or populated or industrial areas. These sites are used to measure long- range transport and urban emissions.

The locations of the monitoring stations in Glasgow, shown in Figure 1.2, shows the spread of the sites, how spatially similar the sites are and give us an idea of which sites we may expect to have similar PM_{10} time series. There is a relatively linear line of eight sites running from the west through the centre to the east of the city along the north side of the River Clyde. There are a further two sites in the south side (Nithsdale Road and Battlefield Road) which are relatively spatially similar and lastly one site which is located on the south-west border (Waulkmillglen Reservoir) which is the site furthest away from the city centre and in fact the only rural classified monitoring site location.

Figure 1.1: Site classification for each site

	Classification
Abercrombie Street (AS)	Roadside
Anderston (A)	Urban
Battlefield Road (BTR)	Roadside
Broomhill (B)	Roadside
Burgher Street (BS)	Undisclosed
Byres Road (BR)	Roadside
Centre (C)	Urban Centre
Dumbarton Road (DR)	Roadside
Kerbside (K)	Kerbside
Nithsdale Road (NR)	Roadside
Waulkmillglen Reservoir (WR)	Rural

1.5.2 Meteorological Data

To accompany the PM_{10} monitoring site data various aspects of meteorological data are available from the Weather Underground website (Weather Underground Network, 2012) which is part of The Weather Channel Companies. This data are publicly available and consist of various simple statistics involving different aspects of meteorology. Unfortunately, meteorological data is not available at each of the monitoring sites that measure air pollution as specified above. The most reliable source of weather data for Glasgow, as a whole, is Glasgow International Airport, Paisley. The historical data dates back to 1994 and a central database collects these weather readings daily and processes and formats them to make them available online. The Glasgow station provides an hourly report of weather events in and around the station.

Various aspects of meteorological data were available for years 2010 to 2012. Temperature and relative humidity have been explored as having a

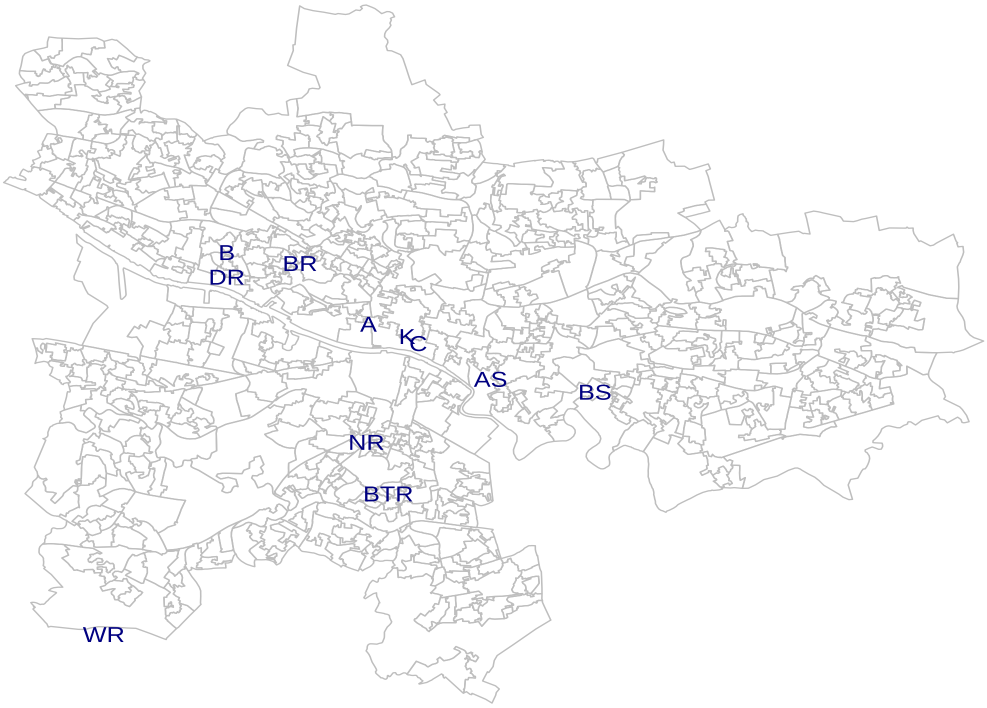


Figure 1.2: Locations of Monitoring Stations in Glasgow

relationship to PM_{10} in papers such as Barmpadimos *et al.* (2011) and Yusof *et al.* (2008) and therefore were included in the study. The temperature variable is measured in °C and hourly mean values are available. Humidity measures the amount of water vapor in the atmosphere and is measured as a percentage. In a general sense, it is the amount of moisture in the air compared to what that specific atmosphere is capable of holding.

1.5.3 Modelled Annual Mean PM_{10} Data

The modelled annual mean PM_{10} data were also obtained from the Scottish Air Quality website (Scottish Air Quality, 2012a). Annual mean PM_{10} concentrations were modelled in 2010 for Scotland at background and roadside locations. The methodology used was based on the UK Pollution Climate Mapping approach explained in the DEFRA website (Defra, 2013a), however, the Scotland specific model used appropriately scaled Scottish PM_{10} monitoring data concentrations along with secondary aerosols, particles from long range transport, iron and calcium based dusts and Scottish meteorological data only to model the concentrations for Scotland. Annual mean concentrations were modelled for the year 2010 then projected forward for years 2015, 2020, 2025 and 2030 with intermediate years being linearly interpolated. The model output data is available for each local authority in Scotland and consists of background concentrations for each 1×1km grid square. Accompanying the background concentrations is the contribution from each emissions sector as well as the grid co-ordinates. The attributing emissions concentrations include motorways, A and B roads, and railroads.

The modelled PM_{10} data values were presented in the form of a lattice shown in Figure 1.3 where each circle represents a location (s_i). In the plot the previously discussed PM_{10} monitoring site locations are also marked, giving an idea of the relative position of these two PM_{10} data sources.

With each of the two main data sets described and the aim of the thesis explained the next chapter focuses on summarising both sets of data before

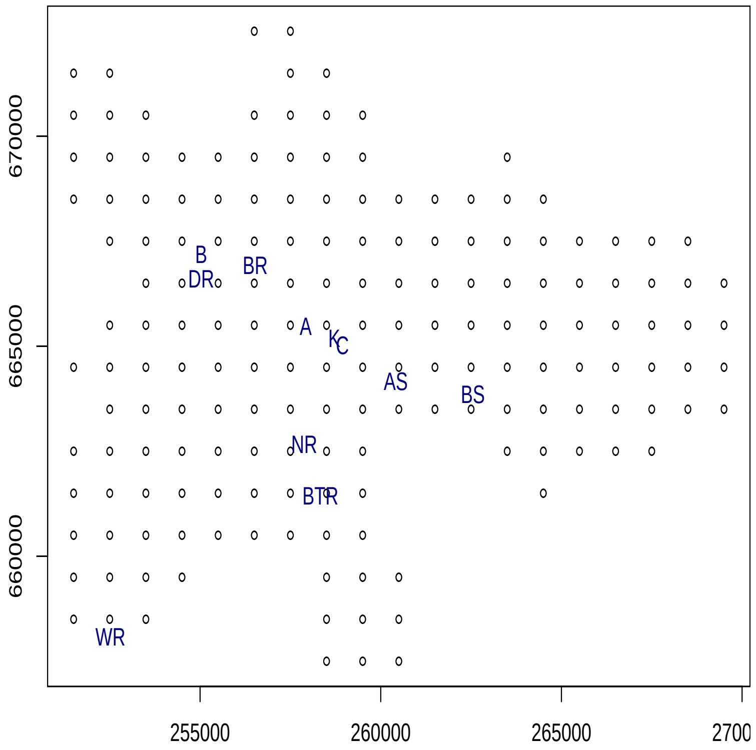


Figure 1.3: 1 km x 1km grid location in Glasgow

any modelling or inferences can be made.

Chapter 2

Exploring Trends and Seasonality of PM_{10} Monitoring Site Data

In order to produce an air pollution indicator for Glasgow using PM_{10} as the indicator pollutant it is necessary to have an idea of how PM_{10} is distributed through time and through space. The PM_{10} monitoring site data, discussed in Section 1.5, is used to explore the distribution of air pollution across time at a number of locations across the city. In this chapter, possible trends and seasonality within the PM_{10} monitoring site data and the relationships between the covariates (humidity and temperature) are explored informally by means of graphical and numerical summaries and linear regression. Linear regression modelling is employed as a more formal exploratory tool, which uses the knowledge gained in exploring the two data sets, to assess the trend and seasonality and the relationship between PM_{10} and the meteorological variables. This method has to relax the assumptions of a traditional linear regression to allow us to examine the dependence in the residuals. The next step after this is to consider a model with a more complicated covariance structure for the errors which allows for autocorrelation. The chapter then moves onto model checking and interpretation of the model output. The analysis provides information about how PM_{10} is distributed temporally and spatially which could hence inform about the distribution of air pollution in Glasgow. The air pollution information from this chapter will be the starting point of an air pollution indicator in Glasgow.

2.1 Methods

2.1.1 Exploratory Methods

Exploring Model Variables Using Linear Regression

Firstly the discussion starts with a brief outline of a simple regression model where y_t is the response variable which in this case is $\log(\mathrm{PM}_{10})_{,t}$ for $t = 1, \ldots, T$. Assuming that the response variable is being influenced by a series of explanatory variables $x_{k,t}$ where $k = 1, \ldots, K$ and $t = 1, \ldots, T$, the relationship between PM_{10} and the explanatory variables is described by the linear regression model

$$y_t = \beta_0 + \beta_1 x_{1,t} + \ldots + \beta_K x_{K,t} + \varepsilon_t. \tag{2.1}$$

Here $(\beta_1, \ldots, \beta_K)$ are the unknown, fixed regression coefficients and $\{\varepsilon_t\}$ is the random error term which, assuming non correlated errors, is assumed to have mean zero. The unknown parameters in the linear regression model were estimated using ordinary least squares (OLS).

Taking the linear model as above where the data consists of T observations of which each has a corresponding response y_t and a number of explanatory variables K, the model can also be written in matrix notation:

$$Y = X^T\beta + \varepsilon, \tag{2.2}$$

where

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix}, X = \begin{pmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,T} \\ x_{2,1} & x_{2,2} & \dots & x_{2,T} \\ \vdots & \vdots & \dots & \vdots \\ x_{K,1} & x_{K,2} & \dots & x_{K,T} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_T \end{pmatrix}.$$

The OLS method computes the regression lines in search of the line of best fit which minimises the sum of squared vertical distances from the line to the observed points. The residual value is the vertical distance between the observed and fitted points and the regression line and therefore can be used to assess the degree of fit of the model. The residual sum of squares (RSS) is a measure of the overall fit denoted by equation (2.3) where β denotes the possible values for the parameter and the value of β which minimises the RSS is the OLS estimator is denoted as $\widehat{\beta}$. The OLS estimator $\widehat{\beta}$ in matrix form is shown in equation (2.4).

$$S(\beta) = (Y - X\beta)^T(Y - X\beta). \tag{2.3}$$

$$\widehat{\beta} = (X^TX)^{-1}X^TY. \tag{2.4}$$

A number of assumptions are made by standard linear regression models which use the estimation technique OLS, these must hold for the model estimates to be accurate. Firstly, the assumption of homoscedasticity means that the errors must have constant variance, this can be checked by looking for a fanning or unequal trend looking at a plot of the residuals. The assumption of normality must also hold which can be checked using a histogram or more formally a normal Q-Q plot. Lastly, the assumption that the errors are uncorrelated with each other; this is often not be the case for time series data with significant autocorrelation

As explained in Section 1.5, there is a high proportion of missing data in the monitoring site time series data. The use of linear regression modelling with time series data, especially data which has a large proportion missing, should be used only with a considerable amount of care. The linear regression function used ignores the missing values. Failure to account for autocorrelation in the regression model means that the standard errors and p-values are unreliable but the OLS fit will be used as a rough guide as to how well the model fits the data.

Harmonic Regression

In the case where there appears to be cyclical or seasonal patterns across time, one or many harmonic functions can be used to attempt to capture the seasonality. Basic harmonic regression comes from the equation discussed in Kupper (1972),

$$y_t = \beta_0 + A\cos(2\pi wt + \psi) + \varepsilon_t, \tag{2.5}$$

where y_t is the response variable which in this case is log(PM$_{10}$), 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 is the 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

$$\cos(\alpha^{+}_{-}\beta) = \cos(\alpha)\cos(\beta)^{-}_{+}\sin(\alpha)\sin(\beta), \tag{2.6}$$

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

$$A\cos(2\pi wt + \psi) = \beta_1\cos(2\pi wt) + \beta_2\sin(2\pi wt). \tag{2.7}$$

Here $\beta_1 = A\cos(\psi)$ and $\beta_2 = -A\sin(\psi)$ and therefore the model can be written in the linear regression form

$$y_t = \beta_0 + \beta_1\cos(2\pi wt) + \beta_2\sin(2\pi wt) + \varepsilon_t. \tag{2.8}$$

Linear terms such as temperature and humidity can be easily included in the model, for example we could have

$$y_t = \beta_0 + \beta_1 \cos(2\pi wt) + \beta_2 \sin(2\pi wt) + \beta_3 \text{Temperature} + \beta_4 \text{Humidity} + \varepsilon_t. \tag{2.9}$$

Amplitude and Phase Estimation

In order to display the harmonic regression terms in a more meaningful way, the estimated amplitude ($\widehat{A}$) and phase values ($\widehat{\psi}$) values were calculated. The amplitude is the height of the wave from zero and the phase explains where in the cycle of the function is the oscillation at t=0, which provides an idea of the angle of the function.

The standard harmonic regression has the components $\widehat{A}$ and $\widehat{\psi}$, where w is the cycle component which determines the frequency of the wave; t is the time component and β_0 is the intercept term. We have

$$\widehat{A} = \sqrt{\widehat{\beta}_1^2 + \widehat{\beta}_2^2} \tag{2.10}$$

with

$$\widehat{\psi} = \arctan(-\frac{\widehat{\beta}_2}{\widehat{\beta}_1}). \tag{2.11}$$

Simulation was used in order to estimate the standard error values of $\widehat{A}$ and $\widehat{\psi}$. Firstly $\widehat{\beta}_1$ was simulated 1000 times using the normal distribution with the mean equal to $\widehat{\beta}_1$ and standard deviation equal to the standard error of $\widehat{\beta}_1$ and was denoted by β_{1sim}. This process was then repeated for $\widehat{\beta}_2$ which was then denoted by β_{2sim}. From this distribution, an A_{sim} was calculated using $A_{sim} = \sqrt{\beta_{1sim}^2 + \beta_{2sim}^2}$and ψ_{sim} was calculated using $\psi_{sim} = \arctan(-\frac{\beta_{2sim}}{\beta_{1sim}})$. Then the standard errors were calculated by calculating the standard deviations of A_{sim} and ψ_{sim}.

Residual Diagnostics

In order to assess the model assumptions after the model has been fit, we examine the results which are defined by $r_t = y_t - \widehat{y_t}$ where $\widehat{y_t}$ is the fitted values at time t. When the residuals are plotted against time t, they should have a mean of 0 and an equal spread above and below the mean with no fluctuations in the variation. The residuals of a model can alert you to problems with assumptions made when modelling. When modelling time series data it is important to look out for autocorrelation in the residuals. Failing to adequately account for the autocorrelation in time series data can lead to biased results. The most common way to check for autocorrelation in the residuals is using a sample autocorrelation function acf and partial autocorrelation function (pacf) plot which is discussed in the next methods section.

2.1.2 Time Series Regression Model Methodology

Stationarity

A stochastic process $\{y_t\}$ is strictly stationary if the joint probability distribution does not change when shifted in time and as a result the mean and variance (when they exist) do not depend on t and are finite and the autocovariance and autocorrelation functions only depend on the lag ((weak) stationarity).

2.1.3 Autocorrelation

When modelling PM_{10} it is reasonable to assume that short term correlation may be present. Short term correlation arises when the level of PM_{10}, for example, on one day is related to the level of PM_{10} the following data or the previous day - this is classed as a lag one autocorrelation. A relationship between values two days apart is classed as lag 2 autocorrelation, and so on. The correlation can be assessed using acf and pacf plots.

Acf and Pacf

As discussed in Box *et al.* (2008), the acf plot considers the linear relationship between two values τ lags apart. The autocorrelation function at lag τ where Y_t is a random variable at time t is as follows

$$\begin{aligned} p_\tau &= \text{corr}[Y_t, Y_{t+\tau}] \\ &= \frac{cov[Y_t, Y_{t+\tau}]}{\sqrt{Var[Y_t]Var[Y_{t+\tau}]}}. \end{aligned} \tag{2.12}$$

The pacf is below,

$$\begin{aligned} \alpha_1 &= \text{Corr}(Y_t, Y_{t+1}) \\ \alpha_\tau &= \text{Corr}(Y_{t+\tau} - P_{t,\tau}(Y_{t+\tau}), Y_t - P_{t,\tau}(Y_t)), \text{for} k \geq 2, \end{aligned} \tag{2.13}$$

where $P_{t,\tau}(x)$ denotes the projection of x onto the space $Y_{t+1}, \ldots, Y_{t+\tau-1}$. Under stationarity the numerator in Equation (2.12) is the autocovariance function for lag τ and the denominator is the autocovariance function for lag 0. In the acf and pacf plots if there is a breach of the confidence bands at a certain lag then there could be correlation remaining in the residuals at said lag. The pattern of lags that breach the confidence bands gives an idea if there is autocorrelation and which combination of autoregressive moving average (ARMA) processes would be appropriate to model this.

If there is autocorrelation of the errors then the assumption that error terms are uncorrelated is breached. Missing values are not allowed for either the acf or the pacf plots and the function merely passes through the missing values and estimates the autocovariance from only the complete values. The large amount of missing values in the data mean that the acf and pacf plots can only be used as a rough guide of autocorrelation.

ARMA

This takes us on to the ARMA process which takes the random error term $\{\varepsilon_t\}$, in equation 2.14, and makes some change to the sequence of the random noise process to allow for autocorrelation. A moving average process ($\varepsilon_t \sim MA(q)$) simply applies a linear function to the errors $\{\varepsilon_t\}$ which can take the form

$$\varepsilon_t = Z_t + \sum_{k=1}^{q} \theta_j Z_{t-k}, \tag{2.14}$$

where $Z_t \sim N(0, \sigma^2)$. In the case of the autoregressive process ($\varepsilon_t \sim AR(p)$), each $\{\varepsilon_t\}$ depends on the value of its' predecessor $\{\varepsilon_{t-1}\}$:

$$\varepsilon_t = \sum_{i=1}^{p} \phi_i \varepsilon_{t-i} + Z_t. \tag{2.15}$$

Taking these two cases together to give an ARMA process, ($\varepsilon_t \sim ARMA(p, q)$) which is

$$\varepsilon_t = \sum_{i=1}^{p} \phi_i \varepsilon_{t-i} + Z_t + \sum_{j=1}^{q} \theta_j Z_{t-j}. \tag{2.16}$$

In the above models p is the autoregressive order and q is the moving average order. The method of fitting an ARMA model used in this thesis is outlined in Gardner *et al.* (1980) and uses an algorithm for Exact Maximum Likelihood (EML) using the state-space approach Kalman filtering. In summary there are two processes being performed with the first transferring the model into state-space form and then calculating the covariance matrix for the first value of the state vectors. The second process computes recursions and prediction errors with the covariance matrix determinant. These two processes combined produce the exact likelihood. This can then be maximised using iterations to yield the EML estimate. The state-space approach of Kalman filtering is a convenient and transparent way of modelling ARMA processes with missing values, these details are outlined in Durbin and Koopman (2001).

2.1.4 Model Checking and Selection

AIC

Akaike's Information Criterion (AIC) provides a measure of the goodness of fit whilst considering the complexity of the model which can be used in model selection (Akaike, 1974). The AIC does not give a measure which is tested against a null hypothesis but a measure to compare models. AIC is defined to be

$$AIC = 2k - 2log(L), \tag{2.17}$$

where k is the number of parameters in the model and L is the maximised likelihood function for the estimated model.

Q-Q plot

The quantile-quantile plot (Q-Q plot) is another method of model checking that compares the empirical quantiles for the data against the quantiles of an assumed model. In this context of time series regression we want to assume that the residuals are normally distributed and hence the quantiles of the residuals are plotted versus the normal quintiles. A straight line for the plots indicates that normality is a reasonable assumption.

Ljung-Box test

Another critical test in determining if the short-term correlation has been modelled when dealing with time series regression is the Ljung-Box test (Ljung and Box, 1978). The Ljung-Box test is one of the portmanteau tests which assesses whether a collection of autocorrelations are different to zero. The hypothesis when the test is used for an ARMA model is defined by:

H_0: Data are independently distributed, ie the residuals of the model have no autocorrelation;

H_1: Data are not independently distributed, ie the residuals of the model have autocorrelation.

The Ljung-Box test statistic is as follows in Equation (2.18), where n is the sample size $\widehat{\rho_{\tau}^{2}}$ is the sample autocorrelation at lag τ and the critical region for the rejection of the null hypothesis is $\chi^2_{1-\alpha,T}$, where α is typically 0.05 and T is the degrees of freedom.

$$Q = T(T+2)\sum_{\tau=1}^{T}\frac{\widehat{\rho_{\tau}^{2}}}{n-\tau} \tag{2.18}$$

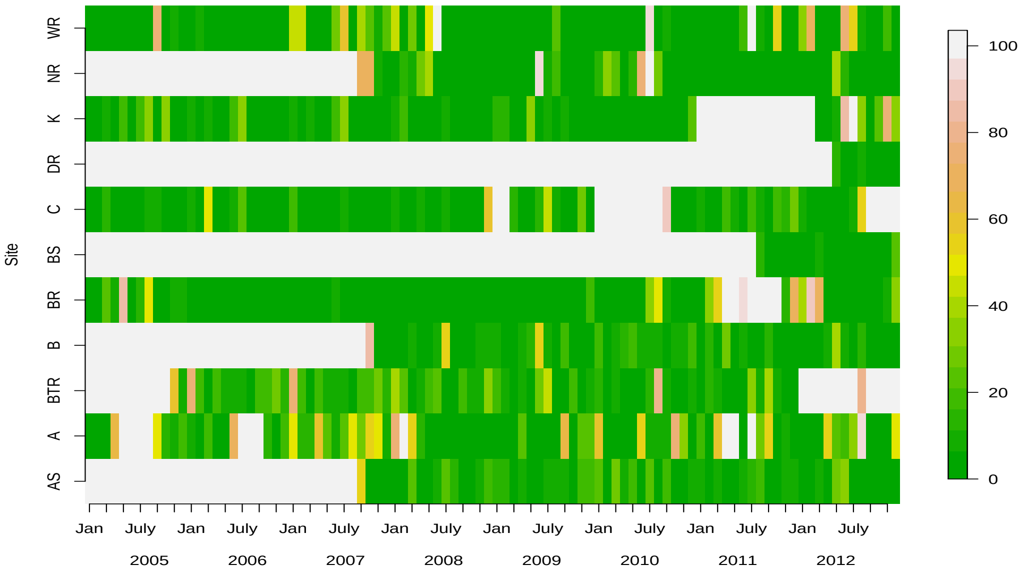
2.2 Site-by-Site Exploratory Data Analysis

This section discusses the different features, trends and patterns of the PM_{10} monitoring site data across the three years. This will notify any features which may pose a problem when summarising and modelling the data and in turn when attempting to produce an air pollution indicator. Firstly, one of the most striking features of the PM_{10} site data is the huge amount of missing data in a number of the sites.

2.2.1 Missing Data

It is common in environmental data to have omitted data and periods of missing values. As the data we are using have come from automatic monitoring sites there are many reasons for missing data including instrument malfunctions, incorrect calibrations, communication failure across the network monitoring system, and in some cases, instances where stations are yet to begin operating or had become disused. A large amount of missing data over a period of time can be problematic: it can reduce the representativeness of the data and therefore distort inferences. Figure (2.1) gives us a clear picture of the monthly percentage of missing data for each site and across eight years, from 2005 to 2012.******* The large white spaces show the sites where there was 100% missing data for that period. With the periods of 100% missing data that span at least one year the issue could be that the station was not yet functional or that it had been closed down. Apart from

Figure 2.1: Image plot for the percentage of missing data in each site for each year 2005 - 2012. The right hand axis indicates the percentage of missing data with 100% coloured white and 0% coloured dark green.



the large spells of completely missing data there are month long spells which appear to be randomly scattered across the months and sites. For the majority of the months shown on the graph, missing data values lay between 0 and 40% (shown in green). For the purpose of exploring the PM_{10} distribution across the city the missing values need not be imputed or interpolated. Each modelling technique has a different way of dealing with missing values, each of which are outlined in the methods which are described earlier in this chapter.

2.2.2 Graphical and Numerical Summaries of PM_{10} Monitoring Site Data

Tables 2.1a, 2.1b and 2.1c display the percentage of missing data and summary statistics for each site at each year. At first glance there is a vast difference in the percentage of missing data site to site with the smallest amount being no missing data and the largest with 100% of the data found to be missing. There is a wide spread of missing values across the sites for each year and this difference will have to be kept in mind throughout the rest of the time series modelling process. The largest mean PM_{10} value for 2010 and 2012 is at the Kerbside site with the value around $28.5\mu g/m^3$ and $23.7\mu g/m^3$ respectively whereas for 2011 the Kerbside site is disused and so the largest mean value is found at Byres Road and is around $23.7\mu g/m^3$. The discrepancies in available site data for each year make it hard to compare sites and so this has to be kept in mind throughout the modelling process and inference. The minimum values are mostly below $10\mu g/m^3$ whereas the maximum values are subject to a much wider spread - this could be due to the rise in PM_{10} around the 5th November which is discussed later in this Chapter.

A boxplot of the PM_{10} values is used as another summary method. The boxplots for the three years in Figure 2.2a, 2.2b and 2.2c show a similar dispersion of positively skewed values across the sites but with a large number of outliers at many of the sites. The outliers suggest that there could be a non-constant variance issue. Kerbisde and Byres Road have consistently high median values with Nithsdale road increasing in 2012 while Waulkmillglen Reservoir has consistently one of the lowest median values. It is unsurprising that Waulkmillglen Reservoir has consistently lower median values, as the site is the only one in a rural location.

To gain an initial impression of how the PM_{10} data are dispersed over time the data were plotted against time to give an insight into the overall trend of the data and to gain a subjective comparison between each of the sites and across the years. There appears also to be a non-constant variance issue for the time series for each site, with most of the values clustered at low PM_{10} levels. Each of the sites have one or two days around day 310 which are subject to a steep increase in PM_{10} levels, this could be due to Bonfire Night on the 309th day. The smoke that is produced by bonfires contain vast amounts

Table 2.1: Summary Statistics for PM_{10} at Each Site.

(a) 2010

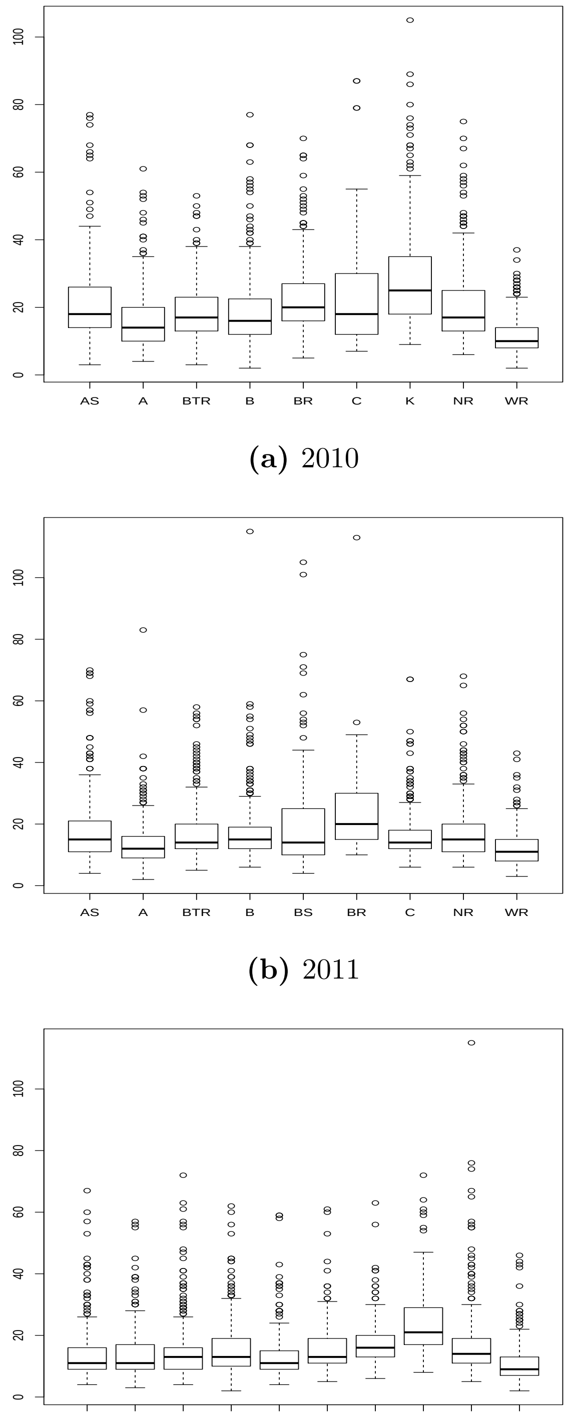
	Min	Q1	Median	Mean	Q3	Max	St.Dev	%NA
Abercrombie Street	3.00	14.00	18.00	21.35	26.00	77.00	11.66	11.51
Anderston	4.00	10.00	14.00	16.47	20.00	61.00	9.45	21.92
Battlefield Road	3.00	13.00	17.00	18.73	23.00	53.00	8.39	11.23
Broomhill	2.00	12.00	16.00	18.88	22.50	77.00	11.16	9.31
Byres Road	5.00	16.00	20.00	22.99	27.00	70.00	10.17	7.95
Burgher Street	-	-	-	-	-	-	-	100
Centre	7.00	12.00	18.00	23.22	30.00	87.00	17.15	73.97
Dumbarton Road	-	-	-	-	-	-	-	100
Kerbside	9.00	18.00	25.00	28.45	35.00	105.00	14.75	2.74
Nithsdale Road	6.00	13.00	17.00	21.42	25.00	75.00	12.30	24.11
Waulkmillglen Reservoir	2.00	8.00	10.00	11.80	14.00	37.00	5.83	8.49

(b) 2011

	Min	Q1	Median	Mean	Q3	Max	St.Dev	%NA
Abercrombie Street	4.00	11.00	15.00	18.15	21.00	70.00	11.17	6.03
Anderston	2.00	9.00	12.00	14.06	16.00	83.00	7.78	40.27
Battlefield Road	5.00	12.00	14.00	17.38	20.00	58.00	9.35	9.59
Broomhill	6.00	12.00	15.00	17.57	19.00	115.00	10.34	5.48
Byres Road	10.00	15.00	20.00	23.70	30.00	113.0	13.79	72.33
Burgher Street	4.00	10.00	14.00	20.19	25.00	105.00	16.84	59.45
Centre	6.00	12.00	14.00	16.53	18.00	67.00	8.44	11.23
Dumbarton Road	-	-	-	-	-	-	-	100
Kerbside	-	-	-	-	-	-	-	100
Nithsdale Road	6.00	11.00	15.00	17.55	20.00	68.00	9.64	0
Waulkmillglen Reservoir	3.00	8.00	11.00	12.14	15.00	43.00	6.13	15.63

(c) 2012

	Min	Q1	Median	Mean	Q3	Max	St.Dev	%NA
Abercrombie Street	4.00	9.00	11.00	13.87	16.00	67.00	8.88	6.28
Anderston	3.00	9.00	11.00	14.24	17.00	57.00	8.50	23.22
Battlefield Road	-	-	-	-	-	-	-	100
Broomhill	4.00	9.25	13.00	15.05	16.00	72.00	9.52	5.46
Byres Road	4.00	9.00	11.00	13.40	15.00	59.00	7.80	19.40
Burgher Street	2.00	10.00	13.00	15.44	19.00	62.00	9.11	2.73
Centre	5.00	11.00	13.00	15.96	19.00	61.00	8.34	39.62
Dumbarton Road	6.00	13.00	16.00	17.68	20.00	63.00	7.30	35.25
Kerbside	8.00	17.00	21.00	23.92	29.00	72.00	11.08	45.90
Nithsdale Road	5.00	11.00	14.00	17.14	19.00	115.00	11.50	4.64
Waulkmillglen Reservoir	2.00	7.00	9.00	11.11	13.00	46.00	6.46	22.13



(a) 2010

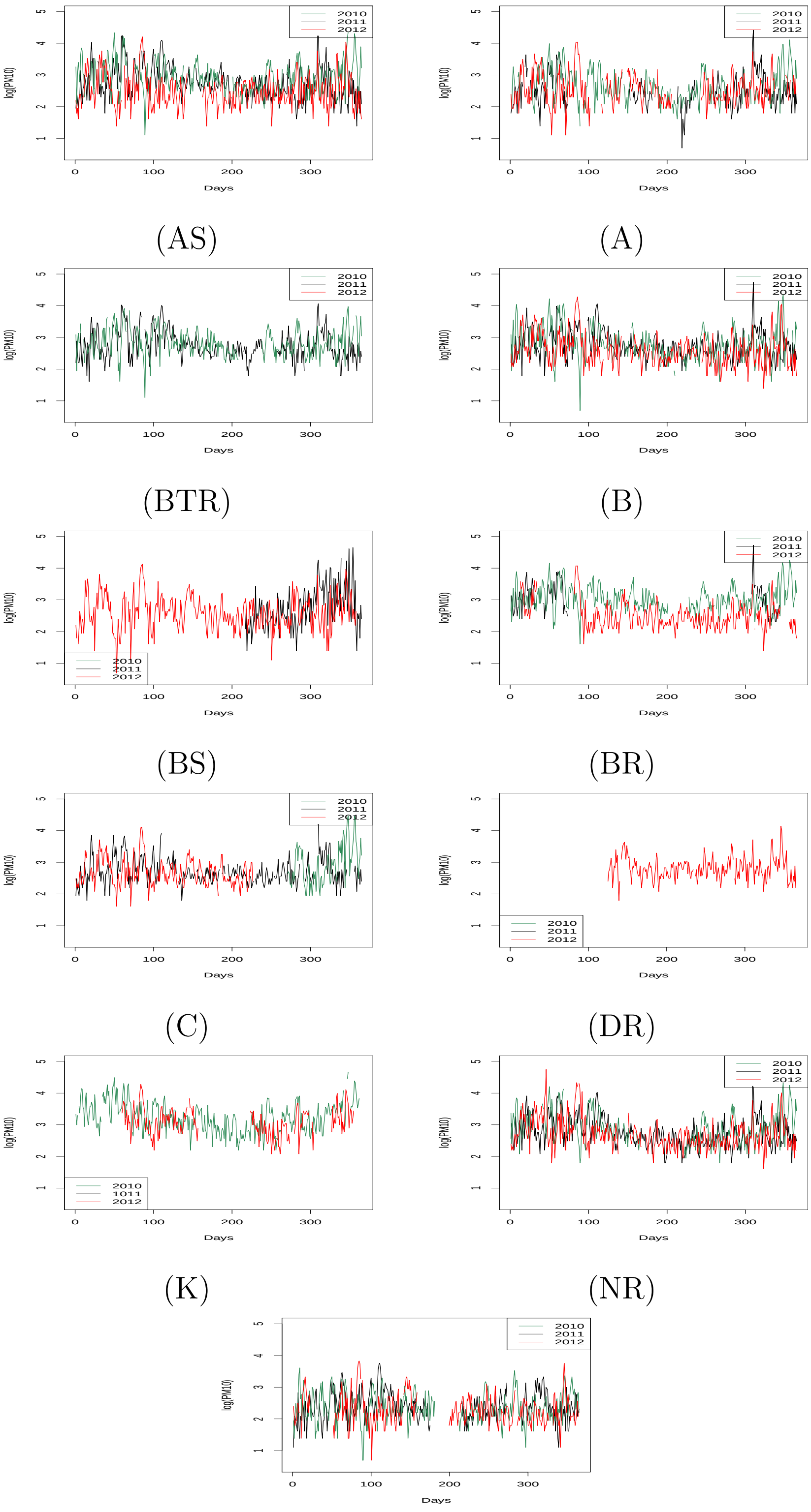
(b) 2011

(c) 2012

Figure 2.2: Boxplot of PM_{10} for Each Site.

of PM_{10} due to the combustion of fuels contains carbon, this results in higher than average levels. Firstly, the non-constant variance issue was addressed by applying different transformations to each site including log, exponential, square root and the Box-Cox transformation where $\lambda = \{-2, \ldots, 2\}$. The log transformation adequately addressed this issue by distributing the distribution in a fashion that resembles the normal distribution. After applying the logarithm transformation to each site in some cases the outliers from the time series plots were integrated into the main body of the distribution whereas for a few of the sites in each year this was not the case. The 310th value for several of the sites was removed to ensure that the modelling process was not compromised by these much higher than average values. Figure 2.4 shows the plots of log PM_{10} against time with the specified outliers removed at each year for each of the three years with the green line showing the spread of values for 2010, the black shows 2011 and the red shows 2012. This provides an obvious comparison between the years. The plot in Figure 2.4 shows the log PM_{10} concentrations across the 3 years. Overall, looking at both the time series plots, log PM_{10} would appear to follow a wave like seasonality with the peaks and dips of each site differing slightly. In the plots with large white space, however, this seasonality is not as apparent due to the huge amounts of missing data. These wave-like sinusoidal seasonality could be due to weekly or daily variations in log PM_{10} levels or it could be linked to a covariate effect. The plots in Figure 2.4 show that the average log PM_{10} levels seem to decrease with time.

To gain another perspective on the relationship between sites - a pairs plot for each year was produced. Figure 2.5a displays the pairs plot for year 2011 as an example as each years' pairs plots are similar. The plot shows that each of the pairs of sites have positively correlated log PM_{10} concentration levels. The correlation between each of the sites is also expressed in Table 2.2 as a numerical value. Some of the monitoring sites' data contains missing values and therefore the correlation coefficient cannot provide an accurate measure



(AS) (A)

(BTR) (B)

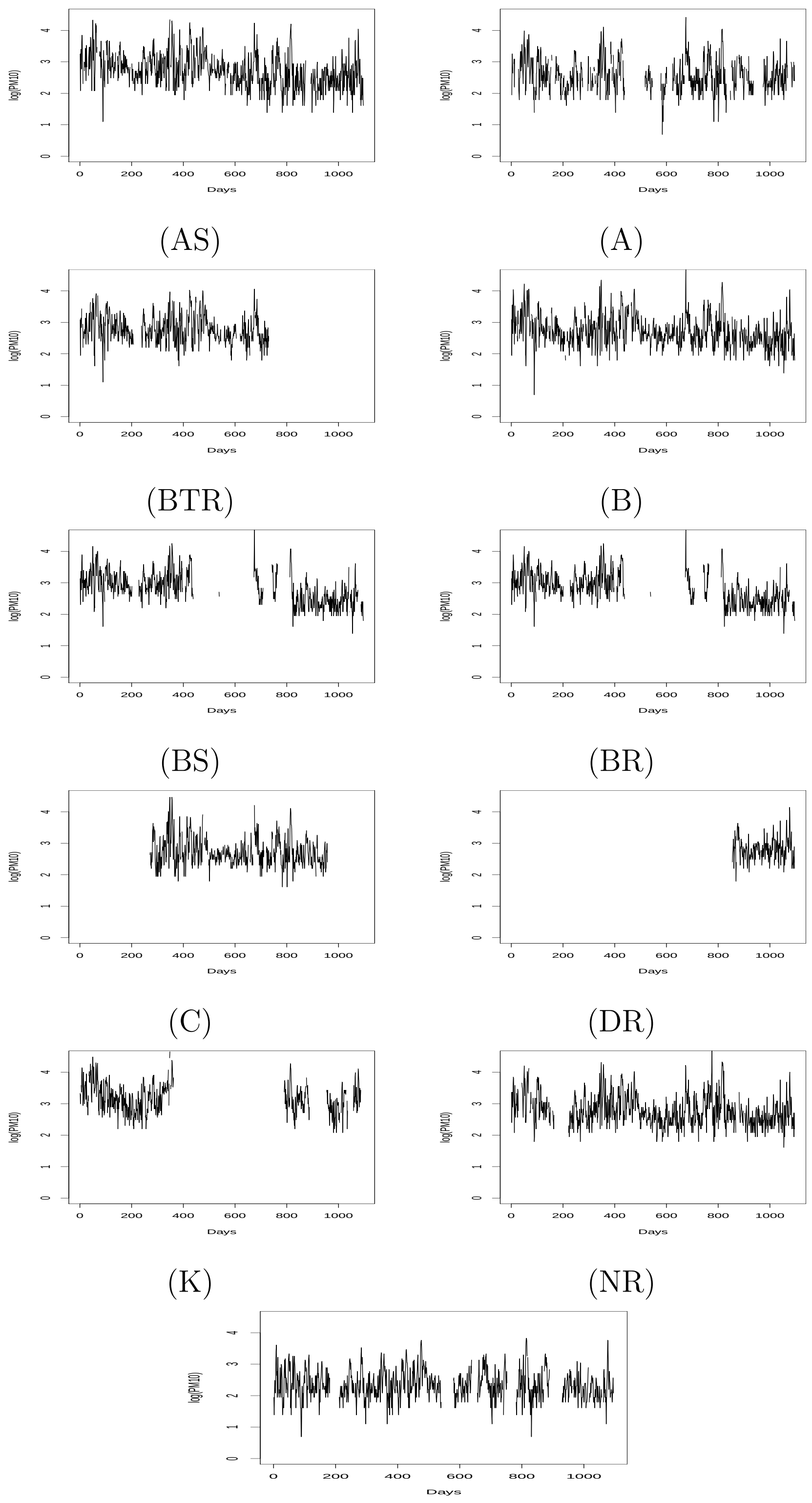
(BS) (BR)

(C) (DR)

(K) (NR)

(WR)

Figure 2.3: Time series plot of log(PM_{10}) for each site location for all three years on the same axis.



(AS) (A)

(BTR) (B)

(BS) (BR)

(C) (DR)

(K) (NR)

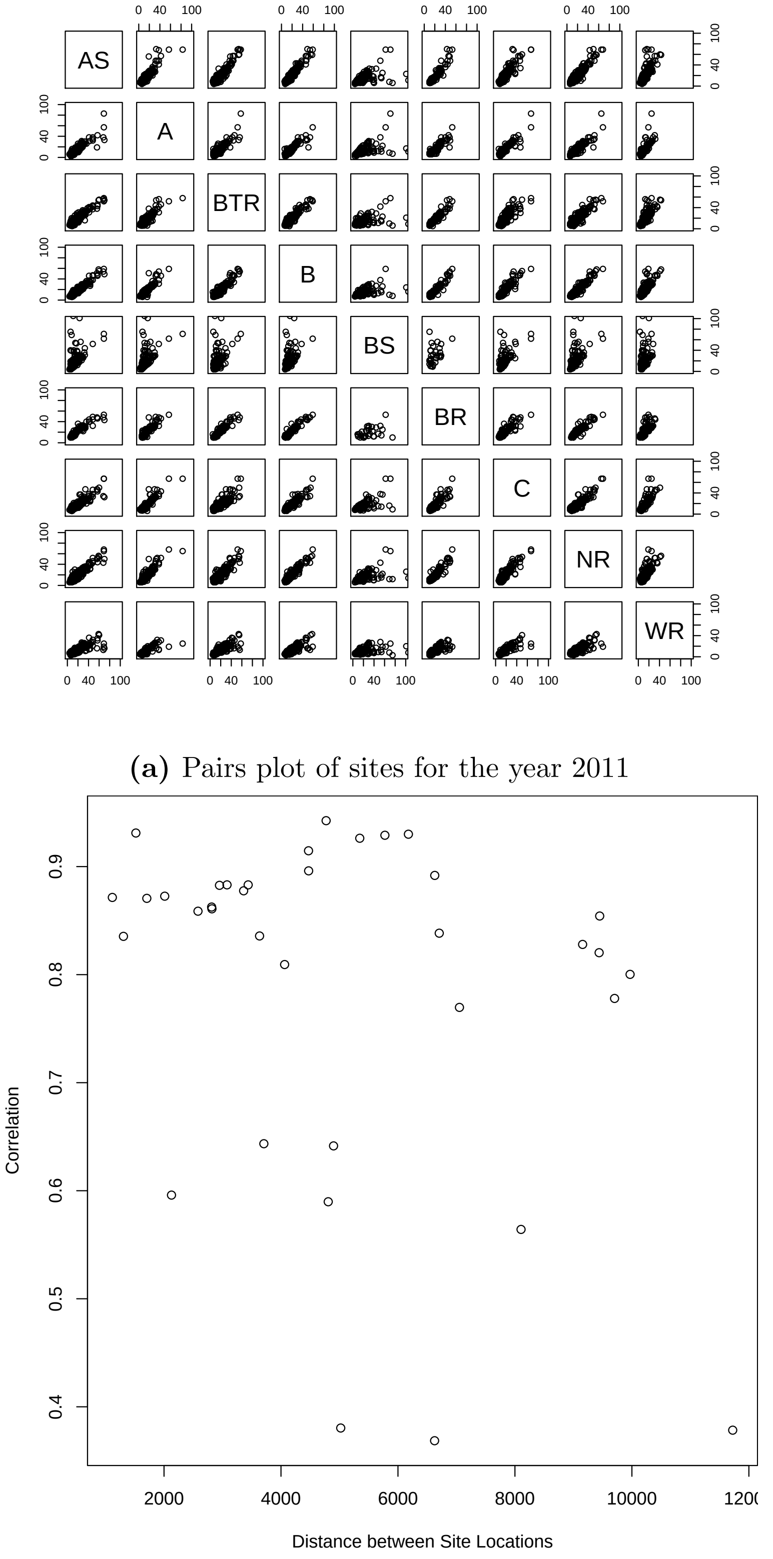
(WR)

Figure 2.4: Time series plot of $\log(\mathrm{PM}_{10})$ for each site location for all there years.

of correlation, however, it can be used as a rough guide. The correlation coefficients range from 0.37 to 0.94 which, further to the pairs plot, suggests that there could be a similar pattern for PM_{10} across the monitoring site locations. Both the plots and the correlation coefficients provide a argument that modelling the sites with the same model is reasonable. In addition to the correlation values the plot in Figure 2.5b gives an idea of the relationship between the level of correlation of log PM_{10} between two monitoring sites and the distance separating the two sites. Typically, it would be assumed that the larger the distance between sites the smaller the correlation would be which is true for most of the cases, however there are seven pairs of sites which have relatively small distances between the sites but relatively small correlation values. At closer inspection it appears that each of these seven pairs are common to the Burgher Street site and so the site pairs are not as strongly spatially correlated as the others. The Burgher Street site is relatively central and is positioned close to an A road, therefore it would be expected that the correlations would be high. The reason for this is uncertain but it could be due to a number of things including a large amount of missing values.

Table 2.2: Table of Correlations between Monitoring Cites, 2011

	AS	A	BTR	B	BS	BR	C	NR	WR
Abercrombie Street	-	0.86	0.88	0.93	0.60	0.94	0.87	0.88	0.80
Anderston	-	-	0.80	0.88	0.59	0.87	0.87	0.86	0.83
Battlefield Road	-	-	-	0.89	0.38	0.93	0.83	0.84	0.84
Broomhill	-	-	-	-	0.56	0.93	0.90	0.92	0.82
Burgher Street	-	-	-	-	-	0.37	0.64	0.64	0.38
Byres Road	-	-	-	-	-	-	0.88	0.91	0.78
Centre	-	-	-	-	-	-	-	0.86	0.85
Nithsdale Road	-	-	-	-	-	-	-	-	0.77
Waulkmillglen Reservoir	-	-	-	-	-	-	-	-	-



(a) Pairs plot of sites for the year 2011

(b) Plot of correlation between sites against the distance between the sites for the year 2011

Figure 2.5: Correlation Plots

Graphical and Numerical Summaries of Meteorological Data

The meteorological data as described in the data description Section 1.5 consist of daily mean values of temperature and humidity at one site in Glasgow. Tables 2.3a, 2.3b and 2.3c summarise each of the potential covariates with a number of summary statistics. The median and mean values for temperature were around 8°C in 2010, they then increased to around 10°C and 9°C respectively in 2011 and then dropped in 2012 to around 8°C. Whereas, the value for the standard deviation is at the highest in 2010 at 6.5 where it then decreases to around 4.5 in 2011 and 2012. This suggests that the temperature in Glasgow is on average at its highest in 2011 and at its most variable in 2010. The average percentage of humidity in Glasgow increased slightly with time. The median and mean went from around 82% and 82.5% in 2010 to 85% and 84.5%, respectively, in 2012. The standard deviation peaked at 8.9 in 2010 and dropped to around 7 in 2010 and 2012. Temperature and humidity are relatively consistent over the three years with slight changes.

The plots of temperature and relative humidity over time in Figures 2.6a and 2.6b explore the individual trends that each of the covariates possess. Both of the meteorological variables appear to follow a strong yearly sinusoidal cycle with temperature peaking during the summer months and dipping in the winter months. Humidity was slightly more variable, however, overall peaked in the winter months and dipped in the summer months. This mirrored effect, when temperature is high, humidity is low, and vice versa, suggests that there could be a strong negative correlation between these variables. The plot in Figure 2.6 further suggests that there could be a negative linear or quadratic relationship between temperature and humidity however there appears to be a large amount of variation. The Pearson's product-moment correlation coefficient was calculated to formally assess the correlation between temperature and humidity. The correlation coefficient was found to be -0.296 which suggests that there is no collinearity issue between

Table 2.3: Summary Statistics for Temperature and Humidity

(a) 2010

	Min	Q1	Median	Mean	Q3	Max	St.Dev
Temperature	-11.00	3.00	8.00	7.50	13.00	18.00	6.50
Humidity	49.00	76.00	82.00	81.55	88.00	100.00	8.90

(b) 2011

	Min	Q1	Median	Mean	Q3	Max	St.Dev
Temperature	-4.00	6.00	10.00	9.14	13.00	18.00	4.46
Humidity	57.00	78.00	84.00	82.88	88.00	98.00	7.11

(c) 2012

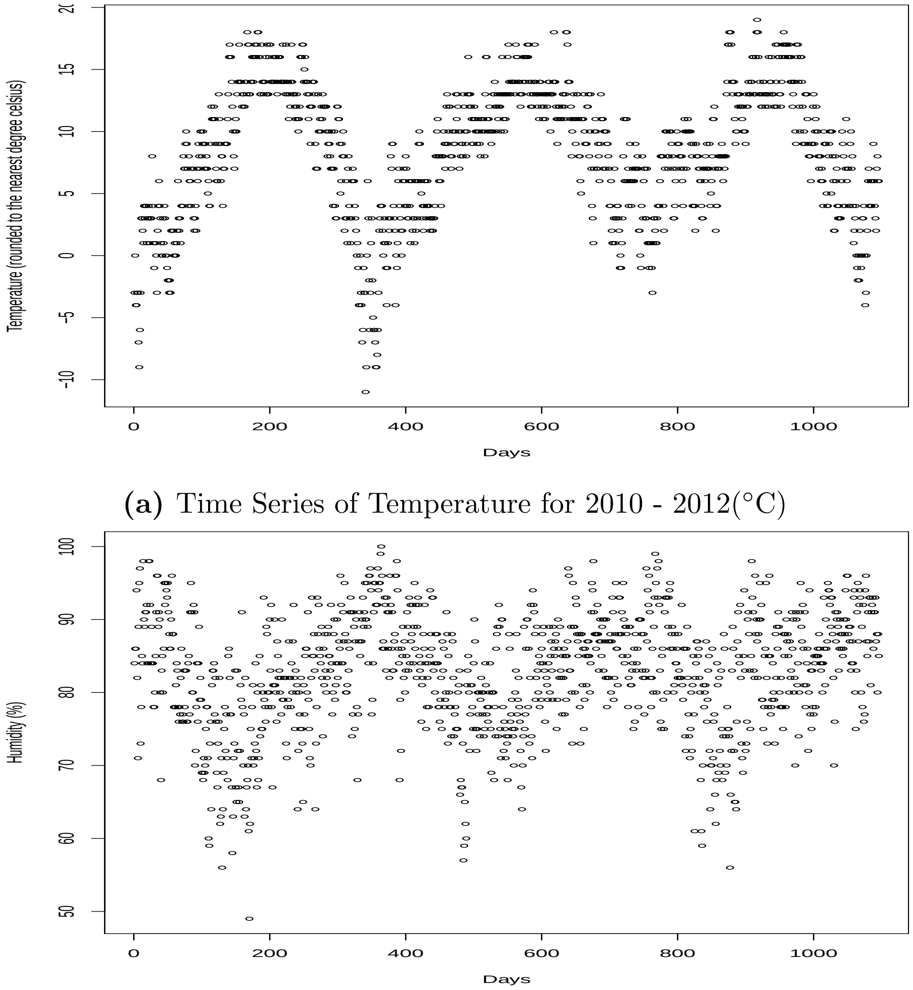
	Min	Q1	Median	Mean	Q3	Max	St.Dev
Temperature	-4.00	6.00	8.00	8.52	12.00	19.00	4.75
Humidity	56.00	79.00	85.00	83.58	89.00	99.00	7.70

temperature and humidity.

The exploratory analysis thus far has used mostly informal methods to assess the relationship between PM_{10} across spatial and temporal domains. The meteorological potential variables, temperature and humidity, have been explored across time and a potential collinearity issue has been discussed. The next chapter quantifies more accurately if and to what extent PM_{10} is related to temperature and humidity using more appropriate regression assumptions.

2.3 Exploring Trends and Seasonality using Linear Regression Modelling

In this section we use a linear regression model to determine the relationship between PM_{10} and the meteorological variables across time. This



(a) Time Series of Temperature for 2010 - 2012(°C)

(b) Time Series of Humidity for 2010 - 2012(%)

Figure 2.6: Time Series Plot of Temperature and Humidity

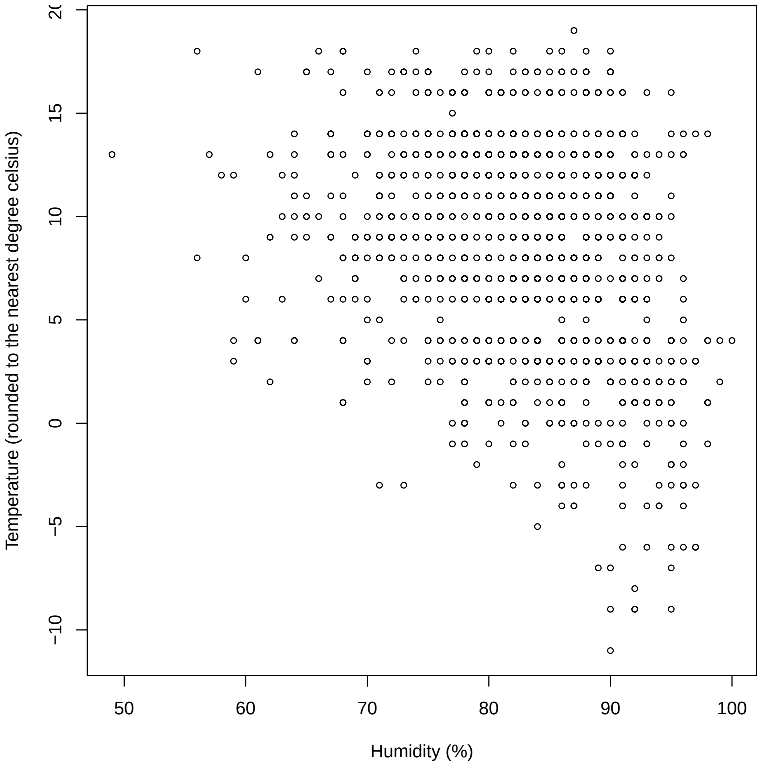


Figure 2.7: Temperature (rounded to the nearest °C) against Humidity (%)

technique is demonstrated using only the available daily mean PM_{10} values for 3 out of the 11 sites for 2011. These sites were chosen to represent different levels of missing data: one site with no missing values, one with a large proportion of missing values (72%) and one site with a medium amount of missing values (40%). This provides an overview of the process from different sites without going into detail for each one.

In order to exploit linear regression modelling, a number of assumptions (outlined in section 2.1.1) had to be relaxed in this case. The main assumption that is breached is the assumption of independent errors. Air pollution time series data, in general, are correlated from one day to the next and the linear model we use first does not account for this covariance. Therefore, a linear regression model is not an accurate method to model the distribution of PM_{10} in this case, however, it can act as an exploratory method to provide a good idea of possible variables to include in our model. The uncorrelated errors linear regression models are fit using OLS and it is assumed that the errors have mean zero.

The previous exploratory analysis suggested a sinusoidal pattern across the year with a possible weekly effect. To model these possible sinusoidal patterns, a regression type known as harmonic regression was used. A harmonic function included regression terms for the pattern over the year (DOY) while the day of the week was modelled as a factor (DOW). These terms were coupled with the meteorological variables and different combinations were fit to gain an idea of what variables were related to PM_{10} and if this differed across each site and across each year. Three models in total were fit and are described in Table (2.4). The model equation are then explained as follows:

$$\begin{aligned} y_t = \beta_0 + \beta_1 \cos(2\pi t/365) + \beta_2 \sin(2\pi t/365) + \beta_3 (\mathrm{DayofWeek})_t \\ + \beta_4 (\mathrm{Humidity})_t + \beta_5 (\mathrm{Temperature})_t + \varepsilon_t, \end{aligned} \tag{2.19}$$

$$\begin{aligned} y_t = \beta_0 + \beta_1 \cos(2\pi t/365) + \beta_2 \sin(2\pi t/365) + \beta_3 (\mathrm{DayofWeek})_t \\ + \beta_4 (\mathrm{Humidity})_t + \varepsilon_t, \end{aligned} \tag{2.20}$$

$$y_t = \beta_0 + \beta_1 \cos(2\pi t/365) + \beta_2 \sin(2\pi t/365) + \beta_3 (\mathrm{DayofWeek})_t + \varepsilon_t. \tag{2.21}$$

In the above models $t = 1, \ldots, 365$.

Starting with Model 1, the plots in Figure 2.8a, 2.8b and 2.8c display the log PM_{10} values with the fitted mean line for Model 1 with pointwise 95% confidence bands for these estimated means. Looking at the plots in Figure 2.8a, 2.8b and 2.8c the model appears to fit the overall trend of the data well, however the large amount of variability is not accounted for. A number

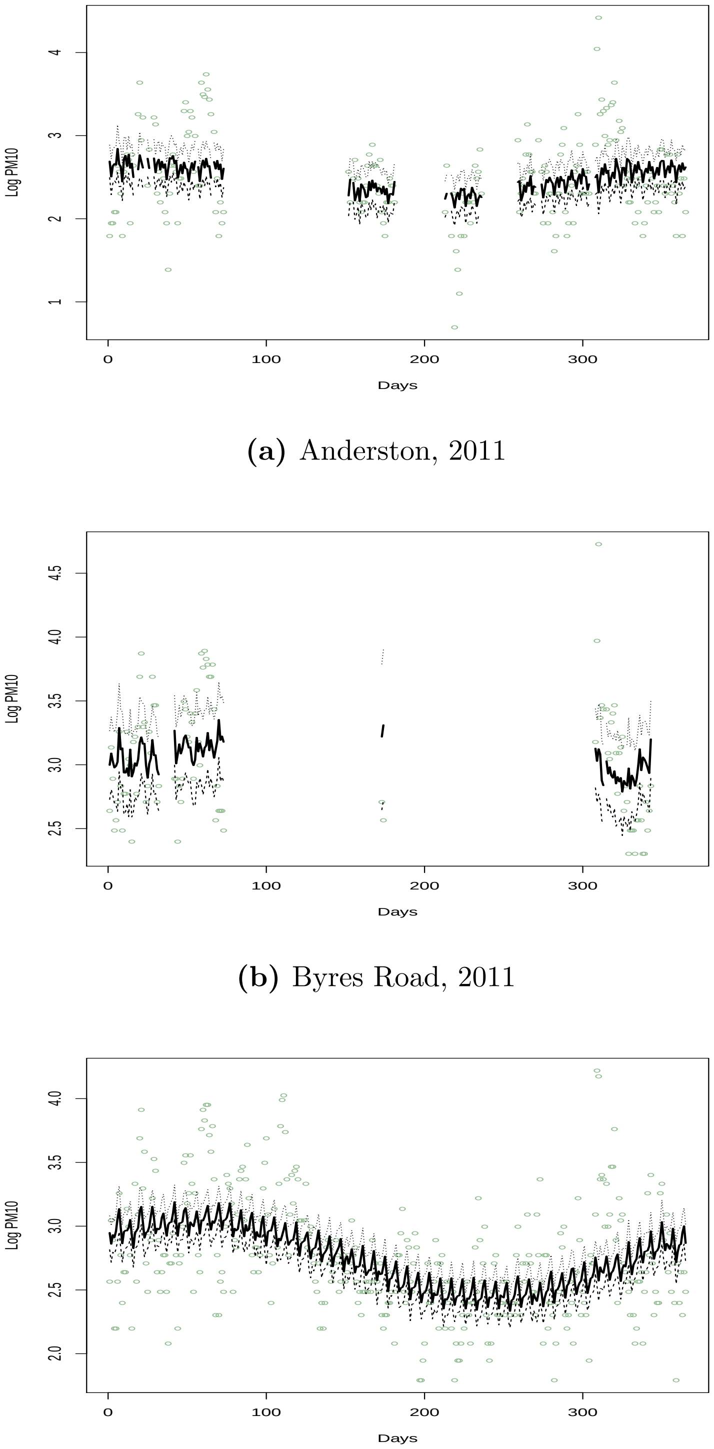
Table 2.4: Description of the three yearly models

Model Number	Model Description
Model 1	DOY, DOW, Humid & Temp
Model 2	DOY, DOW & Humid
Model 3	DOY & DOW

of linear regression model assumptions are breached by not accounting for the temporal correlation, as explained in Section 2.1.1, therefore the standard errors and p-values are not reliable. The intercept terms for each of the models are around 4.9 which demonstrates some similarity between the sites. The DOY term at each of the sites would be significant if the errors were uncorrelated and the DOW factor would be significant at the Nithsdale Road site. However, neither the temperature or humidity terms appear to be significant in the event that the errors were uncorrelated. This is surprising considering that from the graphical summaries it looked like the meteorological variables would be important in the modelling of the daily PM_{10} values. Looking at the estimates and p-values (the p-values are not completely reliable without accounting for the autocorrelated errors) the decision was made to drop temperature from the model.

Model 2 contains the same covariates as in Model 1 minus the temperature variable. The next set of plots in Figure 2.9a, 2.9b and 2.9c show the fitted line and respective confidence bands for Model 2. Compared to the plots of the fitted line for Model 1, the fitted line for Model 2 appears to be very similar. The intercept estimates are, again, similar - between 2.7 and 2.9. Assuming that the errors are uncorrelated the DOY harmonic regression terms are significant for each of the sites, DOW is significant for the Nithsdale road site but the humidity variable is not significant. Therefore the next model to explore has both of the meteorological variables removed.

The last model, Model 3, contains variables DOY and the DOW factor. The plots with the fitted line and confidence bands in Figure 2.9a, 2.9b

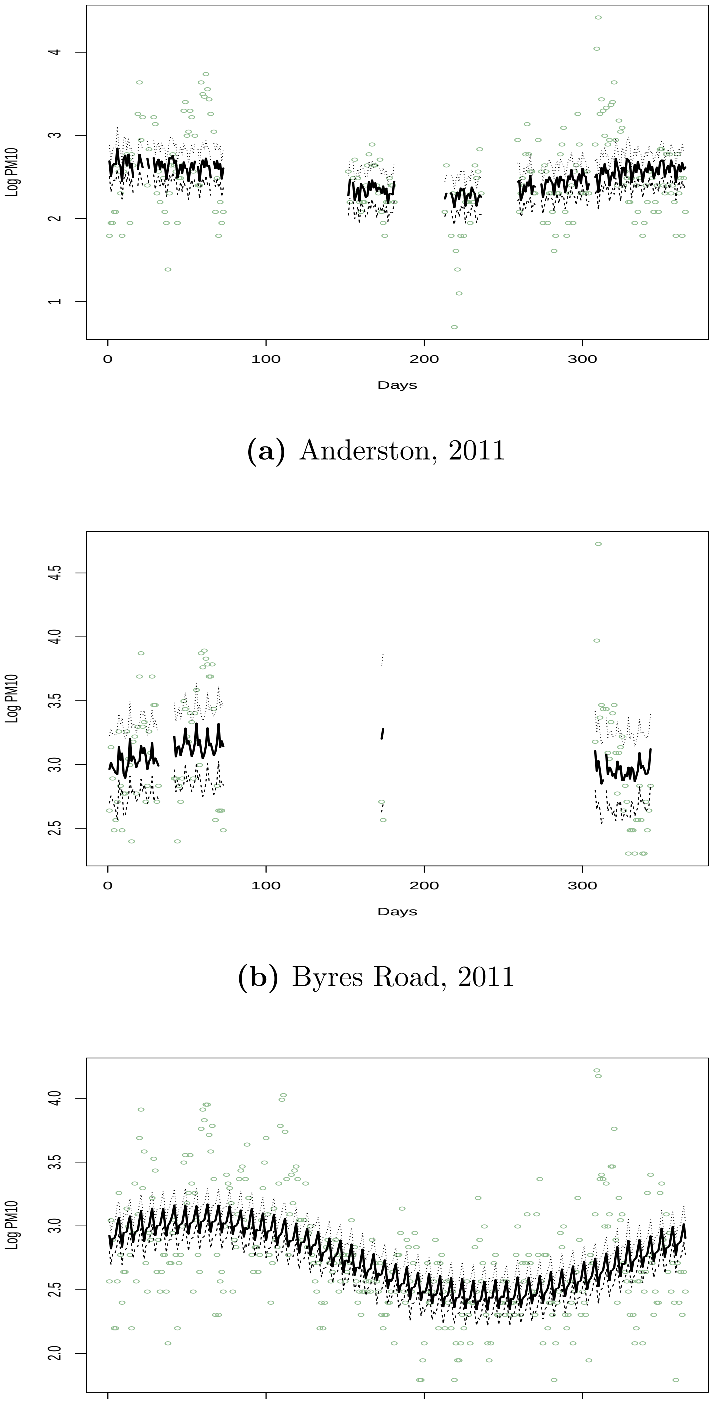


(a) Anderston, 2011

(b) Byres Road, 2011

(c) Nithsdale Road, 2011

Figure 2.8: Logged PM_{10} Values with fitted line plot for Model 1



(a) Anderston, 2011

(b) Byres Road, 2011

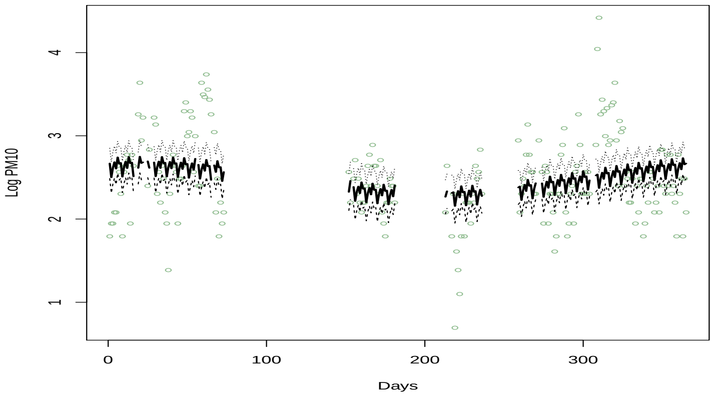
(c) Nithsdale Road, 2011

Figure 2.9: Logged PM_{10} Values with fitted line plot for Model 2

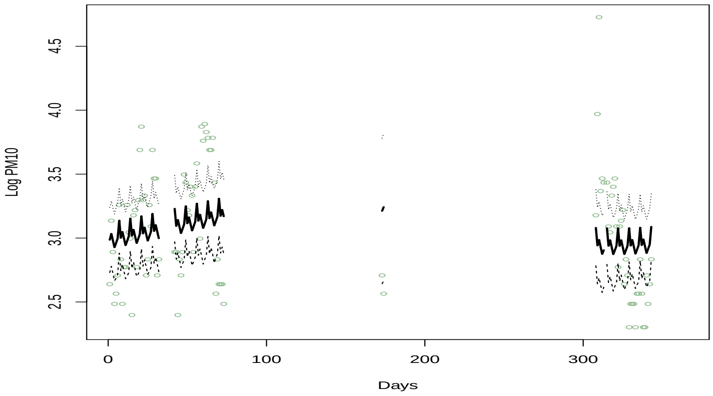
and 2.9c show that the fitted line does not account for the variability in the data. The overall trend is much flatter than the overall trend in Model 2. These plots alone suggest that the meteorological variable accounts for the variability. Without humidity included in the model the model assumes that PM_{10} depends only on time. Without humidity in the model the difference between the intercept estimates for each of the sites grows with the Anderson site intercept estimated to be around 2.4 and the Byres Road site intercept estimated to be around 3.1. Suggesting that humidity could have been accounting for the differences between the sites.

2.3.1 Exploratory Conclusions

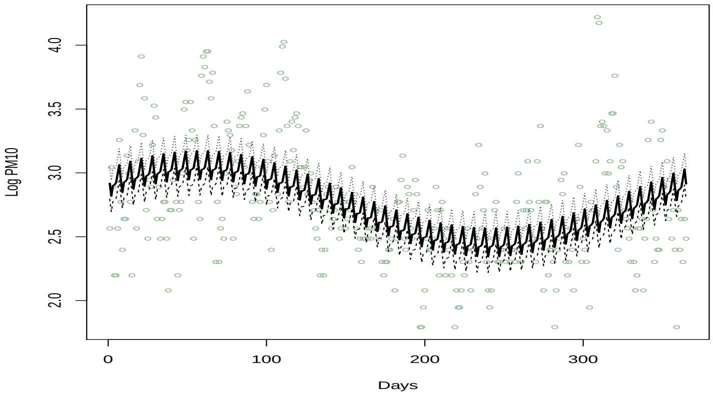
The exploratory section of this chapter introduced the attributes and complications of the PM_{10} and meteorological data. Characteristics such as missing data were discussed and unequal variance and outliers were dealt with using data transformations and outlier removals. The aim of this chapter was to, firstly, explore the PM_{10} monitoring site data characteristics such as missing data, unequal variance, and outliers and then to move onto explore the PM_{10} data against time the relationship with the meteorological variables. Possible seasonalities and trends were then explored across each year separately and combined using simple descriptive statistics and graphical summaries. The distribution of temperature and humidity was also examined and possible collinearity issues considered. A more formal method of data exploration was employed to identify possible model variables for each site and year - linear regression modelling. This concluded that different combinations of DOY, DOW, temperature and humidity could possibly model the log PM_{10} values for one year at each site. This analysis, however, did not account for the time series nature of the data and the likely autocorrelation in the values across time. Therefore a more formal modelling approach is employed in the next section to gain a more definite idea of what variables play a part in modelling PM_{10} in Glasgow to ultimately aid in the



(a) Anderston, 2011



(b) Byres Road, 2011



(c) Nithsdale Road, 2011

Figure 2.10: Logged PM_{10} Values with fitted line plot for Model 3

construction of an air pollution indicator.

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

Moving onto a formal analysis of logged PM_{10} across the monitoring sites in Glasgow, it becomes imperative that the covariance structure accounts for the time series nature of the data. The aim of the modelling process is to model each of the sites across the years with the same model in order to gain an understanding of how PM_{10} is distributed over time across the sites. The three models fit to the time series PM_{10} data at each location for each year were a starting point in the modelling process, however linear regression with uncorrelated errors is not a suitable method to model time series data due to the temporal correlation. This temporal correlation which rendered the linear regression standard errors incorrect will have been left over in the residuals of the previous models as it was not accounted for. Therefore, in order to account for this autocorrelation we must first assess the correlation which was left over in the residuals for Models 1, 2 and 3 from the previous section. Figures 2.11a, 2.11b and 2.11c display the acf and pacf plots for Model 2 at the same three locations as earlier - the plots were very similar for each of the models therefore only the plots for one model are displayed for explanation. Each of the black lines which breach the confidence bands represents that there is significant autocorrelation left in the residuals at that lag. An acf plot with a breaching line at lags 1 and 2 (if there are no significant lags in the pacf) would suggest that an MA(2) process could account for the autocorrelation. A pacf plot with a breaching line at lag 1 (if there are no significant lags in the acf) would suggest that an AR(1) process would account for the autocorrelation. Realistically, this is not always this simple and both process have to work together forming an ARMA(p,q) process. In this case it is not clear which process would adequately model the correlation. Looking at the plots it would suggest that an AR(1) or an ARMA(1,1) process could work.

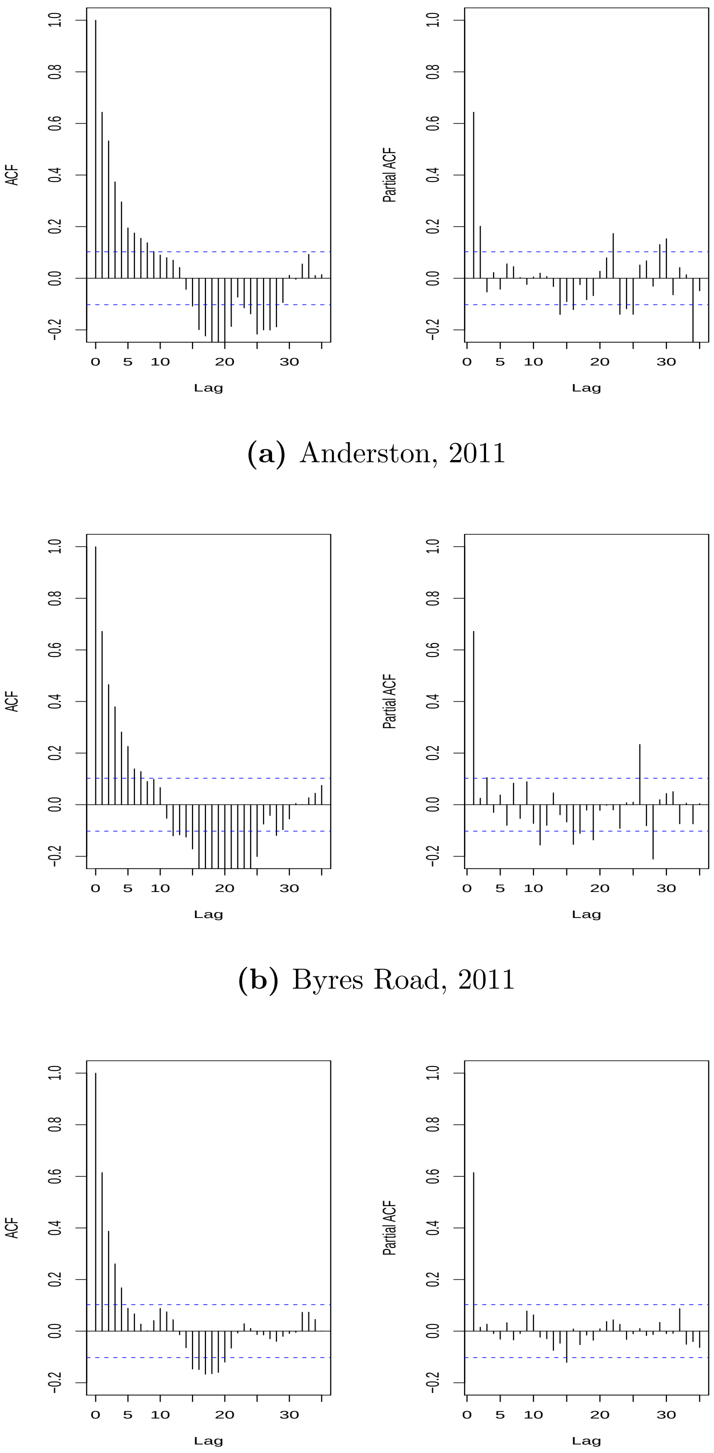
Both the AR(1) and ARMA(1,1) should both be tested in the model to assess which one, if either, account for autocorrelation. Stationarity was also assessed using the residuals of the linear regression model, it was found that there was no systematic change in mean or variance, therefore the data could be assumed to be stationary and a differencing technique would not need to be applied to the data to try to make them stationary.

Table 2.5: Description of the three yearly models

Model Number	Model Description
Model 4	DOY, DOW, Humid & Temp
Model 5	DOY, DOW & Humid
Model 6	DOY & DOW

The three models were fit to each of the sites with similar equations to those in the previous chapter. The same combination of regression terms DOY and DOW along with the meteorological variables temperature and humidity are included in the models named Model 4, 5 and 6 as summarised in Table (2.5). The models are outlined in the previous section in Equations (2.19), (2.20) and (2.21) where ε_t is now a mean zero time series process. Both the AR(1) and the ARMA(1,1) processes were tested in the model to account for the autocorrelation at the different locations. In the interest of consistency we strived to fit the same process to each of the locations. The ARMA(1,1) overfit the model whereas the AR(1) process in each of the locations consistently removed the autocorrelation. The temporal correlation accounted for by the AR(1) process where $\varepsilon_t = \phi\varepsilon_{t-1} + Z_t$ and the residuals Z_t can be assumed to follow a gaussian distribution where $Z_t \sim N(0, \sigma^2)$. In each of the sites, after including the AR(1) process in the model, the acf and pacf plots showed no outstanding temporal autocorrelation.

In order to compare the models adequately, each monitoring site has been



(a) Anderston, 2011

(b) Byres Road, 2011

(c) Nithsdale Road, 2011

Figure 2.11: ACF and Partial ACF Plots

taken in turn to display the estimates and standard errors for each of the models. Tables 2.6a, 2.6b and 2.6c display the estimates and standard errors for the Anderston monitoring site. Model 4 and 5 estimate the intercept term to equal around 2.8 whereas Model 6 estimates the intercept to be around 2.4. Model 4 and 5 are different with regards to significant variables. In Model 4 DOY and the AR (1) term are both significant, whereas the DOY and the DOW 7 and the AR(1) term are significant for both Model 5 and 6. The humidity term was not significant in Model 5, however, we know from the exploratory section that without humidity the volatility is not accounted for. From the analysis of this site alone it would appear that either Model 5 or 6 accounts for the trend, seasonality and volatility appropriately.

The Byres Road estimates and standard errors are displayed in Tables 2.7a , 2.7b and 2.7c for Model 4, 5 and 6 respectively. These models show a very different picture to that of the Anderston site. The intercept and the AR(1) terms are the only significant terms in all three of the models. The Byres road site has only 28% of the data available, however, which makes is less reliable when it comes to estimating an air pollution indicator.

The Nithsdale Road model estimates and standard errors are displayed in Tables 2.8a, 2.8b and 2.8c. Altogether there are more significant variables in these models than in the other sites. The intercept estimates range from around 2.7 for Model 6 to around 3.4 for Model 4 and 5, again suggesting that Model 6 could be underestimating the true intercept term. The DOY, DOW and humidity terms are consistently significant in each of the models. The Nithsdale Road site has no missing values and therefore could be argued to be the most reliable monitoring site.

This inconsistency over each of the sites demonstrates the variation in PM_{10} in space. Each of the sites have different estimates and a slightly different profile of significant variables. Although for ease of interpretation the models for each of the sites are consistent and model PM_{10} as best as possible, each of the sites could as easily have a slightly different mix of

Table 2.6: Estimate and Standard Error for Anderston, 2011

(a) Model 4

	Estimate	Standard Error
Intercept (β_0)	**2.851**	**0.410**
DOY $\widehat{A}$	**0.150**	**0.093**
DOY $\widehat{\psi}$	0.267	0.728
DOW 2 (β_3)	0.044	0.075
DOW 3 (β_4)	-0.029	0.093
DOW 4 (β_5)	0.105	0.099
DOW 5 (β_6)	0.015	0.100
DOW 6 (β_7)	0.042	0.091
DOW 7 (β_8)	-0.137	0.073
Humidity (β_9)	-0.004	0.005
Temperature (β_10)	-0.006	0.012
AR(1) (ϕ)	**0.640**	**0.051**

(b) Model 5

	Estimate	Standard Error
Intercept (β_0)	**2.784**	**0.390**
DOY $\widehat{A}$	**0.183**	**0.087**
DOY $\widehat{\psi}$	0.323	0.599
DOW 2 (β_3)	0.044	0.075
DOW 3 (β_4)	-0.029	0.093
DOW 4 (β_5)	0.105	0.099
DOW 5 (β_6)	0.017	0.100
DOW 6 (β_7)	0.040	0.091
DOW 7 (β_8)	**-0.140**	**0.071**
Humidity (β_9)	-0.004	0.005
AR(1) (ϕ)	**0.640**	**0.051**

(c) Model 6

	Estimate	Standard Error
Intercept (β_0)	**2.464**	**0.093**
DOY $\widehat{A}$	**0.173**	**0.085**
DOY $\widehat{\psi}$	0.373	0.621
DOW 2 (β_3)	0.050	0.075
DOW 3 (β_4)	-0.033	0.093
DOW 4 (β_5)	0.108	0.100
DOW 5 (β_6)	0.017	0.100
DOW 6 (β_7)	0.032	0.091
DOW 7 (β_8)	**-0.141**	**0.073**
AR(1) (ϕ)	**0.635**	**0.051**

Table 2.7: Estimate and Standard Error for Byres Road, 2011

(a) Model 4

	Estimate	Standard Error
Intercept (β_0)	**3.000**	**0.620**
DOY $\widehat{A}$	0.053	0.141
DOY $\widehat{\psi}$	-0.755	0.825
DOW 2 (β_3)	-0.074	0.093
DOW 3 (β_4)	-0.012	0.118
DOW 4 (β_5)	0.058	0.128
DOW 5 (β_6)	0.131	0.124
DOW 6 (β_7)	-0.004	0.114
DOW 7 (β_8)	0.056	0.091
Humidity (β_9)	0.001	0.006
Temperature (β_10)	-0.023	0.014
AR(1) (ϕ)	**0.689**	**0.070**

(b) Model 5

	Estimate	Standard Error
Intercept (β_0)	**2.622**	**0.579**
DOY $\widehat{A}$	0.114	0.135
DOY $\widehat{\psi}$	0.997	0.842
DOW 2 (β_3)	-0.087	0.094
DOW 3 (β_4)	-0.022	0.120
DOW 4 (β_5)	0.050	0.130
DOW 5 (β_6)	0.127	0.126
DOW 6 (β_7)	-0.008	0.115
DOW 7 (β_8)	0.059	0.092
Humidity (β_9)	0.004	0.006
AR(1) (ϕ)	**0.680**	**0.071**

(c) Model 6

	Estimate	Standard Error
Intercept (β_0)	**2.933**	**0.199**
DOY $\widehat{A}$	0.115	0.139
DOY $\widehat{\psi}$	0.908	0.849
DOW 2 (β_3)	-0.092	0.094
DOW 3 (β_4)	-0.025	0.120
DOW 4 (β_5)	0.046	0.130
DOW 5 (β_6)	0.123	0.126
DOW 6 (β_7)	-0.005	0.115
DOW 7 (β_8)	0.051	0.091
AR(1) (ϕ)	**0.683**	**0.070**

Table 2.8: Estimate and Standard Error for Nithsdale Road, 2011

(a) Model 4

	Estimate	Standard Error
Intercept (β_0)	**3.495**	**0.250**
DOY $\widehat{A}$	**0.223**	**0.065**
DOY $\widehat{\psi}$	1.093	0.712
DOW 2 (β_3)	0.002	0.047
DOW 3 (β_4)	0.028	0.058
DOW 4 (β_5)	0.088	0.063
DOW 5 (β_6)	**0.153**	**0.063**
DOW 6 (β_7)	0.054	0.058
DOW 7 (β_8)	-0.065	0.047
Humidity (β_9)	**-0.007**	**0.003**
Temperature (β_10)	-0.017	0.009
AR(1) (ϕ)	**0.629**	**0.041**

(b) Model 5

	Estimate	Standard Error
Intercept (β_0)	**3.310**	**0.240**
DOY $\widehat{A}$	**0.307**	**0.058**
DOY $\widehat{\psi}$	**0.896**	**0.205**
DOW 2 (β_3)	0.002	0.047
DOW 3 (β_4)	0.025	0.059
DOW 4 (β_5)	0.086	0.063
DOW 5 (β_6)	**0.155**	**0.063**
DOW 6 (β_7)	0.046	0.059
DOW 7 (β_8)	-0.069	0.047
Humidity (β_9)	**-0.007**	**0.003**
AR(1) (ϕ)	**0.622**	**0.041**

(c) Model 6

	Estimate	Standard Error
Intercept (β_0)	**2.719**	**0.057**
DOY $\widehat{A}$	**0.299**	**0.060**
DOY $\widehat{\psi}$	**1.019**	**0.243**
DOW 2 (β_3)	0.004	0.048
DOW 3 (β_4)	0.016	0.059
DOW 4 (β_5)	0.082	0.064
DOW 5 (β_6)	**0.147**	**0.064**
DOW 6 (β_7)	0.031	0.059
DOW 7 (β_8)	-0.078	0.047
AR(1) (ϕ)	**0.618**	**0.041**

variables which model pm_{10} at each site specifically. This confirms that when exploring PM_{10} we must not only consider time but also space.

2.4.1 Model Selection

The estimates and standard errors have been displayed and discussed for Model 4, 5 and 6 for the three monitoring sites. The covariates that were significant in the model provide an idea of which combination of covariates model the log PM_{10} monitoring site data best. The AIC, which is a measure of the goodness of fit of the model, can provide a more formal model selection method. Table 2.9 displays the AIC values for each of the models for each of the sites.

There are important factors, specific to the Glasgow PM_{10} monitoring site data, to take into consideration. Firstly, temperature term was not significant for any of the sites and therefore despite the AIC values Model 4 should not be chosen as the most appropriate model. Then taking each site individually, it would appear that the best fitting model for Anderston and Byres Road is between Model 5 and 6 with the difference between the AIC values less than 2. Model 5 for the Nithsdale Road site, which could be argued to be the most reliable site, has the lowest AIC value. Therefore, the best fitting model for log PM_{10} across the sites in Glasgow would appear to be between Model 5 and Model 6. Looking back at the exploratory analysis, Model 6 did not model the variability as well as Model 5 suggesting that overall PM_{10} at each of the sites across Glasgow is modelled best using Model 5. To gain an idea of how this is distribute across the different 11 sites the final estimates and standard errors are discussed later in Section 2.5.

2.4.2 Model Diagnostics

As discussed in section 2.1, there are a number of assumptions and diagnostics checking which must be performed after the model has been fit to

Table 2.9: Summary of the three models and their corresponding AIC value at each site

	AIC		
	Anderston	Byres Rd	Nithsdale Road
Model 4	222.554	88.690	199.404
Model 5	220.833	89.126	202.930
Model 6	219.544	87.452	207.3042

the data. The assumption of homoscedasticity can be checked by looking for any trend or patterns left in the residuals, the assumption of normality can be checked using a Q-Q plot and the Ljung-Box test can be used to test if the data are independently distributed and if the short-term correlation has been accounted for by the AR(1) process.

Model Diagnostics for Model 5

The residual plots in Figure 2.12 provide an idea of the level of trend left in the residuals. In each of the plots the values appear to be equally spread around the zero line with no clear underlying leftover pattern or trend. The residual plots for Anderston and Byres Road are more difficult to interpret due to the large amount of missing values but the model appears to remove the trend and patterns in the data relatively well. The Q-Q plots in Figure 2.13 show that the residuals appear to be normally distributed as the points for each site run along the x=y line almost perfectly except for a slight deviation at the tails of the distribution. Deviation from the x=y line, however, is not unusual and despite this the normality of the residuals can be assumed. The Ljung-Box test tests the null hypothesis that the data are independently distributed, ie the residuals of the model have no autocorrelation. If the p-value is less than the 0.05 critical value then the null hypothesis of independent and identically distributed (iid) process can be rejected in favour of the alternative hypothesis, ie the data are not iid. The Ljung-Box p-value in Table 2.10 for each of the sites is more than 0.05 and therefore it can be assumed that there is no autocorrelation left in the residuals.

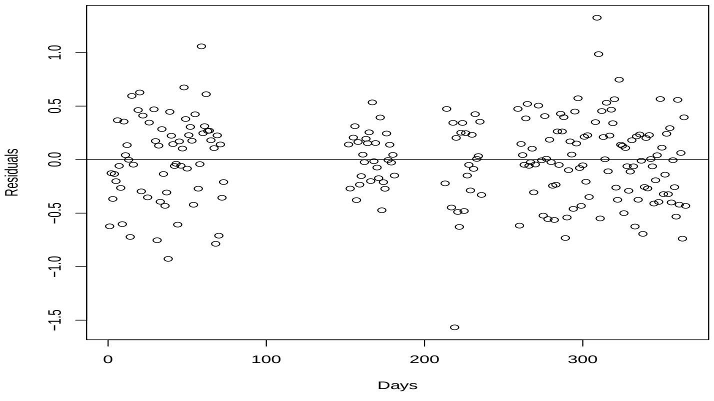
Table 2.10: The Ljung-Box P-Value for Each of the Three Sites

	Anderston	Byres Rd	Nithsdale Rd
Ljung-Box Q stat	0.079	0.722	0.883

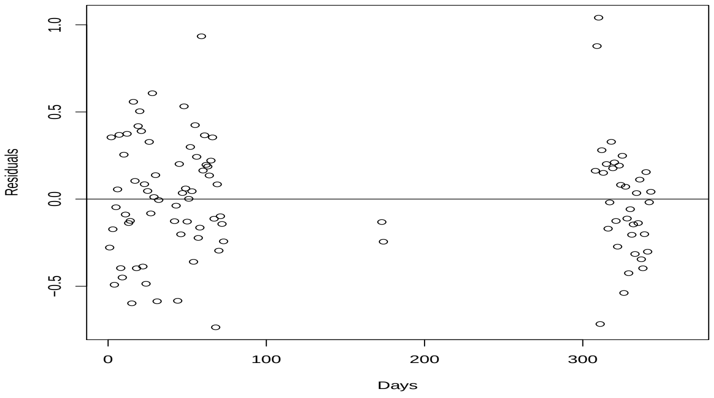
2.5 PM_{10} Monitoring Site Data Conclusion

The yearly modelling process thus far has used only three of the 11 sites for 2011 only to illustrate each step. This process has concluded Model 5 appears to model the log PM_{10} values across each of the three sites best. This means that log PM_{10} depends on the regression terms for different periods in time - day of the year and week and the meteorological variable humidity. The aim of the modelling process is to model log PM_{10} concentrations for each of the sites and years with the same model in order to compare the similarities across time and space. Table 2.11, 2.12 and 2.13 display the estimates and standard errors for Model 5 (which includes DOY, DOW and humidity) for 2010, 2011 and 2012 respectively to provide an overall idea of how the variable dependence differs across the sites and years. The variables which are significant in the model for each of the sites are displayed in bold in the tables. Table 2.11 displays the estimates and standard errors for 2010 for each of the available nine sites. There is a difference in the independent variables across each of the sites. The DOY variable is significant for seven out of the nine available sites for 2010, humidity is significant for only one of the sites and the DOY factor is significant for fove of the sites with the days of the week varying from Thursday to Sunday.

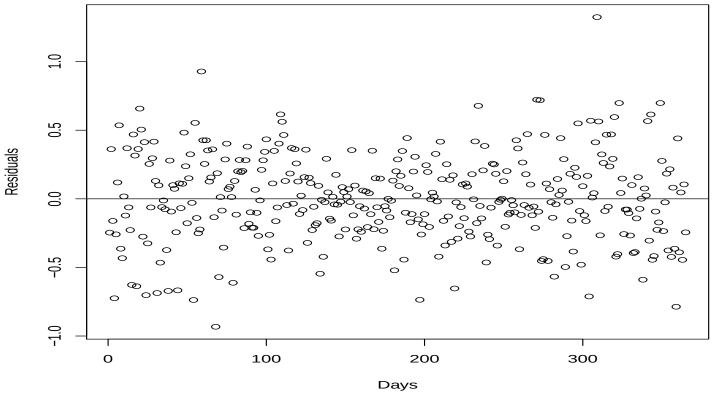
The estimates and standard errors for 2011, in Table 2.12, tell a similar story to those for 2010 however the humidity variable is significant for four of the sites. The Ljung-Box p-value is more than 0.05 for each of the monitoring



(a) Anderston, 2011

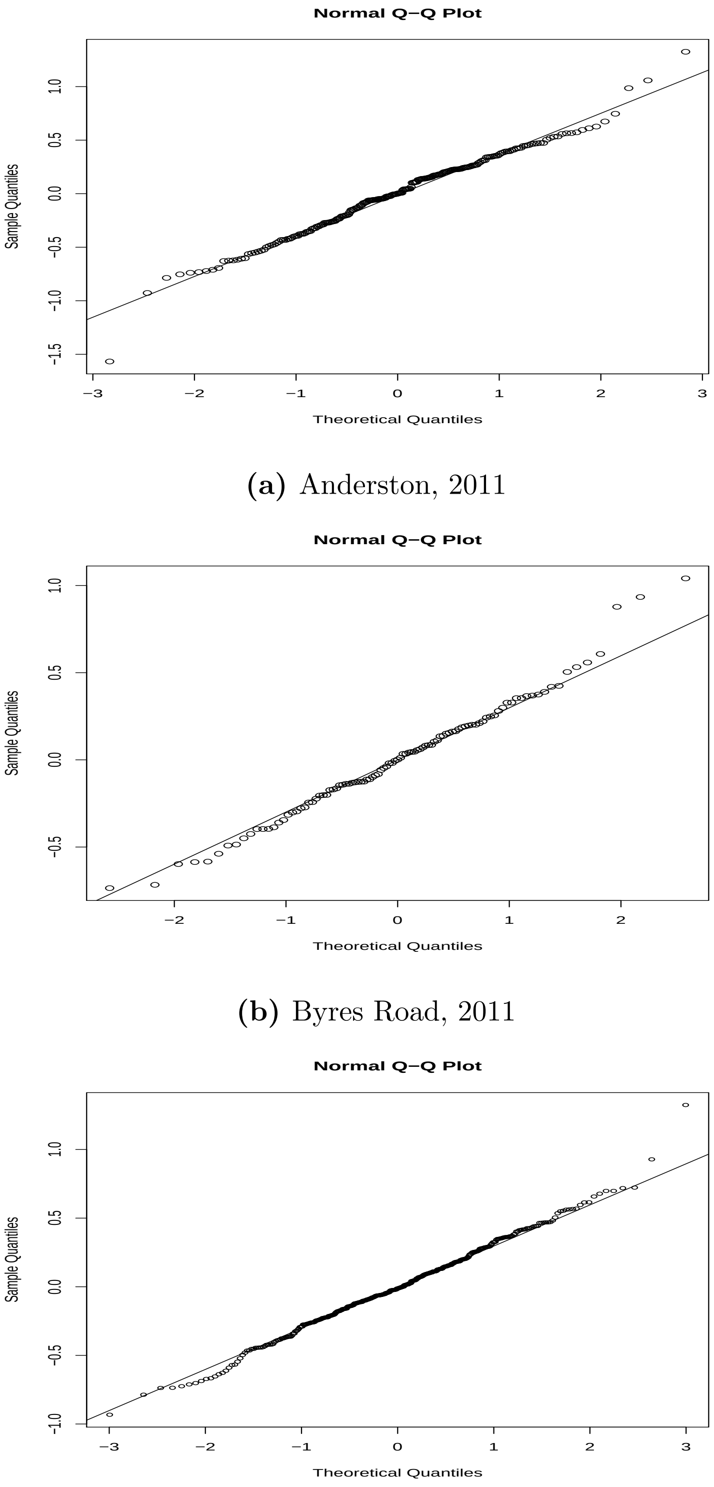


(b) Byres Road, 2011



(c) Nithsdale Road, 2011

Figure 2.12: Logged PM_{10} Residual Values with Zero Line



(a) Anderston, 2011

(b) Byres Road, 2011

(c) Nithsdale Road, 2011

Figure 2.13: Logged PM_{10} Residual Values with Zero Line

sites suggesting that there is no trend or autocorrelation left in the residuals. The table in 2012, Table 2.13, displays the estimates and standard errors for 2012. Again, the 2012 table tells a similar story to the two previous years, however, the DOY factor is significant on for Sunday. In addition to this, three of the sites have Ljung-Box p-value less than 0.05 which suggests that there is not no trend or autocorrelation left in the residuals.

Table 2.11: Estimates, standard errors, AIC and the Ljung box test statistic (2010)

	AS	A	BTR	B	BR	BS	C	DR	K	NR	WR
	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate
Intercept (β_0)	**3.046**	**2.611**	**3.404**	**2.560**	**2.846**		**2.455**		**3.027**	**2.425**	**1.824**
S.E Intercept (β_0)	**0.284**	**0.301**	**0.246**	**0.287**	**0.219**		**1.252**		**0.224**	**0.285**	**0.271**
DOY $\widehat{A}$	**0.224**	**0.260**	**0.214**	**0.261**	**0.148**		1.235		**0.362**	**0.337**	0.074
S.E DOY $\widehat{A}$	**0.065**	**0.071**	**0.060**	**0.068**	**0.057**		0.668		**0.058**	**0.080**	0.052
DOY $\widehat{\psi}$	0.600	**0.671**	0.580	**0.723**	0.703		-1.249		**0.436**	**0.751**	1.322
S.E DOY $\widehat{\psi}$	0.315	**0.327**	0.311	**0.287**	0.512		1.023		**0.156**	**0.268**	0.984
DOW 2 (β_3)	0.007	0.043	0.041	0.092	0.059		0.192		0.061	0.013	-0.002
S.E DOW 2 (β_3)	0.056	0.063	0.065	0.055	0.059		0.192		0.061	0.013	-0.002
DOW 3 (β_4)	0.071	0.076	0.083	0.044	0.060		0.153		0.062	0.077	0.073
S.E DOW 3 (β_4)	0.079	0.080	0.069	0.079	**0.152**		**0.240**		0.083	0.101	0.071
DOW 4 (β_5)	**0.199**	0.116	0.134	0.103	**0.065**		**0.161**		0.067	0.083	0.079
S.E DOW 4 (β_5)	**0.084**	0.087	0.074	0.086	**0.065**		**0.161**		0.067	0.083	0.079
DOW 5 (β_6)	0.106	-0.036	0.126	0.047	**0.151**		0.083		0.110	0.034	0.007
S.E DOW 5 (β_6)	0.083	0.088	0.073	0.085	**0.064**		0.164		0.066	0.083	0.078
DOW 6 (β_7)	-0.014	-0.117	0.049	-0.011	0.053		0.159		0.031	0.011	-0.041
S.E DOW 6 (β_7)	0.079	0.080	0.068	0.080	0.061		0.154		0.062	0.077	0.074
DOW 7 (β_8)	-0.063	**-0.148**	-0.016	-0.047	-0.013		0.138		**-0.117**	-0.084	-0.054
S.E DOW 7 (β_8)	0.064	**0.065**	0.056	0.064	0.049		0.124		**0.050**	0.062	0.059
Humidity (β_9)	-0.063	**-0.148**	-0.016	-0.047	-0.013		0.138		-0.117	-0.084	-0.054
S.E Humidity (β_9)	0.064	**0.064**	0.056	0.064	0.049		0.124		0.050	0.062	0.059
AR(1) (ϕ)	**0.524**	**0.568**	**0.554**	**0.544**	**0.590**		**0.599**		**0.573**	**0.595**	**0.596**
S.E AR(1) (ϕ)	**0.048**	**0.048**	**0.047**	**0.046**	**0.044**		**0.082**		**0.043**	**0.047**	**0.044**
Ljung -box	**0.370**	**0.955**	**0.221**	**0.134**	**0.489**		**0.842**		**0.981**	**0.429**	**0.300**
AIC	321.302	277.837	242.108	354.981	177.470		122.158		222.263	236.425	304.967

Table 2.12: Estimates, standard errors, AIC and the Ljung box test statistic (2011)

	AS	A	BTR	B	BR	BS	C	DR	K	NR	WR
	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate
Intercept (β_0)	**3.715**	**2.783**	**3.198**	**3.431**	**2.622**	**2.144**	**2.827**			**3.310**	**2.682**
S.E Intercept (β_0)	**0.293**	**0.390**	**0.267**	**0.247**	**0.579**	**0.749**	**0.254**			**0.240**	**0.309**
DOY $\widehat{A}$	**0.263**	**0.183**	**0.198**	**0.208**	0.114	**0.674**	**0.129**			**0.307**	0.090
S.E DOY $\widehat{A}$	**0.069**	**0.086**	**0.067**	**0.066**	0.137	**0.165**	**0.060**			**0.062**	0.060
DOY $\widehat{\psi}$	1.423	0.323	1.311	1.054	0.997	-0.459	0.817			**0.896**	1.551
S.E DOY $\widehat{\psi}$	1.262	0.616	1.099	0.637	0.844	0.366	0.696			**0.244**	1.048
DOW 2 (β_3)	0.007	**0.044**	-0.021	0.024	-0.087	-0.109	0.016			0.002	0.044
S.E DOW 2 (β_3)	0.056	**0.075**	0.053	0.048	0.094	0.138	0.049			0.047	0.060
DOW 3 (β_4)	0.069	-0.029	0.057	0.035	-0.022	-0.199	0.062			0.025	0.042
S.E DOW 3 (β_4)	0.071	0.093	0.066	0.061	0.120	0.155	0.062			0.059	0.074
DOW 4 (β_5)	**0.157**	0.105	**0.156**	0.115	0.050	-0.084	0.117			0.086	0.127
S.E DOW 4 (β_5)	**0.076**	0.099	**0.071**	0.065	0.130	0.160	0.065			0.063	0.079
DOW 5 (β_6)	**0.157**	0.017	0.125	**0.139**	0.127	-0.033	**0.157**			**0.155**	0.127
S.E DOW 5 (β_6)	**0.075**	0.100	0.071	**0.065**	0.126	0.159	**0.065**			**0.063**	0.079
DOW 6 (β_7)	0.010	**0.040**	0.069	0.035	-0.008	**-0.302**	0.084			0.046	0.060
S.E DOW 6 (β_7)	0.069	**0.091**	0.066	0.060	0.115	**0.155**	0.060			0.059	0.074
DOW 7 (β_8)	-0.102	-0.138	0.000	-0.030	0.059	**-0.598**	-0.009			-0.069	-0.006
S.E DOW 7 (β_8)	0.055	0.073	0.054	0.047	0.092	**0.136**	0.048			0.047	0.059
Humidity (β_9)	**-0.012**	-0.004	**-0.006**	**-0.009**	0.004	0.006	-0.002			**-0.007**	-0.004
S.E Humidity (β_9)	**0.003**	0.005	**0.003**	**0.003**	0.006	0.009	0.003			**0.003**	0.004
AR(1) (ϕ)	**0.626**	**0.640**	**0.627**	**0.650**	**0.680**	**0.315**	**0.632**			**0.622**	**0.612**
S.E AR(1) (ϕ)	**0.041**	**0.051**	**0.043**	**0.041**	**0.071**	**0.080**	**0.044**			**0.041**	**0.046**
Ljung -box	0.765	0.038	0.453	0.728	0.596	0.842	0.165			0.838	0.286
AIC	299.056	220.833	230.670	199.817	89.126	241.491	175.373			202.930	260.447

Table 2.13: Estimates, standard errors, AIC and the Ljung box test statistic (2012)

	AS	A	BTR	B	BR	BS	C	DR	K	NR	WR
	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate
Intercept (β_0)	**3.300**	**2.628**		**2.673**	**2.983**	**1.978**	**2.534**	**3.343**	**2.903**	**3.059**	**2.256**
S.E Intercept (β_0)	**0.306**	**0.329**		**0.278**	**0.311**	**0.231**	**0.273**	**0.254**	**0.306**	**0.270**	**0.298**
DOY $\widehat{A}$	**0.146**	0.102		**0.211**	0.118	**0.246**	0.144	0.061	**0.237**	**0.283**	0.106
S.E DOY $\widehat{A}$	**0.069**	0.069		**0.068**	0.065	**0.078**	0.098	0.053	**0.080**	**0.061**	0.066
DOY $\widehat{\psi}$	0.177	0.786		1.048	0.884	0.944	1.105	0.036	0.753	**0.905**	-1.103
S.E DOY $\widehat{\psi}$	0.538	0.850		0.666	0.854	0.488	0.976	0.855	0.493	**0.234**	0.957
DOW 2 (β_3)	-0.031	-0.030		-0.060	-0.015	-0.026	-0.032	-0.007	-0.024	-0.011	-0.097
S.E DOW 2 (β_3)	0.063	0.066		0.057	0.067	0.049	0.056	0.054	0.066	0.057	0.062
DOW 3 (β_4)	-0.024	-0.044		-0.020	0.035	-0.006	-0.024	0.004	0.079	-0.014	-0.099
S.E DOW 3 (β_4)	0.079	0.082		0.071	0.082	0.061	0.071	0.067	0.084	0.070	0.077
DOW 4 (β_5)	0.074	0.147		0.101	0.127	0.092	0.130	0.007	0.150	0.075	0.011
S.E DOW 4 (β_5)	0.084	0.088		0.076	0.087	0.065	0.076	0.071	0.090	0.074	0.082
DOW 5 (β_6)	-0.060	-0.052		-0.039	-0.002	0.017	0.085	-0.110	0.067	-0.021	-0.060
S.E DOW 5 (β_6)	0.084	0.088		0.075	0.088	0.065	0.077	0.071	0.089	0.074	0.083
DOW 6 (β_7)	-0.104	-0.023		-0.028	-0.084	0.000	0.031	-0.116	0.042	-0.053	-0.063
S.E DOW 6 (β_7)	0.078	0.083		0.070	0.082	0.060	0.071	0.066	0.083	0.070	0.077
DOW 7 (β_8)	-0.134	**-0.120**		-0.091	**-0.149**	-0.055	-0.085	**-0.200**	-0.095	**-0.139**	-0.047
S.E DOW 7 (β_8)	0.063	**0.063**		0.056	**0.066**	0.047	0.056	**0.054**	0.068	**0.056**	0.062
Humidity (β_9)	**-0.009**	-0.001		-0.001	-0.005	**0.007**	0.001	**-0.006**	0.002	0.001	0.001
S.E Humidity (β_9)	**0.003**	0.003		0.003	0.003	**0.003**	0.003	**0.003**	0.003	0.003	0.003
AR(1) (ϕ)	**0.585**	**0.625**		**0.607**	**0.576**	0.693	**0.689**	**0.564**	**0.591**	**0.564**	**0.632**
S.E AR(1) (ϕ)	**0.044**	**0.046**		**0.043**	**0.043**	0.043	**0.048**	**0.053**	**0.057**	**0.045**	**0.045**
Ljung -box	**0.279**	**0.110**		**0.399**	**0.008**	**0.727**	**0.384**	**0.376**	**0.845**	**0.615**	0.041
AIC	367.697	278.948		301.235	421.666	132.365	111.237	91.652	139.936	289.457	248.289

Earlier in this chapter, Section 2.2.2, a plot of correlations between sites against the distance of the sites, Figure 2.5b was discussed to assess if the distance between the sites had an effect on the correlation. This plot was shown only for 2011 as all three of the plots showed a similar trend. It was found that this was found to be mostly true with the exception of seven of the Burgher Street pairs. With the PM_{10} data modelled, the plot in Figure 2.14 shows the correlation values against the distances for pairs of sites with the trend removed. This plot is very similar to the previous, this would suggest that even with the trend removed from the log PM_{10} data, the correlation between the sites are strongly related to the distance between them. Suggesting, again, that there is a spatial trend which has not been exploited.

To conclude, this chapter has explored and modelled trend and seasonality in PM_{10} across 11 sites for three years. Fitting the same model to each of the sites across the three years has shown the differences and similarities in the distribution of PM_{10} over time and space. PM_{10} appears to have a seasonal pattern which is modelled using a harmonic regression day of the year term and a day of the week factor. The meteorological variable humidity also appears to account for some of the variation in time. PM_{10}, however, has hugely volatile and the models do not account for all of the variation. By fitting the same models across the sites it shows the stark difference between them. Although PM_{10} was modelled somewhat for all of the sites there were vast differences between the profile of significant variables. This suggests that PM_{10} has a spatial variation as well as a temporal variation. Exploring these sites across the three years has given us an initial understanding of how PM_{10} is distributed across time and somewhat across space. This leads us onto looking at the second available dataset for the gridded modelled annual mean PM_{10} data to explore the spatial distribution of PM_{10} further.

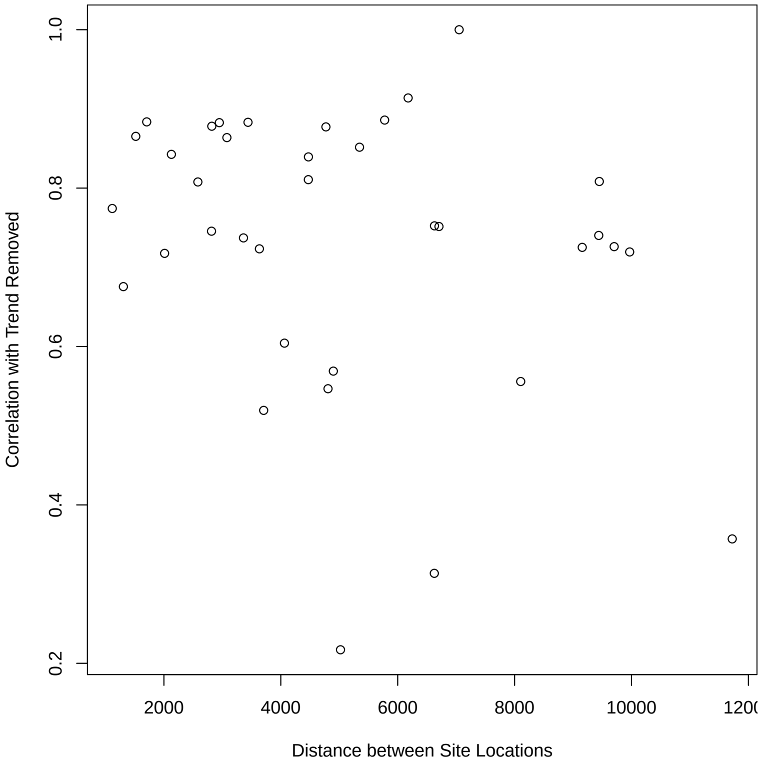


Figure 2.14: Plot of correlation between sites against the distance between the sites with the trend removed for the year 2011

Chapter 3

Modelling the Spatial Trend and Dependence in the Gridded Modelled Annual Mean PM_{10} Data

The previous chapter explored the distribution of PM_{10} across time at 11 different monitoring sites across Glasgow. It was found that the distribution of PM_{10} is not constant over space. However these monitoring sites are not placed uniformly across Glasgow but appear to be arranged mostly in the centre of the city. The gridded modelled annual mean PM_{10} data, introduced in Section 1.5.3, comes in a grid format and can provide more of an insight into the spatial aspect of PM_{10} uniformly in 1×1km grids across Glasgow. These spatial data, while lacking in temporal accuracy, should provide a more accurate description of PM_{10} levels across the city and can also be incorporated into the air pollution indicator. In addition to the annual mean PM_{10} estimates, variables that were used to initially model the PM_{10} levels at each grid square location were included in the data set. This included a spatially located binary variable for motorways and main A and B roads, named in this context as the *motorway* covariate. This chapter starts by explaining the main methods used to analyse this spatial dataset. After an exploration of possible spatial patterns of PM_{10} the gridded values are investigated formally using a spatial statistical model. This chapter ends with a discussion of the main spatial trends seen in PM_{10} across Glasgow.

3.1 Methods Used to Explore the Gridded Modelled Annual Mean PM_{10} Data

3.1.1 Geostatistical Modelling

Spatial Process

To explain geostatistical modelling, a spatial process must first be defined. Spatial data can be thought of being generated by the stochastic process, Y, but with a spatial index (instead of time) indicating locations or regions:

$$\{Y(\underline{s}) : \underline{s} \epsilon D\} \tag{3.1}$$

Here D denotes the spatial domain which in this case is a fixed subset of 2-dimensional space, R^2. For PM_{10} we observe a finite number of locations $(\underline{s_1}, \ldots, \underline{s_n})^T$ where, in this case, the number of spatial locations is n=175. The mean function of a geostatistical process $Y(\underline{s})$ is defined to be $\mu(\underline{s}) = E(Y(\underline{s}))$ for each $\underline{s} \epsilon D$.

Stationarity and Isotropy

In geostatistics if a process is *stationary* then the absolute coordinates that we observe the process are unimportant but the direction and difference between locations are important. If only the distance is important then the process is said to *isotropic.*

A geostatistical process is *strictly stationary* if the random vectors $(Y(\underline{s_1}), \ldots, Y(\underline{s_n}))$ and $(Y(\underline{s_1} + \underline{h}), \ldots, Y(\underline{s_n} + \underline{h}))$ have the same joint distribution for all $n \geq 1$

where $\underline{h}$ denotes some displacement.

$$(Y(\underline{s_1}), \ldots, Y(\underline{s_n})) =_d (Y(\underline{s_1} + \underline{h}), \ldots, Y(\underline{s_n} + \underline{h})), \tag{3.2}$$

If a process is strictly stationary then $\{Y(\underline{s}) : \underline{s} \epsilon D\}$ is identically distributed and the locations themselves do not affect the distribution - only the displacement between locations matter.

A process is (weakly) stationary if $E(Y(\underline{s})) = \mu(\underline{s}) = \mu$ if μ is a constant which does not depend on $\underline{s}$ and the covariance function is a finite constant which depends on $\underline{h}$ but does not depend on $\underline{s}$; we have $\text{cov}(Y(\underline{s}), Y(\underline{s}+\underline{h})) = C(\underline{s}, \underline{s} + \underline{h}) = C(\underline{h})$. Note that a strictly stationary process is also weakly stationary as long as $\mu(\underline{s})$ is finite. The covariance function at a displacement $\underline{h}$ of a (weakly) stationary process is defined as:

$$\begin{aligned} C(\underline{h}) &= \text{cov}(Y(\underline{s}), Y(\underline{s} + \underline{h})) \\ &= E((Y(\underline{s}) - \mu)(Y(\underline{s} + \underline{h}) - \mu)). \end{aligned} \tag{3.3}$$

The correlation function of a stationary process is defined as:

$$\rho(\underline{h}) = \frac{C(\underline{h})}{\sqrt{C(\underline{0})C(\underline{0})}} = \frac{C(\underline{h})}{C(\underline{0})} \tag{3.4}$$

A stationary process is said to be an isotropic process if $C(\underline{s}, \underline{s'})$ only depends on the distance between the locations, $\|\underline{s} - \underline{s'}\|$, and not the direction. The covariance function of an isotropic process can be written as:

$$C(\underline{s}, \underline{s} + \underline{h}) = C(\|\underline{h}\|) \tag{3.5}$$

Variogram

A variogram is one way of measuring of spatial dependence - it measures the variance between two spatial locations in a geostatistical process. The variogam is denoted as $\gamma(\underline{s}, \underline{s'})$ and the commonly used semi-variogram is denoted by $2\gamma(\underline{s}, \underline{s'})$. Both the variogram and semi-variogram can be written in terms of the covariances, when they exist:

$$\begin{aligned} 2\gamma(\underline{s}, \underline{s'}) &= \mathrm{var}(Z(\underline{s}) - Z(\underline{s'})) \\ &= \mathrm{cov}(Z(\underline{s}) - Z(\underline{s'}), Z(\underline{s}) - Z(\underline{s'})) \\ &= C(\underline{s}, \underline{s}) + C(\underline{s'}, \underline{s'}) - 2C(\underline{s}, \underline{s'}) \qquad (3.6) \end{aligned}$$

and when Z is stationary

$$C(\underline{h}) = \lim_{\|\underline{u}\| \to \infty} \gamma(\underline{u}) - \gamma(\underline{h}) \qquad (3.7)$$

Both the variogram and semi-variogram have the following descriptive parameters: the nugget (ϕ^2) which is the difference between the origin line and the limiting value of the variogram as t → 0, the sill which is the limiting value of the variogram as t → 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 (Cressie and Hawkins, 1980).

A binned empirical variogram is often used in conjunction with the variogram in order to identify a spatial structure more clearly. The binning process partitions the distances into H intervals, called bins, where

$$I_l = (t_{l-1}, t_l], l = 1, \ldots, L. \qquad (3.8)$$

If we let $t_l^m = (t_{l-1} + t_l)/2$ denote the midpoint the pairs of distances for each of the L intervals then the binned empirical variogram is given by

$$2\widehat{\gamma}(t_l^m) = \frac{1}{jN(t_l)} \sum_{(s_i, s_j) \varepsilon N(t_l)} [y(s_i) - y(s_j)]^2, \qquad (3.9)$$

where $N(t_l) = \{(si, sj) : \|si - sj\| \epsilon I_l\}$. Caution should be used when interpreting binned empirical variograms, however, as measures of uncertainty are not easily calculated. Typically there may not be enough pairs in the bins especially for the bins at longer distances and therefore care must be taken in interpretation. To ensure an accurate representation of potential correlation structure it is advised in some cases that empirical variograms should only be trusted at half the maximum distance.

There are a number of parametric models for the variogram and covariance function that can be used for geostatistical modelling. The most common parametric model used is the exponential variogram and covariance function. The exponential variogram and the covariance function are expressed below where and $t = \|s_i - s_j\|$:

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

and

$$\gamma(t) = \begin{cases} \phi^2 + \sigma^2(1 - \exp(\frac{-t}{\lambda})) & \text{if } t \geq 0; \\ 0 & \text{if } t = 0. \end{cases}$$

The Gaussian covariance/ variogram is another example of a parametric model which can be used for geostatistical modelling. The Gaussian covariance/ variogram gives a much smoother process then the exponential one. The Gaussian covariance function and variogram are respectively:

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

and

$$\gamma(t) = \begin{cases} \phi^2 + \sigma^2(1 - \exp(\frac{-t}{\lambda})^2) & \text{if } t \geq 0; \\ 0 & \text{if } t = 0. \end{cases}$$

In a case where the covariance increases and decreases with time, the wave exponential covariance/variogram could be used:

$$C(t) = \begin{cases} \sigma^2[\frac{\sin\frac{t}{\phi}}{\frac{t}{\phi}}] & \text{if } t \geq 0; \\ \phi^2 + \sigma^2 & \text{if } t = 0, \end{cases}$$

and

$$\gamma(t) = \begin{cases} \phi^2 + \sigma^2[\frac{1-\sin\frac{t}{\phi}}{\frac{t}{\phi}}] & \text{if } t \geq 0; \\ 0 & \text{if } t = 0. \end{cases}$$

These are just three examples of parametric models which can be used for covariances/variograms which give an idea of how the spatial dependence and relationships between ϕ^2, σ^2 and λ differ as a function of distance.

Empirical Variogram

A variogram assumes isotropy - that the variogram depends only on the distance, not the direction, however isotropy is not always a reasonable assumption. A directional variogram is one of the most simple methods to test this assumption. A directional variogram combines multiple different angled variograms into a single variogram, if each of these variograms follow the same trend then isotropy can be assumed.

3.2 Estimating Model Parameters

3.2.1 Maximum Likelihood Estimation

Suppose $y(\underline{s}) : \underline{s}\epsilon D$ is a Gaussian geostatistical process with mean $\mu(\underline{s}) = \underline{x}^T(\underline{s})\underline{\beta}$ and covariance $C_\theta(\underline{s}, \underline{t})$. We can write this as a regression model

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

where $\varepsilon(\underline{s})$ has mean zero. Given the mean parameters $\underline{\beta}$ and covariance parameters $\underline{\theta}$ the likelihood of the data $\underline{y} = (y_1, \ldots, y_n)^T$ at locations $s_i(i = 1, \ldots, n)$ is explained in Equation (3.10), where n equals the sample size and $\Sigma_{\underline{\theta}}$ is the covariance matrix of $y(\underline{s})$ with (i, j) element $C_\theta(s_i, s_j)$:

$$L(\underline{\beta},\underline{\theta}) = (2\pi)^{(-n/2)}(\text{det}\Sigma_{\underline{\theta}})^{-1/2}\text{exp}(-\frac{1}{2}(\underline{y}-\underline{X}\underline{\beta})^T\Sigma_{\underline{\theta}}^{-1}(\underline{y}-\underline{X}\underline{\beta})). \tag{3.10}$$

The log-likelihood is then calculated as:

$$l(\underline{\beta},\underline{\theta}) = -\frac{1}{2}\text{log}(2\pi) - \frac{1}{2}\text{log}(\text{det}\Sigma_{\underline{\theta}}) - \frac{1}{2}(\underline{y}-\underline{X}\underline{\beta})^T\Sigma_{\underline{\theta}}^{-1}(\underline{y}-\underline{X}\underline{\beta})). \tag{3.11}$$

If the derivative of $l(\underline{\beta},\underline{\theta})$ is then calculated with respect to $\underline{\beta}$ and set equal to zero then the MLE of $\underline{\beta}$ is the Generalised Least Squares (GLS) estimator:

$$\widehat{\underline{\beta}}(\underline{\theta}) = (\underline{X}^T\Sigma_{\underline{\theta}}^{-1}\underline{X})^{-1}\underline{X}^T\Sigma_{\underline{\theta}}^{-1}\underline{y}. \tag{3.12}$$

As you can see the MLE of $\underline{\beta}$ is dependent on the spatial parameters θ. This can simply be plugged back into the log-likelihood in Equation (3.11) and maximised with respect to $\underline{\theta}$ to get the MLE for $\underline{\theta}$. However, this method would mean that the estimate of $\underline{\beta}$ may introduce a bias in θ. The Restricted Maximum Likelihood (REML) approach is an alternative approach which can minimise the bias when estimating $\underline{\theta}$ Patterson and Thompson (1971).

3.2.2 Restricted Maximum Likelihood

The REML approach is a form of maximum likelihood estimation which again requires that y follows has a multivariate normal distribution. This method is used to estimate the spatial model parameters $\underline{\theta} = (\phi^2, \sigma^2, \lambda)^T$ where the parameters in $\underline{\theta}$ denote the nugget, sill and the range respectively. In place of the standard maximum likelihood, the restricted maximum likelihood can be used to ensure less biased estimates of $\underline{\theta}$ by calculating the likelihood function from a transformed set of data which ensures that the nuisance parameters have no effect on the estimates. As explained the Gaussian random fields model is defined in Equation (3.13) where $\mu(s) = x(s)^T\beta$, Z(s) denotes a stationary Gaussian process with variance σ^2 and the correlation defined by λ and the ε is the error term which had the variance parameter ϕ^2 (Ribeiro and Diggle, 2013).

$$Y(s) = \mu(s) + Z(s) + \varepsilon. \tag{3.13}$$

Under this model for $E[Y] = X\beta$, the data can be transformed linearly to $Y^* = AY = X(X^TX)^{-1}X^TY$ where Y^* does not depend on β, (Diggle and Ribeiro, 2007). The model remains multivariate Gaussian after Y is linearly transformed. The constraint imposed that Y^* not depending on β means that the dimensionality of y is reduced from n to n-p, where p denotes the rank of X. The REML estimates for $\underline{\theta}$ are then computed by maximising the likelihood for $\underline{\theta}$ based on Y^*.

3.3 Exploring Spatial Trends of Gridded Modelled Annual Mean PM$_{10}$ Data

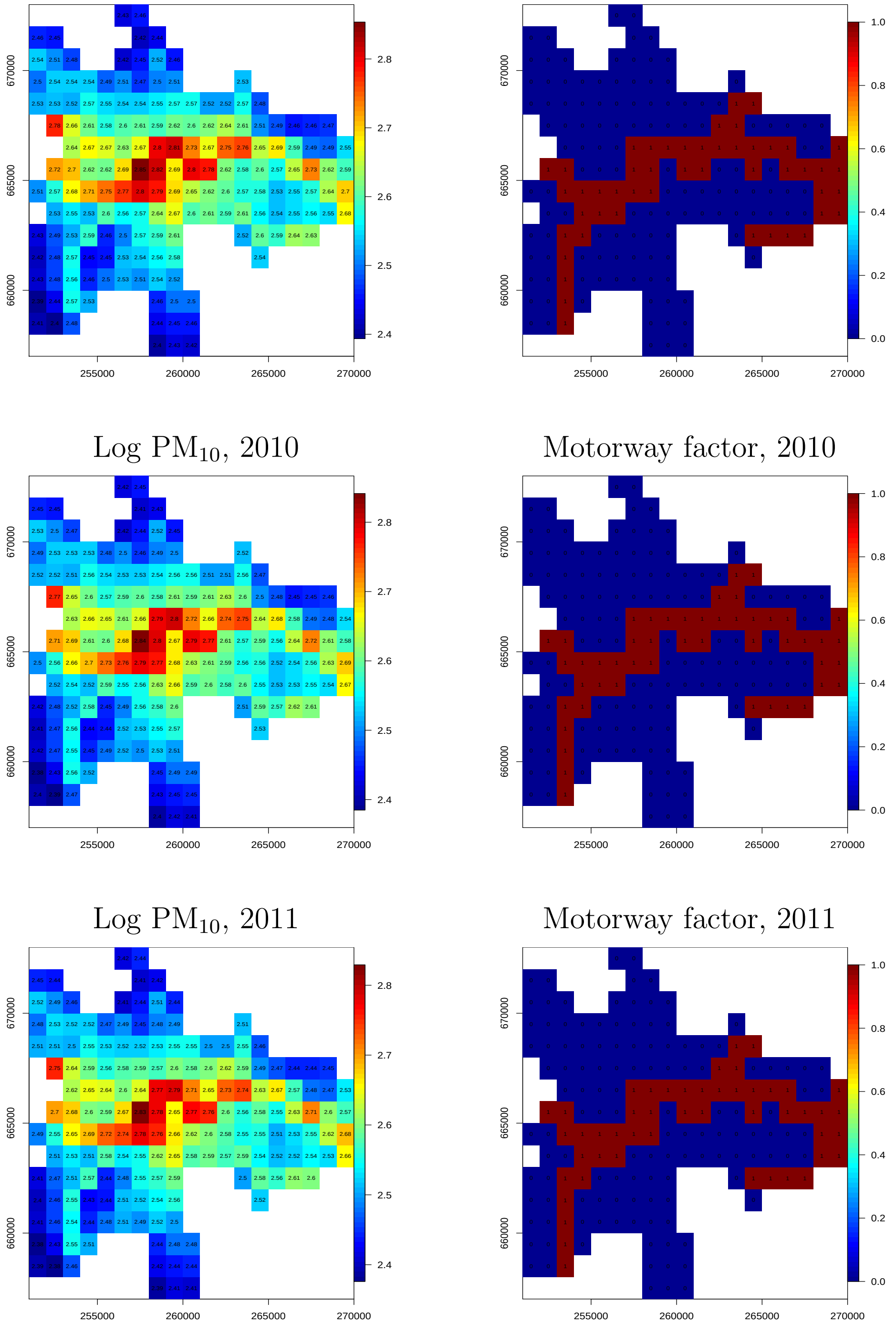
This section explores the spatial distribution of the gridded PM$_{10}$ model across Glasgow for the three years assuming that each year is independent of the other years. It provides an idea of where in Glasgow appears to have the highest and lowest concentrations of PM$_{10}$ and the form of the spatial trend. Each grid square gives an annual mean modelled concentration. Table 3.1 displays a summary of the gridded modelled concentrations for each year. It would appear that across the minimum, median, mean and maximum values that PM$_{10}$ appears to be slowly decreasing by year as is the standard deviations. This would suggest that overall PM$_{10}$ levels could be decreasing and that variability is also decreasing. The range of values from the minimum to the maximum and also the difference between the 1st and 3rd quantiles appears constant over time.

Looking at the spread of PM$_{10}$ values the decision was made to log the PM$_{10}$ concentrations, partly to make the concentrations more normally dis-

Table 3.1: Summary of the Previously Modelled Annual Mean PM_{10} Data for 2010 - 2012

	Min	1st Qu	Median	Mean	3rd Qu	Max	St.Dev
2010	10.95	12.15	12.96	13.12	13.75	17.35	1.33
2011	10.86	12.04	12.82	12.99	13.60	17.14	1.30
2012	10.76	11.93	12.68	12.85	13.43	16.93	1.27

tributed but also to remain consistent with the previous modelling process in which the concentrations were also logged. The map of logged, previously modelled annual mean concentrations are displayed in the three plots on the lefthand side of Figure 3.1 for 2010, 2011 and 2012 respectively. The colour scale on the left hand side of each plot explains that the deep red colour denotes the high log PM_{10} concentrations and the deep blue denotes the lowest levels. The three maps appear very similar. The outskirts of Glasgow tend to have the lowest levels and the very centre appears to have the highest levels of PM_{10}. The grid square six down and two from the left has a large mean PM_{10} values with respect to the surrounding grids. This is true for all the three years. Interestingly, as each of the maps are almost identical, there appears to be a strong spatial trend across Glasgow across time. A motorway covariate which includes motorways and A and B roads are displayed in the maps down the right hand side of Figure 3.1. The motorway covariate is a binary spatially varying factor with 0 denoting no motorway and 1 denoting motorway. These maps are identical for each of the three years which means that the main motorways and roads remain unchanged as expected. The shape of the motorway covariate shows that the main motorways and roads seem to stretch across the city centre but does not stretch far north nor does it lie in the middle of the southside. The highest PM_{10} levels in the log PM_{10} maps follow a similar shape to that of the motorway factor. Both have a sinusoidal trend from the very west to the east through the city centre. However, the unusual grid to the very west of the city does not lie in the path of the motorway factor.



Log PM_{10}, 2010

Motorway factor, 2010

Log PM_{10}, 2011

Motorway factor, 2011

Log PM_{10}, 2012

Motorway factor, 2012

Figure 3.1: Map of log PM_{10} and corresponding motorway covariate map

3.4 Spatial Trend Estimation of the Gridded Modelled Annual Mean PM_{10} Data

This section aims to more formally explore the previously modelled annual mean PM_{10} data in order to gain an idea of how PM_{10} is distributed spatially. The exploratory analysis has displayed graphical and numerical summaries of the previously modelled annual mean data for the three years and for the motorway covariate. This analysis suggests that PM_{10} appears to be higher in the city centre and lower the further out of the city travelled, therefore, the modelling process should take this into account. To explain the modelling process thoroughly in this section the process is outlined for year 2010 only to reduce repetition as a similar process was used for each of the three years.

As discussed, the exploratory analysis would suggest that there is a spatial trend in PM_{10} as well as a dependence with the motorway factor. The first model that was fit included the latitude and longitude values called eastings and northings respectively and the motorway factor. The second model included these variables as well as eastings^2, northings^2 and the interaction between eastings and northings. There were two models fit in total which are summarised in Table 3.2 which follow the equation $y(s_i) = x(s_1) + \varepsilon(s_i)$. The models are then explained in Equations (3.14) and (3.15) where $y(s_i)$ corresponds to the PM_{10} value at each spatial location $i = 1, \ldots, n$, $x(s_i)$ is the design matrix which is made up of different covariates, $\underline{\beta}$ which corresponds to the regression coefficients and $\varepsilon(s_i)$ denotes the residuals which are assumed to follow a normal distribution with mean zero.

Table 3.2: Description of the Two Geostatistical Models

Model Number	Model Description
Model 1	East, north & motorway factor
Model 2	East, north, east^2, north^2 & motorway factor

$$\log y(s_i) = \beta_0 + \beta_1\text{eastings}(s_i) + \beta_2\text{northings}(s_i) + \beta_3\text{motorway(factor)}(s_i) + \varepsilon(s_i) \tag{3.14}$$

$$\log y(s_i) = \beta_0 + \beta_1\text{eastings}(s_i) + \beta_2\text{northings}(s_i) + \beta_3\text{eastings}^2(s_i) + \beta_2\text{northings}^2(s_i) + \beta_5\text{motorway(factor)}(s_i) + \varepsilon(s_i) \tag{3.15}$$

Beginning with Model 1 a linear model can be fit to the annual mean PM_{10} data in order to estimate parameters using OLS. OLS assumes independent errors however when dealing with geostatistical data the errors are dependent but if we can assume that the $\underline{\varepsilon}(\underline{s}) = 0$ then the estimates $\beta_0, \ldots, \beta_k$ are unbiased. Figure 3.2 displays the residual map. This map shows that there appears to be a clear spatial trend left in the residuals. The red colour still centres around the city centre while the values steadily decrease the further out of the city travelled with the lowest values at the border. The motorway factor appears to have accounted for the higher values that lie along the outline of the motorway. This map suggests that there could be a more complex spatial structure which including the simple eastings and northings values have not accounted for.

Figure 3.3 displays four diagnostic plots: the residuals, the Q-Q plot, the residuals against eastings and the residuals against northings. These plots provide an idea of how well these covariates estimate the spatial trend in the data. The top left plot displays the residuals map which places the residual value in the spatial location, this shows the residual value without the coloured image. This gives another clear indication that the residual values are much higher in the centre than the outskirts of the city with the highest value reaching 0.2 and the lowest values -0.2. The normal Q-Q plot

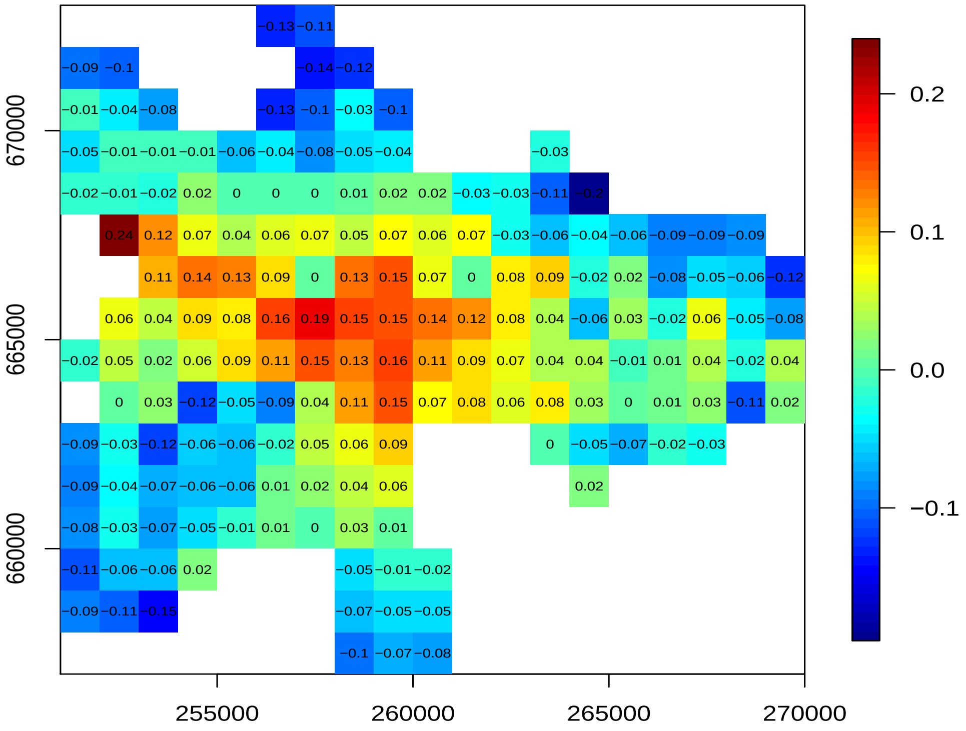
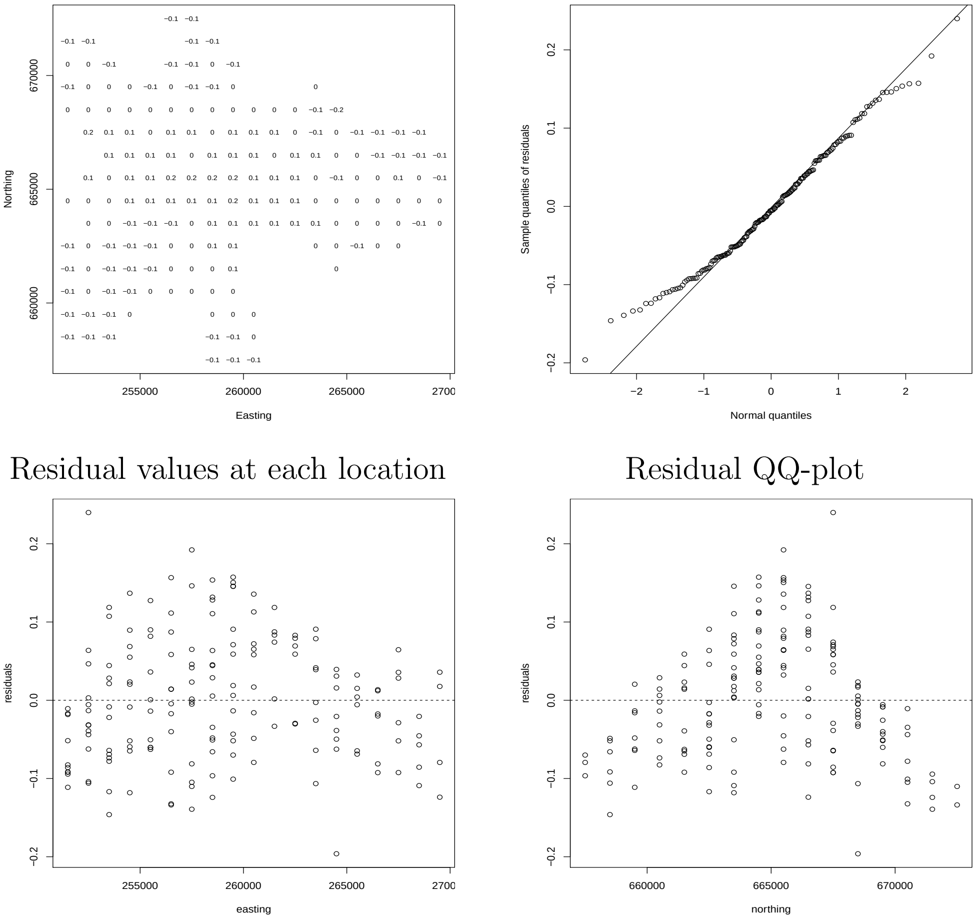


Figure 3.2: Residual map of model with eastings, eastings^2, northings, northing^2 and northings and the motorway factor (2010).

in the top right hand corner shows that the tails of the residuals do not follow the line suggesting that this is not a perfect normal fit. The bottom left plot shows the residuals against eastings and the bottom right shows the residuals against northings. Neither plot has constant variance and both plots suggest that there is spatial trend left in the residuals that could be argued to be a quadratic effect in the easting and northing variables.

Secondly, Model 2 includes the covariates contained in the previous model along with more complex spatial variables: eastings^2 and northings^2. Similar to the previous modeling process, an exploratory linear model is fit using OLS to estimate the parameters under the assumption that $\varepsilon(s) = 0$. Figure 3.4 displays the residual map. The residual map for Model 2 compared to



Residual values at each location

Residual QQ-plot

Residual by eastings

Residuals by northings

Figure 3.3: Residual plots for model 1 (2010)

the residual map for the Model 1, in Figure 3.2 shows a much more even distribution of the residuals. The city centre does not have any red grid squares which denotes very high residual values relative to the grids further out of the centre. The grids in the city centre have slightly higher residual values than the grids on the border in some cases, however, the difference is much smaller than for the previous model. The grid on the west border which did not follow the motorway covariate, however, remains geographically alone with the largest residual value. The much more even distribution of residuals suggest that Model 2 captures the spatial trend better than Model 1.

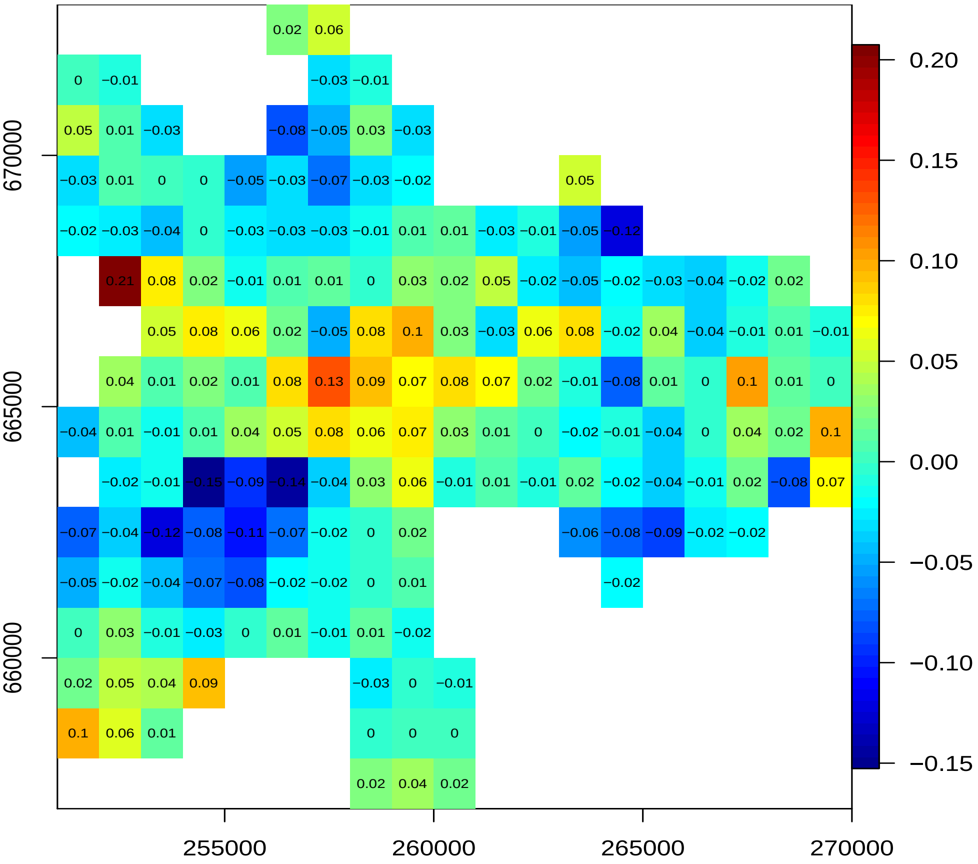
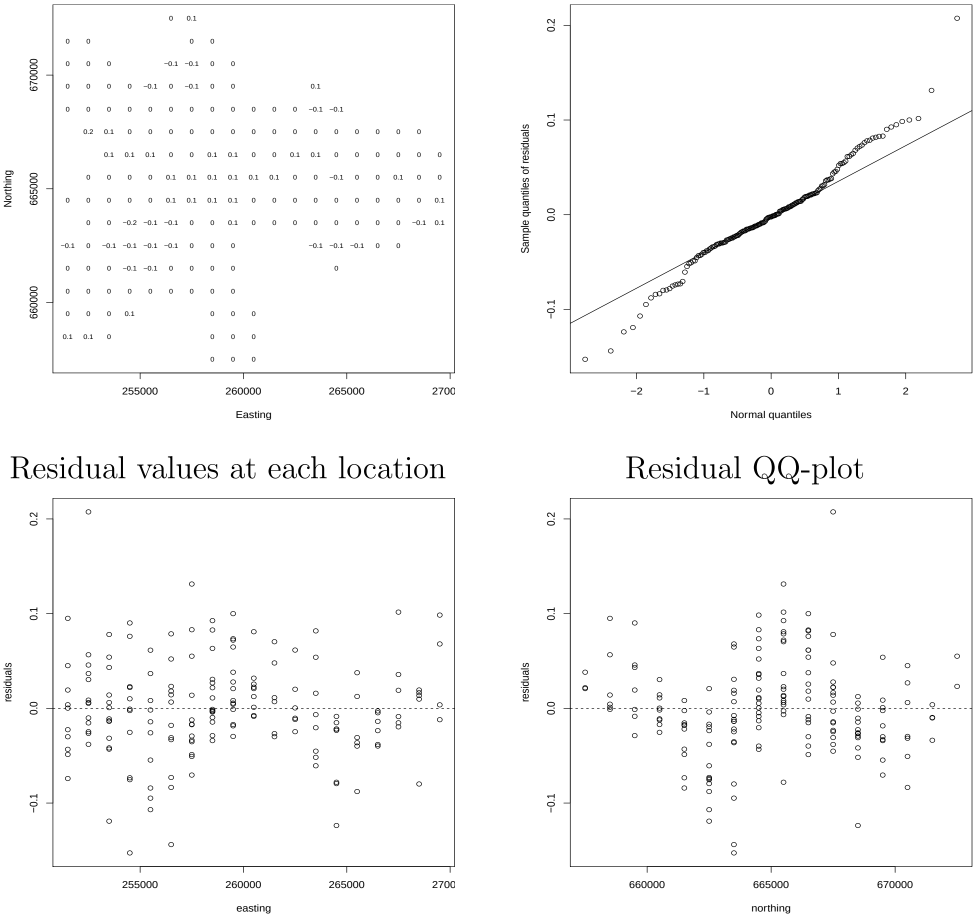


Figure 3.4: Residual map of model with eastings, eastings^2, northings, northing^2 and northings and the motorway factor (2010).

The diagnostic plots for Model 2 are displayed in Figure 3.5 and explore how well the model fits the log PM_{10} data. The residual plot in the top left hand corner shows that the actual residual values are not as extreme as they were for Model 1. The highest value in the city centre is 0.1 whilst the lowest value which occurs in the border is -0.2. The Q-Q plot in the top right hand corner shows that the residuals do not fit the Q-Q line around the tails of the distribution suggesting that this is not a perfectly normal fit. The bottom two plots display the residuals against eastings and against northings. These plots compared to their Model 1 counterparts are much less variable but there does appear to be a sinusoidal trend which had not been captured in the model which is most prominent in the northings plot.



Residual values at each location

Residual QQ-plot

Residual by eastings

Residuals by northings

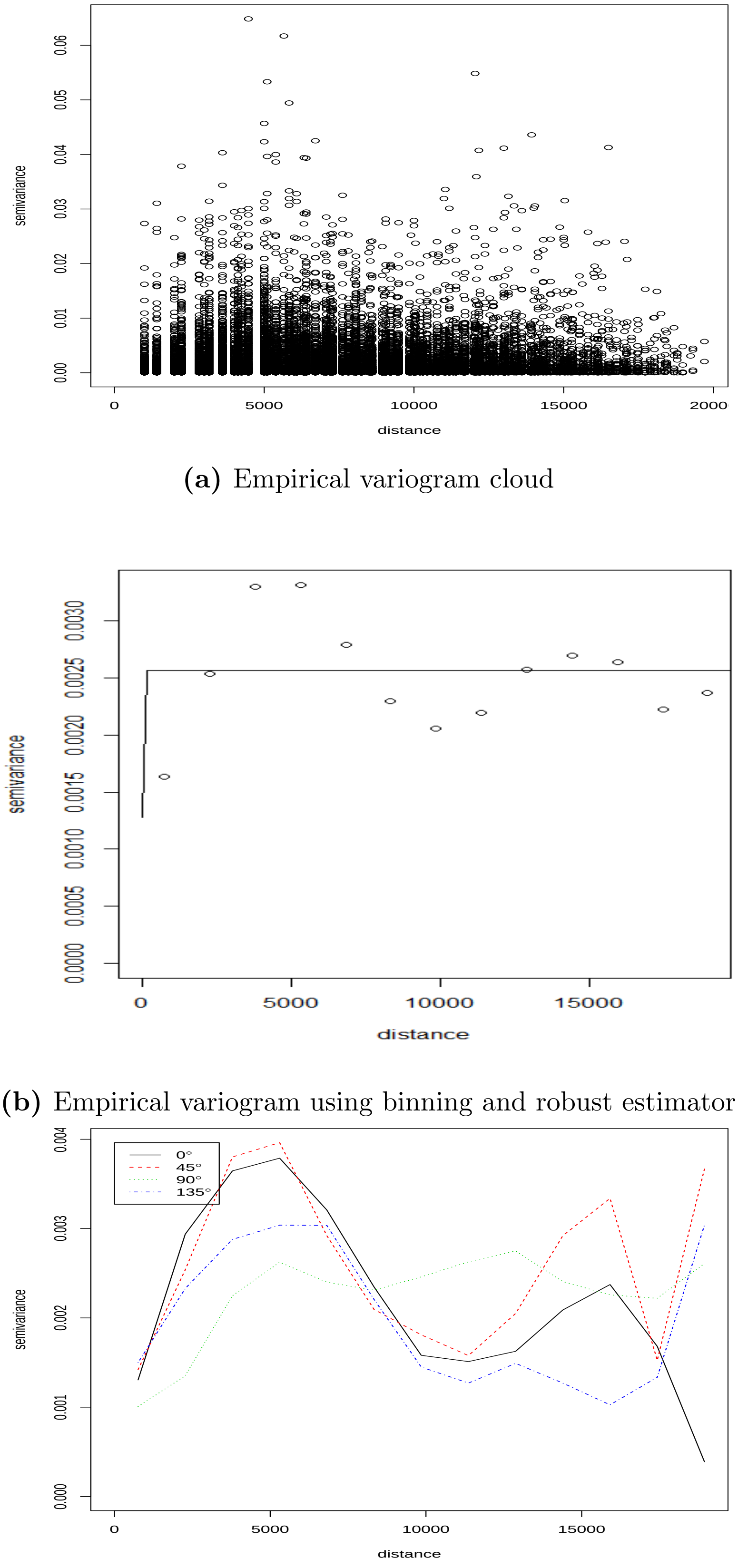
Figure 3.5: Residual plots for Model 2 (2010)

A number of plots, found in Figure 3.6, are displayed in order to assess the correlation structure and isotropy of the process. The estimated empirical variogram cloud and the binned semi-variogram in Figure 3.6a and 3.6b can be used to determine what covariance/ variogram function is suitable for the process. The semi-variogram cloud and semi-variogram shows that the smaller the distance the lower the semi variance which then tails off with a slight sinusoidal wave. This indicates that a plausible choice of spatial correlation structure of the errors could be exponential or the wave exponential covariance/ variogram. The directional variogram in Figure 3.6 shows that regardless of which angle the variogram is estimated at they appear similar, especially up to distance 8000. This suggests that the process can be assumed to be isotropic therefore the $C_{(}\underline{s}, \underline{s}')$ only depends on the distance between the locations, $\|\underline{s} - \underline{s}'\|$, and not the direction.

Model 2 would appear to contain variables which have estimated the spatial trend in the data somewhat. The process can be assumed to be isotropic and the variogram cloud and the binned empirical variogram have provided an idea of the correlation structure of the errors. This modelling process so far, however, has merely explored PM_{10} using a linear model. The next stage in the geostatistical modelling process is to take Model 2 and using the most appropriate correlation structure for the errors to estimate the model parameters using a form of maximum likelihood.

3.4.1 Estimating the Model Parameters

In order to estimate $\underline{\beta} = (\beta_0, \ldots, \beta_k)$ and the spatial model parameters $\underline{\theta} = (\phi^2, \sigma^2, \lambda)$ a maximum likelihood approach is used. The MLE and REML approach is used to estimate $\underline{\beta}$ and the REML approach is used to estimate $\underline{\theta}$. As explained in the previous chapter, either the exponential covariance function or the wave exponential covariance function could be used to estimate the correlation structure of the errors. At first, looking at the binned empirical variogram, it looks like a wave function would be the most appropriate model to estimate the correlation structure. However, the binned empirical



(a) Empirical variogram cloud

(b) Empirical variogram using binning and robust estimator

(c) Directional variogram

Figure 3.6: Multiple variograms for Model 2 (2010)

variogram should only be trusted to the midpoint of the graph, as explained in the methods section of this chapter. Therefore, excluding the distance 10,000 or more the variogarm suggests that an exponential model, which is a more simple parametric model, would estimate the correlation structure of the errors as there are no signs of a wave like function until after the mid point. The exponential covariance function:

$$C_z(t) = \begin{cases} \sigma^2 \exp(\frac{-||s_i - s_j||}{\lambda}) & \text{if } ||s_i - s_j|| \; 0; \\ \phi^2 + \sigma^2 & \text{if } ||s_i - s_j|| = 0. \end{cases}$$

Using the correlation structure for the errors above, the estimates for $\underline{\beta}$ and $\underline{\theta}$ and the standard errors are calculated and they can be found in Table 3.3. This table shows that all of the variables are significant in the model except eastings^2 and the interaction term. The modelling process has been outlined for only year 2010 therefore to gain an overall idea of how PM$_{10}$ is distributed across Glasgow each of the three years should be discussed.

Table 3.3: Table of Estimates and Standard Errors, 2010

	Estimate	St.Error
$\widehat{\beta}_0$ (Intercept)	**2.369**	**0.063**
$\widehat{\beta}_1$ (Motorway)	**0.087**	**0.009**
$\widehat{\beta}_2$ (Easting)	**0.215**	**0.135**
$\widehat{\beta}_3$ (Northing)	**0.673**	**0.211**
$\widehat{\beta}_4$ (Easting^2)	-0.233	0.154
$\widehat{\beta}_5$ (Northing^2)	**-0.658**	**0.175**
$\widehat{\phi}^2$	0.0003	
$\widehat{\sigma}^2$	0.004	
$\widehat{\lambda}$	0.197	

3.5 Previously Modelled Annual Mean PM_{10} Three Years Conclusion

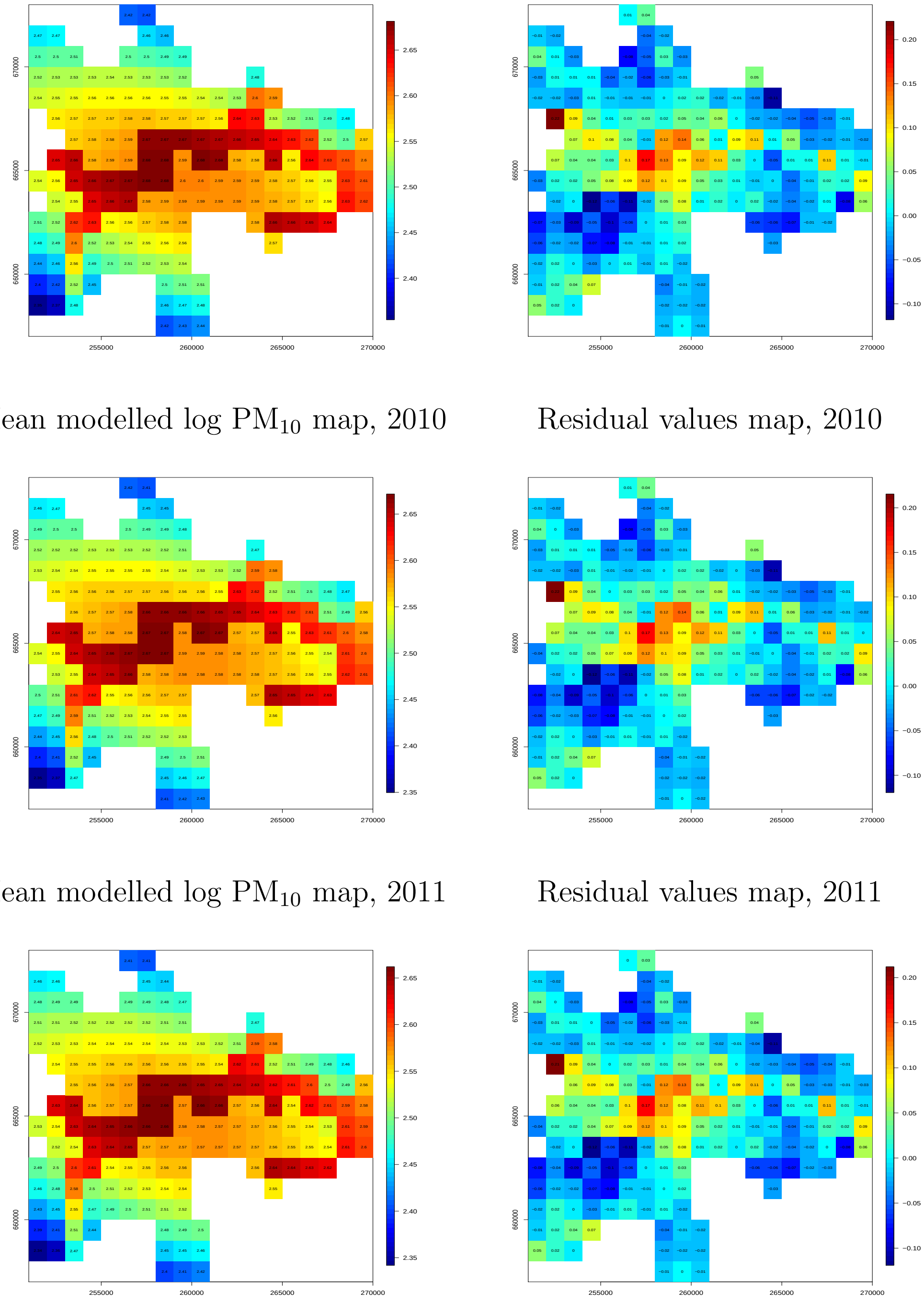
Table 3.4 displays the estimates and standard errors for each of the three years. The estimates for each of the years are very similar and have the same significant variables. Although the estimates appear similar for each of the years each of the estimates are decreasing very slightly with time. This could suggest a decreasing trend over time. The nugget and partial sill parameters remain constant while the range parameter increases only slightly.

Table 3.4: Table of Estimates for each year 2010 - 2012

	2010	2011	2012
	Est (St.Error)	Est (St.Error)	Est (St.Error)
$\widehat{\beta}_0$ (Intercept)	**2.369 (0.063)**	**2.360 (0.060)**	**2.351 (0.061)**
$\widehat{\beta}_1$ (Motorway)	**0.087 (0.009)**	**0.087 (0.009)**	**0.087 (0.008)**
$\widehat{\beta}_2$ (Easting)	**0.215 (0.135)**	**0.213 (0.134)**	**0.211 (0.131)**
$\widehat{\beta}_3$ (Northing)	**0.673 (0.211)**	**0.666 (0.210)**	**0.659 (0.208)**
$\widehat{\beta}_4$ (Easting^2)	-0.233 (0.154)	-0.232 (0.157)	-0.230 (0.158)
$\widehat{\beta}_5$ (Northing^2)	**-0.658 (0.175)**	**-0.650 (0.187)**	**-0.642 (0.185)**
$\widehat{\phi}^2$	0.0003	0.0003	0.0003
$\widehat{\sigma}^2$	0.004	0.004	0.004
$\widehat{\lambda}$	0.379	0.396	0.391

The standard errors and the residual values maps which can be found in Figure 3.7. The maps down the left hand side are the mean modelled log PM_{10} maps for each of the years and the maps down the right hand side are the residual values maps. It is apparent in the mean maps and the residual maps that each of the three years are very similarly spatially distributed therefore although there is a slight decrease with time, the spatial correlation structure could possible be assumed to be constant. Residuals reveal that most of the spatial trend has been modelled there remains two grid cells which have large residual values. It would appear that both of these grid cells which are coloured as red have been under estimated. Apart from these two grid cells there remains a slight trend in that the grid cells in the centre of the city have slightly higher residuals than the grid cells towards the outskirts but the difference is not huge. The mean map shows that those grids that lie along the motorway path have been estimated as having much higher PM_{10} levels than those not on the pathway. The binary motorway covariate, however, does not take into account that those locations that are spatially close to the grid cells with motorways would also be affected and so a smoothed function could have been more appropriate. Apart from the motorway path, the estimates are highest in the city centre and lowest in the outskirts.

This modelling process has provided an idea of how PM_{10} is distributed across time and that although the estimates decrease with time the spatial structure could be assumed constant. However, this modelling process has assumed each year to be independent and not considered the time series aspect. What has been learned in this modelling process could be used in conjunction with what was learned in the PM_{10} monitoring site modelling process to produce an indicator. The strength of this set of PM_{10} data was that it was able to explore the spatial aspect of PM_{10} across Glasgow but there were only three time points across the three years to work with. Whereas, the monitoring site data had data for daily time points across the three years but only at 11 monitoring sites. Combining what was learned from both of these analyses could produce a much more reliable indicator for Glasgow.



Mean modelled log PM_{10} map, 2010

Residual values map, 2010

Mean modelled log PM_{10} map, 2011

Residual values map, 2011

Mean modelled log PM_{10} map, 2012

Residual values map, 2012

Figure 3.7: Mean modelled log PM_{10} map and residual values map for each of the three years.

Chapter 4

Producing an air pollution indicator for Glasgow

The goal of this thesis is to produce an air pollution indicator for Glasgow. In this chapter the aim is to use statistical principles to produce air quality indicators that may be aggregated over time or space to give the flexibility to produce indicators at relevant spatial (e.g citywide) or temporal (e.g. yearly) scales. We previously explored levels of air pollution across Glasgow through the statistical modelling of monitoring site and gridded modelled PM_{10} data. In this chapter we start by reviewing additional literature relevant to producing an air quality index in Glasgow. We then construct some naive indicators for Glasgow, that fail to account for the spatial and temporal dependence in pollution. After criticising these indicators, we next consider a spatio-temporal model for the modelled PM_{10} data. From these model results we construct a regional air quality index for Glasgow, with an associated measure of uncertainty. We finish the chapter with a discussion of our results, and some future directions.

4.1 Constructing air quality indexes - a review of selected works

In order to gain an idea of how to construct an air quality indicator, we will review the techniques already used in a few selected works, chosen to indicate the different approaches being taken. These works will help identify how to overcome issues, such as how to choose appropriate space and time indexes.

A study by Stieb *et al.* (2008) proposes a new air quality health index which captures the additive effects of multiple pollutants and the relationship between air pollution and health. The analysis concludes that this approach is valid in allowing people to judge how likely they are to experience health effects day to day. A further study which combines the relationship between health and air pollution is by Cairncross *et al.* (2007). This article proposes an index which is based on the relative risk of increased mortality associated with common air pollutants. The index is constructed by assigning each of the pollutants an index value ranging from 1 to 10 which denotes the risk of exposure. To account for the simultaneous exposure to common pollutants the index is defined to be the sum of the normalised values of the individual indices for the pollutants. In theory, a given index value or given index values correspond to the mortality risk associated with the combined pollutants. In Kyrkilis *et al.* (2007), an attempt is made to combine the health effects of five common pollutants into an index which accounts for European standards. Our study in Glasgow considers health as a driving force for the indicator but does not attribute any part of the modelling process to the effects of the pollutant on health. Also, our study only considers one pollutant. If multiple pollutants were to be included a health weighting could be used to aggregate the pollutants according to the health risk.

As discussed in Chapter 1, Lee *et al.* (2011) propose an index based on geostatistical modelling which allows for uncertainty to be calculated at the spatial aggregation stage and therefore for the final indicator. The details of this have been explored in Chapter 1. From this article the inclusion of an uncertainty measure was deemed important for the overall understanding of the indicator itself. Similarly, later in this chapter we calculate an uncertainty measure for a temporally varying index of pollution for Glasgow, based on a spatio-temporal model for the modelled PM_{10} data.

Incorporating the lessons learned from reviewing the literature, we move onto a discussion of our spatio-temporal index for Glasgow. We could produce a summary index for each monitoring site, however this is not very representative of space or we could produce a global (over space) Glasgow figure from the modelled data but this had very limited time information. The index could only be compared with other cities if the same analysis and indicator construction was conducted for that city but it can compare the state of air pollution in Glasgow across time. One commonality across most of the literature is that there are three main components of an indicator: the pollutants, the time indexes, and the space indexes. Unlike most of the literature our spatial-temporal index incorporates only one pollutant, PM_{10}, and therefore the aggregation of multiple pollutants need not be considered. There are various time intervals an indicator can be produced at: daily, weekly, monthly or annual scales. The gridded modelled annual mean PM_{10} data which our index is based on has yearly time points which span three years. If the daily monitoring site PM_{10} data were to be incorporated into the spatio-temporal model then there would be more flexibility in the time index. An indicator could be produced at different spatial indexes including for a specific geostatistical location, a small region or for the whole of Glasgow. We can produce an index of air pollution at different spatial scales by aggregating over the gridded locations in the modelled PM_{10} data.

4.2 Producing naive air quality indexes

4.2.1 Daily Mean Monitoring Site PM_{10} Indicator Estimation Discussion

In order to produce an indicator for the daily mean monitoring site PM_{10} data the issue of missing values must be discussed. At some monitoring locations there is a large percentage of missing values, which could result in a biased indicator. However, with a large amount of missing data the interpolation of missing values can be a complex and time consuming problem. Due to the time constraints of this study an interpolation technique such as an expectation maximisation (E-M) algorithm was not employed. We will discuss an interpolation technique briefly in our further work section.

Averaging

In the situation where an interpolation technique cannot be employed, a crude indicator could be constructed by simply averaging over time and space. A simple indicator could easily be calculated by simply averaging each of the daily mean PM_{10} concentrations for all of the 11 sites for each year. This could be calculated using Equation (4.1) where y_{it} denotes the PM_{10} values for time $t = 1, \ldots, T$ and the sites $i = 1, \ldots, n$.

This method assumes that the monitoring sites accurately represent Glasgow's air pollution levels. There is only one rural monitoring station as most of the stations are clustered around the city centre, suggesting that this selection of monitoring sites are not a representative sample for all of Glasgow. In this case it could be argued that to compensate for the unrepresentative nature of the sites the sites outwith the city centre should be assigned a higher weighting. However, without more spatial information it would be difficult to assign a weight to each of the sites. This could follow the Equation (4.1) below, where the weighting, w_i, represents the weight given to each site i.

Our estimated index over sites and times with a weighting of w_i for each site is:

$$\hat{s} = \frac{\sum_i w_i \sum_t y_{it}}{IT}. \tag{4.1}$$

An estimated standard error (SE) for this index is given by

$$SE(\hat{s}) = \sqrt{\sum_i \sum_{i'} w_i w_{i'} \sum_t \sum_{t'} \text{cov}(y_{i,t}, y_{i',t'})}, \tag{4.2}$$

where the covariance between the daily values $\text{cov}(y_{i,t}, y_{i',t'})$ needs to be estimated. However, a naive estimate that ignores covariance and uses equal weights ($w_i = 1$) at each site is produced as an example. In this case our estimated index over sites and times is as follows:

$$\hat{s} = \frac{\sum_i \sum_t y_{it}}{IT}. \tag{4.3}$$

The estimated standard error for this index is given by

$$SE(\hat{s}) = \sqrt{\frac{\sum_i \sum_t \text{var}(y_{i,t})}{IT}}. \tag{4.4}$$

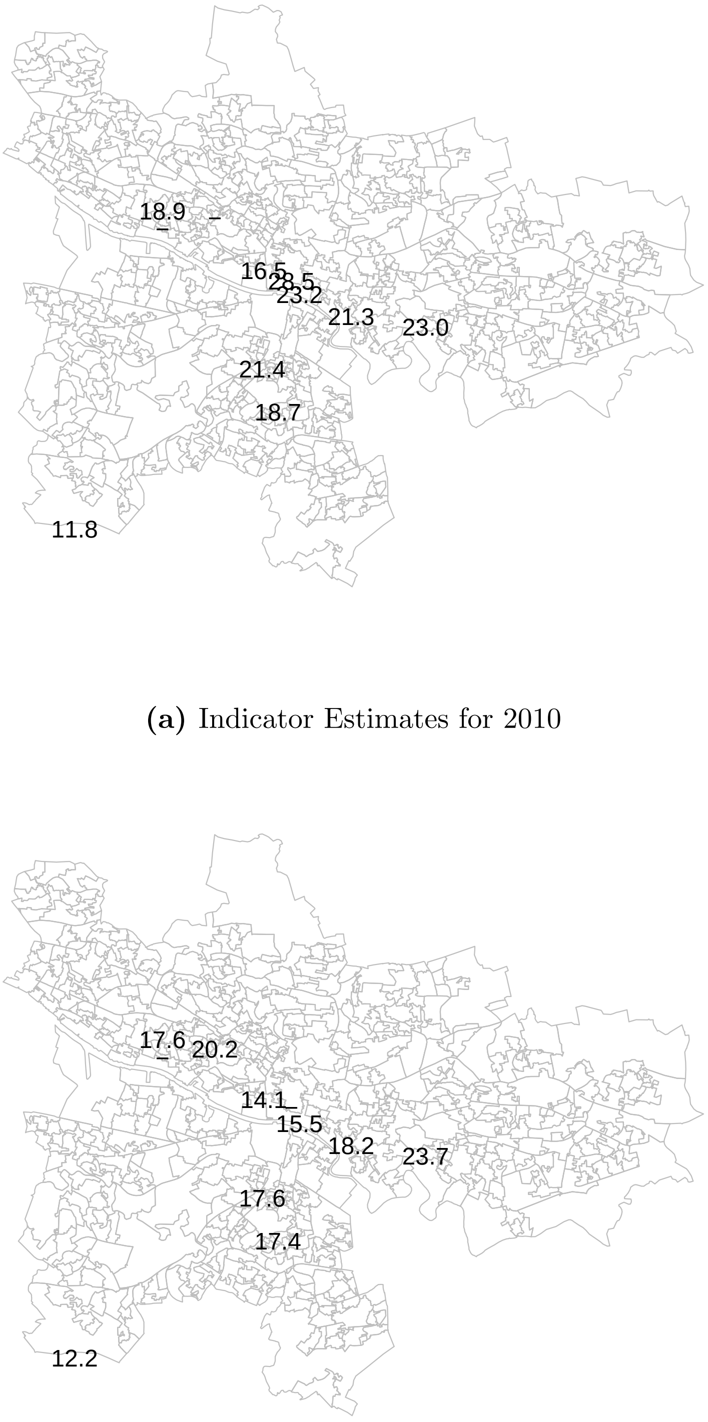
Daily Mean Monitoring Site PM_{10} Indicator

The above approach, Equations 4.3 and 4.4, produced indicator estimates which are summarised in Table 4.1. These values should be interpreted as a geographical indicator of air quality for 2010-2012 at 11 different sites across Glasgow. Although these indicators are based on values with a huge amount of missing data they can still provide a rough idea of the difference in air pollution between the years and across the sites. Table 4.1 indicates that air pollution is decreasing with time across a number of the sites. Overall the monitoring stations located in the city centre are higher than for those stations outwith that area. The only rural station, Waulkmillglen Reservoir has a noticeably smaller indicator estimate. This would suggest that PM_{10} is dependent on space as well as time and to produce a comprehensive indicator for Glasgow a spatial analysis should be conducted. The standard deviations are quite large in comparison to the indicator estimates which is unsurprising due to the volatile nature of the data. The naive indicator estimates are displayed over a map of Glasgow for the years 2010-2012 in Figure 4.1. This figure partially summarises the spatial distribution of pollution over Glasgow and confirms that overall the higher indicator estimates tend to be centered around the city centre with the lower estimates mostly found further outside of the city centre. However, the 2011 map in Figure 4.1b shows that the Anderston and Centre monitoring sites have much lower estimates compared to the surrounding monitoring sites. This may be due to the fact that they are not classified as roadside sites. This correlates well with the conclusions drawn from the annual mean gridded data which also showed that the higher values of PM_{10} were mostly found in the city centre and the lower values in the outskirts or the city.

Table 4.1: Naive Indicator for Glasgow - Temporal Model

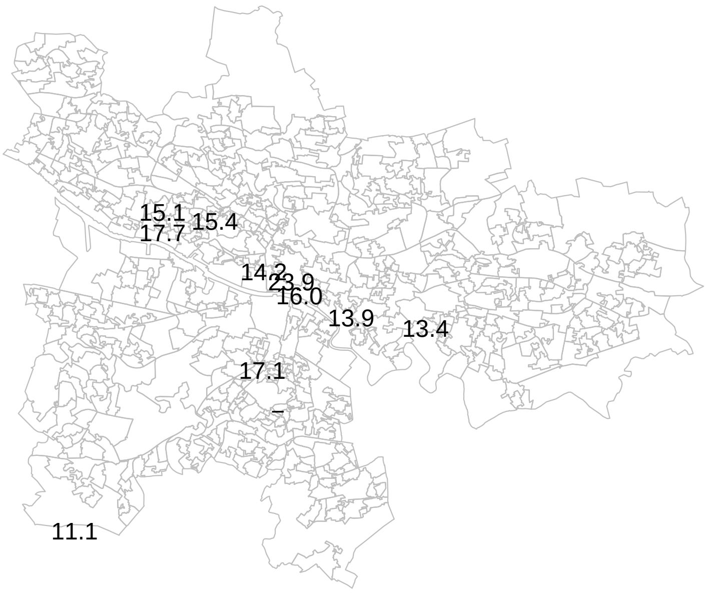
	2010	2011	2012
	Indicator (S.E)	Indicator (S.E)	Indicator (S.E)
Abercrombie St	21.347 (0.649)	18.146 (0.603)	13.866 (0.480)
Anderston	16.474 (0.560)	14.060 (0.614)	14.242 (0.507)
Battlefield Rd	18.735 (0.466)	17.379 (0.514)	-
Broomhill	18.880 (0.613)	17.570 (0.556)	15.055 (0.512)
Burgher St	-	20.189 (1.384)	15.435 (0.483)
Byres Rd	22.991(0.555)	23.703 (1.372)	13.400 (0.454)
Centre	23.221 (1.750)	15.534 (0.469)	15.959 (0.561)
Dumbarton Road	-	-	17.675 (0.474)
Kerbside	28.445 (0.783)	-	23.924 (0.787)
Nithsdale Rd	21.422 (0.739)	17.548 (0.505)	17.140 (0.621)
Waulkmillglen Res	11.796 (0.319)	12.136 (0.349)	11.105 (0.383)

The above indicator is easy to interpret and simple to construct. Air quality at each geographical location can be easily compared across the three years and across Glasgow. However, the large amount of missing data and the assumption that each of the sites were not correlated in space or in time



(a) Indicator Estimates for 2010

(b) Indicator Estimates for 2011



(c) Indicator Estimates for 2012

Figure 4.1: Indicator Estimates Displayed on Map of Glasgow, where - denotes that there was no data for this site.

are somewhat naive. Therefore, although the indicator is simple it has not accounted for the spatial or temporal variation or the huge amount of missing data. This would suggest that a more comprehensive indicator which takes into account time and space dependencies should be explored.

4.2.2 Gridded Modelled Annual Mean PM_{10} Data Indicator Estimation Discussion

Chapter 3 provides an insight into the spatial distribution of PM_{10} across Glasgow for the years 2010 - 2012. The discussion and analysis concludes that the distribution of PM_{10} depends heavily on spatial location and the binary motorway factor. Concentrating solely on the annual mean PM_{10} data, a simple average could be taken across Glasgow where there is no weighting, as each of the grid cells are assumed to be equal in weight. We are assuming spatial and temporal independence. The indicator is constructed by summing all of the grid squares for each year as follows:

$$\hat{c}_t = \frac{\sum_i y_t(s_i)}{I}. \tag{4.5}$$

An uncertainty estimate could also be calculated using the following equation:

$$\text{S.E}\hat{c}_t = \sqrt{\frac{1}{I-1}\sum_i \text{var}(y_t(s_i))}. \tag{4.6}$$

Where $\text{var}(y_t(s_i)) = \frac{1}{I}\sum_i((y_t(s_i)) - \bar{y}_t(s))$ and c_t denotes the crude indicator at time index $t = 1, 2, 3$ (corresponding to 2010, 2011, or 2012 respectively).. The yearly average, with uncertainty measures, are displayed in Figure 4.4 and Table 4.2. These yearly indicator values should be interpreted as the measure of air pollution in that year across the whole of Glasgow. The summaries display a steady decrease in PM_{10} concentrations for each year. However, this modelling process and, in turn, the indicators assume spatial and temporal independence. Air pollution as demonstrated in modelling the daily mean monitoring site data is temporally correlated and therefore by assuming that each year is independent is unreasonable. If these years were assumed not to be independent but to have some temporal correlation then the standard deviations around the indicators would likely increase, making the estimates less certain.

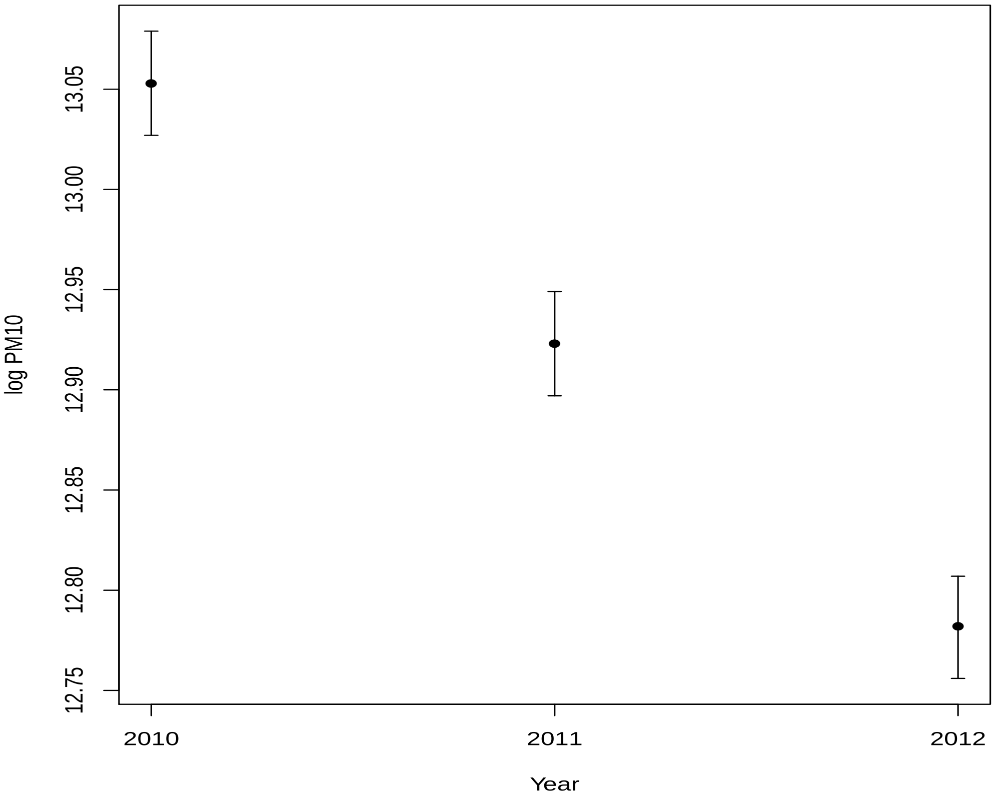


Figure 4.2: Crude indicator estimate with confidence interval

Table 4.2: Naive Indicator for Glasgow - Spatial Model

	Estimate	St. Error
2010	13.053	0.013
2011	12.923	0.013
2012	12.782	0.013

The above indicator gives a simple and effective indication of overall air quality in Glasgow with an uncertainty measure. However, the crude construction has not considered the temporal or spatial correlation between the years and therefore gives an indication of a possible bias in either the indicator estimates or the uncertainty estimates.

This thesis thus far has concentrated mainly on the spatial and temporal aspects of PM_{10} separately. Each of the crude estimator attempts have assumed either no spatial or temporal correlation. PM_{10} in Glasgow, however, is spatially correlated as well as temporally correlated. Therefore an indicator can only be calculated once the trends and patterns in PM_{10} in Glasgow has been modelled spatially and temporally.

4.3 A Spatio-Temporal Model for Modelled PM_{10}

The attempts at building an indicator for both of the sets of data have resulted in temporal and spatial dependence not being accounted for. In this section, after the discussion of the different space and time indexes which can be considered when constructing an indicator, a model and then an indicator which accounts for the spatial and temporal dependence within the gridded modelled annual mean PM_{10} data is discussed. The spatio-temporal model extends the geostatistical model discussed in Chapter 3. From Chapter 1, the modelled concentrations for the gridded annual mean modelled data are calculated using PM_{10} concentrations and meteorological data for the year 2010 and are then projected forward for years 2015, 2020, 2025 and 2030 with intermediate years being linearly interpolated. This means that years 2011 and 2012 are a linear product of 2010. In order to account for the temporal correlation, three models for each of the years do not have to be constructed, as was produced in the geostatistical model discussed in Chapter 3.

If we let $y(\underline{s}_i)$ denote the log PM_{10} value at grid box location $\underline{s}_i(1 = 1, \ldots, I)$, a linear trend or yearly changing mean will account for the correlation in time t where $t = 2010, 2011, 2012$. We assume that $\{y_t(\underline{s}_i\}$ is a Gaussian process with mean $\mu_t(\underline{s}_i) = E(y_t(\underline{s}_i))$ and covariance $C_{t,t'}(\underline{s}_i, \underline{s}_{i'}) = \text{cov}(y_t(\underline{s}_i), y_{t'}(\underline{s}_{i'}))$.

In our spatio-temporal model we assume that for a set of p-dimensional spatio-temporal covariates $\{\underline{X}_t(\underline{s}_i)\}$ that,

$$\mu_t(\underline{s}_i) = X_t^T(\underline{s}_i)\underline{\beta} \tag{4.7}$$

where β is an unknown p-dimensional coefficient vector that must be estimated from the data. Despite the fact that. From the analysis in the previous chapter we assume that the spatial distribution of PM$_{10}$ is constant over time, therefore, (for any $t \neq t'$, $C_{t,t'}(\underline{s}_i, \underline{s}_{i'}) = 0$ regardless of the spatial locations s_i and $s_{i'}$) we assume the same spatial covariance at each time point:

$$C_{t,t'}(\underline{s}_i, \underline{s}_{i'}) = \sigma^2 \exp\left(\frac{-\|s_i - s_{i'}\|}{\lambda}\right). \tag{4.8}$$

In the spatial covariance equation above, σ^2 denotes the spatial sill and λ the spatial range parameter, both of which need to be estimated from the PM$_{10}$ data.

This model can be written in matrix notation.Let $\underline{y}_t = (y_t(s_1), \ldots, y_t(s_i))^T$ denote the vector of log PM$_{10}$ values at year t. Let X_t denote the $I \times p$ design matrix of covariates with row $X_t(\underline{s}_i)$. Then we can say that $\{\underline{y}_t : t = 2010, 2011, 2012\}$ follow the multivariate normal distribution $N_I(X_t\underline{\beta}, \Sigma_{\sigma^2,\theta})$ where $\Sigma_{\sigma^2,\theta}$ is an $I \times I$ covariance matrix with the (i, i') element $C_{t,t'}(\underline{S}_i, \underline{S}_{i'})$.

4.4 Parameter Estimation

In our spatio-temporal model we need to estimate the coefficient vector β and spatial parameters σ^2 and λ. With $\underline{y}_t = (\underline{y}_{2010}, \underline{y}_{2011}, \underline{y}_{2012})$, the log likelihood function for the parameters is as follows,

$$
\begin{aligned}
l(\underline{\beta}, \sigma^2, \lambda|y) = & \frac{-3I}{2}\log(2\pi) - \frac{3I}{2}\log\sigma^2 - \frac{3}{2}\log\det R_\lambda \\
& - \frac{1}{2\sigma^2}\sum_{t=2010}^{2012}(\underline{y}_t - \underline{X}_t\beta)^T R_\lambda^{-1}(\underline{y}_t - \underline{X}_t\beta),
\end{aligned}
\tag{4.9}
$$

where $R_\lambda = \frac{\sum_{\sigma^2,\lambda}}{\sigma^2}$ is the spatial correlation matrix. The derivative of the log likelihood with respect to σ^2 is

$$
\frac{dl}{d\sigma^2} = -\frac{3I}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{t=2010}^{2012}(\underline{y}_t - X_t\underline{\beta})^T R_\lambda^{-1}(\underline{y}_t - X_t\underline{\beta}), \tag{4.10}
$$

which when setting equal to zero and solving for $\widehat{\sigma}^2$, yields the ML estimates of σ^2:

$$
\widehat{\sigma}^2 = \frac{\sum_{t=2010}^{2012}(\underline{y}_t - X_t\widehat{\underline{\beta}})R_{\widehat{\lambda}}^{-1}(\underline{y}_t - X_t\widehat{\underline{\beta}})}{3I}. \tag{4.11}
$$

This estimate is written in terms of the ML estimates for $\underline{\beta}$, $\widehat{\underline{\beta}}$ and for λ, $\widehat{\lambda}$ say. By independence over the years the ML estimate of β is

$$
\widehat{\underline{\beta}} = (\sum_{t=2010}^{2012} X_t R_\lambda^{-1} X_t)^{-1}(\sum_{t=2010}^{2012} X_t R_\lambda^{-1} Y_t). \tag{4.12}
$$

Whereas, the ML estimate of λ, $\widehat{\lambda}$, is solved by minimizing the log likelihood with respect to λ when we plug in $\widehat{\beta}$ and $\widehat{\sigma}^2$ - we minimise numerically using the Nelder-Mead(1965) algorithm.

4.5 Estimating the Spatio-Temporal Model Parameters

Using the methods described in the previous section, the regression parameters, $\underline{\beta} = (\beta_0, \ldots, \beta_m)$, and the spatial model parameters θ and σ^2 were estimated using ML; table 4.3 displays the ML estimates and standard errors. Each of the β terms are all significant in this model with relatively small standard errors. Comparing to the spatial model in Chapter 3, the intercept term is similar however each of the other regression parameter estimates are quite different. This model takes into account the temporal correlation using a yearly changing mean whereas the model in Chapter 3 assumed independence between the three years.

Table 4.3: Estimates and Standard Errors for Spatio-Temporal Model

	Estimate	St. Error
$\widehat{\beta_0}$ (Intercept)	**2.250**	**0.012**
$\widehat{\beta_1}$ (Eastings)	**0.483**	**0.035**
$\widehat{\beta_2}$ (Northings)	**1.072**	**0.043**
$\widehat{\beta_3}$ (Eastings^2)	**-0.352**	**0.029**
$\widehat{\beta_4}$ (Northings^2)	**-0.907**	**0.039**
$\widehat{\beta_5}$ (Eastings * Northings)	**-0.400**	**0.045**
$\widehat{\beta_6}$ (motorway)	**0.106**	**0.005**
$\widehat{\beta_7}$ (2011)	**-0.010**	**0.005**
$\widehat{\beta_8}$ (2012)	**-0.021**	**0.005**
$\widehat{\sigma^2}$	0.003	-
$\widehat{\theta}$	24.194	-

Each of the coefficients are significant in the model with values β_1 to β_5 describing the effect of space while the β_6 estimate denotes the effect of the motorway covariate.

The β_1 to β_5 estimate the effects of space in the model. Each of these estimates are significant and give us an idea of the relationship between PM$_{10}$ and space. For example, β_1 shows that by moving one grid cell east, on average log PM$_{10}$ increases by 0.483 and β_2 shows that by moving one grid cell north, on average log PM$_{10}$ increases by 1.072. The β_6 coefficient shows that on average log PM$_{10}$ increases by 0.106 if the grid cell happens to contain a motorway in comparison to the baseline - no motorway. The residual coefficients β_7 and β_8 show that compared to a baseline year of

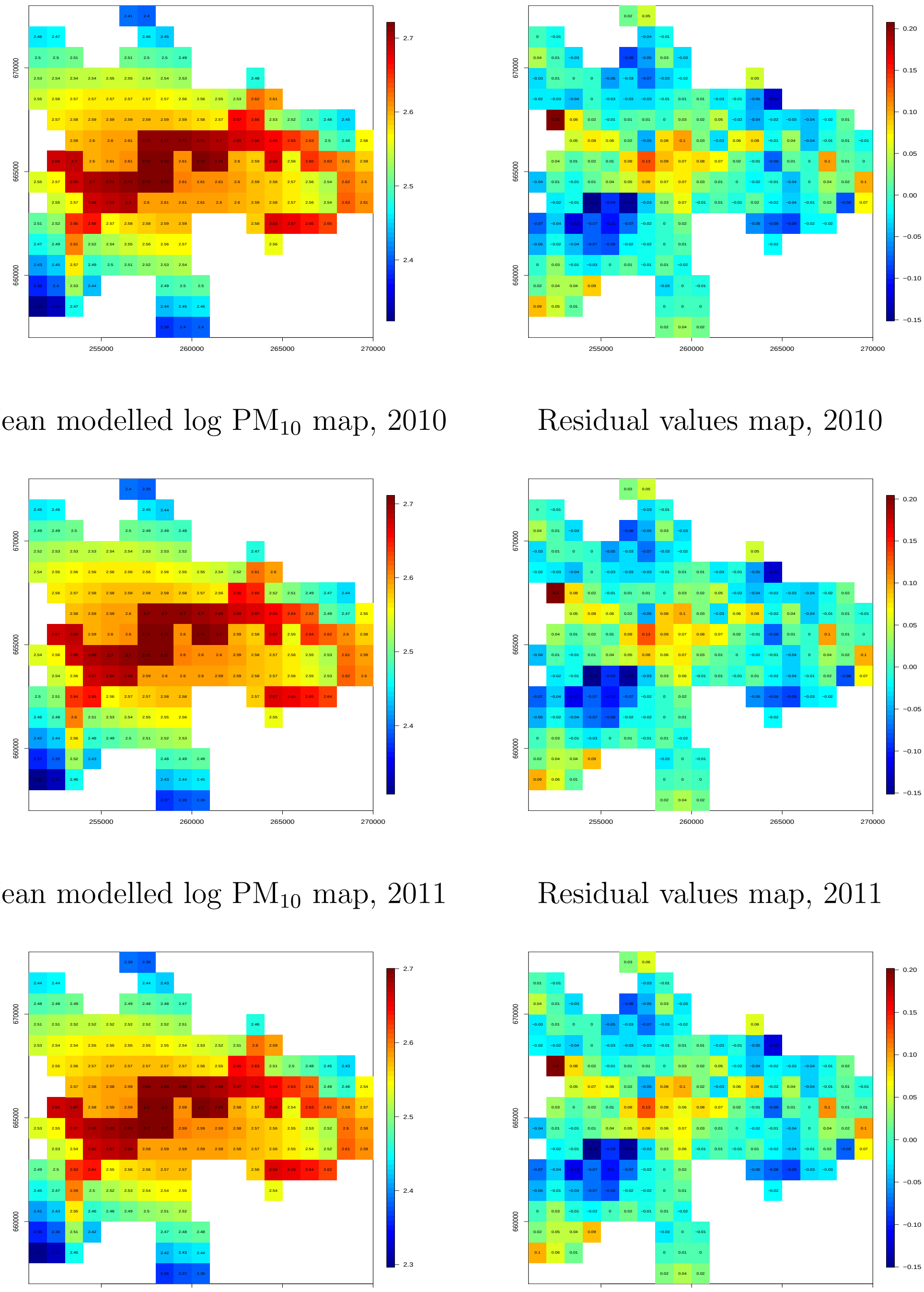
2010 that on average log PM_{10} decreases by -0.01 in 2011 and by -0.021 in 2012. The mean and residual values maps are displayed in Figure 4.3. These maps show that the different years are identically spatially distributed but with the means decreasing slightly with time. This could be due to a reduction in emissions. As discussed earlier in this chapter, the 2010 data was collected and the 2011 and 2012 years were linearly interpolated leaving almost identically distributed data. This does not give us a clear indication of what the annual mean for 2011 and 2012 was, but provides us only with a predicted case.

4.6 Building a Yearly Index of Air Pollution for Glasgow

The spatio-temporal model accounts for both the spatial and temporal correlation in the gridded modelled annual mean PM_{10} data while incorporating interesting covariate effects such as the location of motorways. This should improve on the indicators which were calculated earlier. The following plot, Figure 4.2, and Table 4.4 show the indicator estimates for the whole of Glasgow and the standard errors for each of the three years back transformed to their original scale. This should be interpreted as an indicator which estimates air pollution in Glasgow across the three years. Each of the years are very similar, with 2010 having the highest indicator estimate and 2012 with the lowest indicator estimate and each of the estimates have equally small standard errors.

Table 4.4: Naive Indicator for Glasgow - Spatio-Temporal Model

	Estimate	St. Error
2010	13.053	0.065
2011	12.923	0.065
2012	12.782	0.065



Mean modelled log PM_{10} map, 2010

Residual values map, 2010

Mean modelled log PM_{10} map, 2011

Residual values map, 2011

Mean modelled log PM_{10} map, 2012

Residual values map, 2012

Figure 4.3: 2010-2012 estimated means and residual values map for spatio-temporal model.

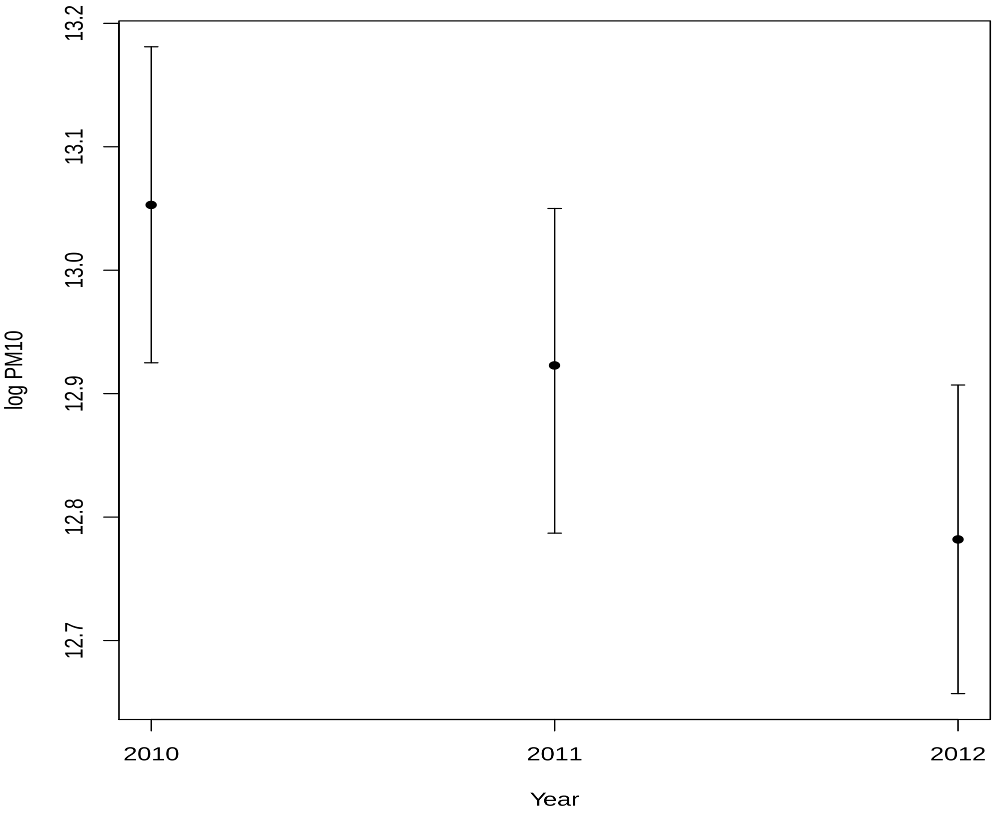


Figure 4.4: Crude indicator estimate with confidence interval

4.7 Discussion

The main aim of this research was to produce an indicator based on PM_{10} for Glasgow to be used as a convenient way to asses Glasgow's PM_{10} levels. By exploring simple indicators for each of the data sets - daily mean monitoring site and annual gridded PM_{10} levels throughout this chapter it became apparent that PM_{10} levels in Glasgow are too complex for a simple average. By averaging over time for the daily mean monitoring site PM_{10} levels the spatial distribution of the site was being overlooked. Similarly, by averaging over space for the annual gridded PM_{10} levels the annual values were assumed to be temporally independent. This motivated the study of a spatio-temporal model. The review of selected works raised the question of which time and space indexes should be used. Again, the daily mean monitoring site PM_{10} level data are rich over time at random locations. Whereas, the annual gridded set is data rich over space with 1 x 1 km modelled estimates but these exist only on an annual scale. Ideally, we would have combined the two sets of complimentary data and modelled these datasets over time and space. However, with the short time period allocated for this research this was not feasible. Alternatively, with the short time scale we used the information gained from modelling the daily mean monitoring data and applied this to the annual gridded data. By applying this knowledge to the annual gridded values a spatio-temporal model and in return and indicator could be produced.

The spatio-temporal model was based on the modelling in Chapter 3 but with the inclusion of a linearly decreasing yearly term. The enabled the model to not only model the spatial distribution of PM_{10} but allowed for the estimates to differ across time. This resulted in an indicator, with uncertainty, which states that the annual average log PM_{10} value for 2010 sits at 13.053 and linearly decreases with over the three years. These indicator estimates have an uncertainty measure of 0.065.

Chapter 5

Conclusions and Further work

5.1 Conclusions

The main aim of this research was to explore several statistical approaches to produce an air pollution indicator for Glasgow using routinely available PM_{10} data. The study has conducted an initial investigation using two datasets into how PM_{10} is distributed across time and space. The first of these datasets is made up of daily mean values at 11 different monitoring sites across Glasgow for the years 2010 to 2012. The second set of data contains gridded modelled annual mean values for each 1×1 km grid cell across Glasgow, also for the years 2010 to 2012.

Chapter 2 explored the trends and seasonality found within the daily mean PM_{10} data at each of the 11 sites for three years. Initially, the most striking feature of this set of data is the large amount of missing values across a number of the sites, some of which may be missing at random. The maximum percentage of missing values is 72% with one or two of the sites each year not operational. In this study we continued exploring and drawing conclusions from this data by ignoring the possible effect of these missing values. However, looking back I would have liked to interpolate these missing values to improve the reliability and robustness of the conclusions drawn from this set of data. The exploratory conclusions found that there was likely to be some seasonality within each of the years and some relationship between PM_{10} and the meteorological variables, temperature and humidity. This exploratory analysis, however, was conducted assuming that there is no temporal correlation in the PM_{10} data. A series of models with various combinations including harmonic regression terms to model the daily effect, a day for the week factor, the meteorological variables, humidity and temperature were fit. The temporal autocorrelation was assessed and an autoregressive, AR(1), process was incorporated into the models. The three models that were discussed in Section 2.4 are as follows,

$$\begin{aligned} y_t = {} & \beta_0 + \beta_1 \cos(2\pi t/365) + \beta_2 \sin(2\pi t/365) + \beta_3 (\text{DayofWeek})_t \\ & + \beta_4 (\text{Humidity})_t + \beta_5 (\text{Temperature})_t + \varepsilon_t, \end{aligned} \tag{5.1}$$

$$\begin{aligned} y_t = {} & \beta_0 + \beta_1 \cos(2\pi t/365) + \beta_2 \sin(2\pi t/365) + \beta_3 (\text{DayofWeek})_t \\ & + \beta_4 (\text{Humidity})_t + \varepsilon_t, \end{aligned} \tag{5.2}$$

$$y_t = \beta_0 + \beta_1 \cos(2\pi t/365) + \beta_2 \sin(2\pi t/365) + \beta_3 (\text{DayofWeek})_t + \varepsilon_t, \tag{5.3}$$

where $t = 1, \ldots, 365$ and ε_t is a mean zero AR(1) time series process. Model 4 was deemed not to be the most suitable as the temperature variable was not significant at any of the sites. Looking at the AIC values and the exploratory fitted line plots it was apparent that Model 5 accounted for more of the variability in the data than Model 6. The estimates, standard errors, Ljung-Box statistics and AIC values for Model 5 for each of the 11 sites across the 3 years is summarised in the three tables 2.11, 2.12 and 2.13.

There is a considerable difference in the significant covariates across the 11 sites which suggests that PM_{10} varies across space. This analysis showed us that PM_{10} has not only a seasonal pattern in time but also that there is a spatial variation in pollution.

The next analysis focussed on the gridded annual mean PM_{10} data analysis, in which the spatial distribution of PM_{10} was explored across the same three years, from 2010 to 2012. In this dataset, the annual mean concentrations were modelled for 2010 and then linearly projected for the following years. Looking back, our analysis assumed a constant spatial distribution over the three years with a yearly linear term, which does not accurately represent the ever-changing air pollution environment. At this point an alternative data source could have been explored to give us a suitable data set in which to analyse the spatial distribution across the years. Image plots of each year, assuming no temporal correlation, were produced to explore the distribution of the annual mean across Glasgow. In hindsight I would not have explored the PM_{10} data for each year individually but I would have assumed a constant spatial distribution and modelled the yearly trend linearly. The image plots showed that for each of the three years the highest annual mean grid cell values were concentrated mainly in the city centre, north of the River Clyde. At this point, a binary motorway covariate was used to explore the distribution of the grid cells with the highest PM_{10} concentrations. This suggested that as well as space, the presence of a motorway could have an effect on PM_{10}. A binary motorway covariate, however, does not reflect the true effect of a motorway on PM_{10} levels. The effect of a motorway on PM_{10} levels is not binary, but a smooth function. The binary variable does not account for raised pollution levels if a motorway is not within the 1km × 1km grid square. If I were to model this data again I would have smoothed this function as it underestimates the effects of a motorway. Two models, found in Equations 3.14 and 3.15 were fit. These models were firstly fit as linear regression models using OLS, assuming independence, as an exploratory measure. These models were discussed in Section 3.4 and are as follows,

$$\begin{aligned}\log y(s_i) = \beta_0 + \beta_1 \text{eastings}(s_i) + \beta_2 \text{northings}(s_i) \\ + \beta_3 \text{motorway(factor)}(s_i) + \varepsilon(s_i)\end{aligned} \tag{5.4}$$

$$\begin{aligned}\log y(s_i) = \beta_0 + \beta_1 \text{eastings}(s_i) + \beta_2 \text{northings}(s_i) + \beta_3 \text{eastings}^2(s_i) \\ + \beta_2 \text{northings}^2(s_i) + \beta_5 \text{motorway(factor)}(s_i) + \varepsilon(s_i)\end{aligned} \tag{5.5}$$

where $y(s_i)$ corresponds to the PM$_{10}$ value at each spatial location $i = , \ldots, I$, $x(s_i)$ is the design matrix which is made up of different covariates, $\underline{\beta}$ which corresponds to the regression coefficients and $\varepsilon(s_i)$ denotes the residuals which are assumed to follow a normal distribution with mean zero. Using diagnostic plots, it was apparent that Model 2 fits the data best. The correlation structure was assessed using various variograms. The correlation structure of the data was assumed to follow an exponential covariance function. ML and REML methods were used to estimate the model parameters. The table of estimates for each year shows that each of the covariates are significant, except the eastings and northings interaction, confirming that PM$_{10}$ is dependent on space and the presence of a motorway. The intercept estimates decreased over time. This modelling did not, however, take into account any temporal correlation.

A naive indicator of air (PM$_{10}$) quality was constructed for each of the 11 monitoring sites for each of the three years using the daily mean monitoring site data. This indicator was constructed by taking a simple average of the time series data and calculating the standard error, assuming independence. As discussed, without adequate interpolation and accounting for the spatial and temporal dependence, the indicator results and conclusions are biased.

The naive indicators, overall, showed a decreasing trend across time and a varied distribution across space.

An annual naive indicator was constructed using the annual mean gridded data for the whole of Glasgow for the three years. As discussed, without accounting for temporal dependence the indicator and uncertainty estimates would not reflect the true value. The indicator does, however, decrease with time which confirms the reduction in air pollution over the years studied.

Combining the knowledge from modelling both of the sets of data it would appear that PM_{10} data is both spatially and temporally correlated. In order to account for both of these correlations, a spatio-temporal model was constructed. A number of selected works were reviewed to gain an idea of the different ways air pollution indicators are constructed. The spatio-temporal model simply extended the geostatistical model which was fit to the annual mean gridded data. The model discussed in Section 4.3 explains that a linear trend in time would account for the temporal correlation in addition to the spatial correlation. Maximum likelihood estimation via general-purpose optimization was used in order to estimate the spatio-temporal model parameters. This model was then used to build, using a simple average, a yearly index of air pollution for Glasgow with an uncertainty measure with the spatially - and temporally- varying covariates and residual spatial dependence accounted for.

The two naive air pollution indicators and the spatio-temporal indicators overall follow a similar trend over time. All of the naive indicators for the temporal model decrease with time with a few exceptions - Anderston, Byres Road, Centre and Waulkmillglen Reservoir. These exceptions, however, could be due to the large amount of missing data at some of the sites possibly skewing the results. The spatio-temporal model indicator estimates range from 13.1$\mu g m^3$ in 2010 to 12.8$\mu g m^3$ in 2012, whereas the indicator estimates for the daily mean data range from between 11.1$\mu g m^3$ to 28.4$\mu g m^3$. The range of indicator values for the daily mean data is much larger than that of the annual mean gridded data although they do overlap. The large amount of missing values in the daily mean data may allow us to question the consistency, although there is also the fact that we are comparing indicators constructed from daily and yearly values. Moving on from this study, the spatio-temporal indicator estimate could be an initial indicator. Bearing in mind that the daily mean data naive indicator estimate suggests that the annual mean indicator may be underestimating the true value of air pollution in Glasgow.

5.2 Further Work

In addition to the literature discussed in the previous section, other literature can provide a direction for further work. A few papers are referenced with regards to further work to give an example of how indicators can be used on a much larger, global scale and in dealing with a large number of pollutants. In addition to improving the PM_{10} indicator, different air pollutants should be considered. Looking back at the literature review most of the existing air pollution indexes have more than one air pollutant - with some studies using the five most common pollutants. The more pollutants included and thus more data could improve the reliability and robustness of the indicator.

The issue of subjectivity is only touched upon in this study, however Sowlat *et al.* (2011) discuss an air quality index which is produced using fuzzy logic. The article states that conventional methods for an air quality assessment are inaccurate due to the large number of parameters which contribute to air pollution. The proposed fuzzy index system appears to be more reliable when dealing with such a large number of contributing factors, although in this study we look only at one pollutant, the fuzzy based index could be an appropriate way to combine a large number of pollutants.

Zujic *et al.* (2009), discuss individual pollutant indices which could be used to compare pollutants in an area while reflecting population exposure. To demonstrate this an existing index scale for Belgrade metropolitan area was modified and extended to include elements of population done using weighting according to population densities. This method does two things - it makes inter-pollutant comparisons possible and aids in assessing the overall exposure of pollutants to the whole city population. Our study does not include population densities in the index nor does it include more than one pollutant, however, with more time these could be introduced using this study as a starting point.

Most recently, Hsu *et al.* (2013) states that in order to construct the "next generation" of air quality indicators, the needs of stakeholders and policymakers must be at the forefront of the discussion. Firstly, the choice of air pollutants to include in the model must consider available data and the impact to the health of humans and the environment. Instead of addressing the pollutants separately, they argue that there is a need for improved measurement and monitoring of pollutants as well as the impact that the pollutants have. By better understanding these factors we can produce better indicators. This is an interesting discussion and one that would need to be had throughout the world if we are to strive for a global indicator. A global indicator would reduce bias when comparing air quality across the world. Currently, countries have different ways of monitoring air pollution and so if there were to be a global indicator a standard air quality monitoring guide would have to be introduced across the world with one overall governing body. This would be hugely costly in terms of money and time and each country across the world would have to agree and contribute. Although, in theory, it would make tracking air pollution over time and space transparent and simple it is unrealistic.

One of the main issues in this study was the quality of the data. The large amount of missing data in the daily mean dataset meant that no reliable conclusions could be drawn from the analysis of that set of data. The fact that the daily mean data is not representative of the whole of Glasgow means that with most of the sites located in the city centre, any Glasgow averages would have been skewed. On the other hand, the gridded data was spatially representative but did not contain data to explore the seasonal trends within the years. The daily mean data used actual values whereas the gridded data used modelled values. Comparing indicators that are the result of two very different datasets could be the cause of the disparity between the indicator values. One interesting challenge going forward would be to merge these datasets with very different properties. Data fusion using data with different properties could enhance the an indicator my having temporally and spatially representative data.

Developments would include an attempt to collect a full set of multiple pollutant data which is measured on a regular basis at regular spatial intervals across Glasgow for a reasonable number of years using a different source or data fusion. This would enable the true temporal and spatial distribution of pollution to be explored. An air pollution indicator (and uncertainty measure) could then be constructed from a spatio-temporal model that accounts for the interesting covariate effects (e.g., traffic), as well as the spatial and temporal variation that explains pollution in Glasgow.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on 19*, 716 – 723.

Barmpadimos, I., Hueglin, C., Keller, J., Henne, A., and Prevot, A. (2011). Infuence of meteorology on PM10 trends and variability in Switzerland from 1991 to 2008. *Atmospheric Chemistry and Physics 11*, 1813–1835.

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

Bell, M., Davis, D., and 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.

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

Bruno, F. and Cocchi, D. (2002). A unifed strategy for building simple air quality indices. *Environmetrics 13*, 243–261.

Cairncross, E., John, J., and Zunckel, M. (2007). A novel air pollution index based on the relative risk of daily mortality associated with short-term exposure to common air pollutants. *Atmospheric Environment 41*, 8442–8454.

Cherchye, L., Moesen, W., Rogge, N., Puyenbroeck, T., Saisana, M., Saltelli, A., Liska, R., and Tarantola, S. (2007). Creating composite indicators with DEA and robustness analysis: the case of the technology achievement index. *Journal of the Operational Research Society 59*, 239–251.

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

Defra (2007). The Air Quality Strategy for England, Scotland, Wales and Northern Ireland.

Defra (2013a). Air modelling for Defra. http://uk-air.defra.gov.uk/research/air-quality-modelling?view=modelling.

Defra (2013b). Effects of air pollution. http://uk-air.defra.gov.uk/air-pollution/effects.

Diggle, P., Menzes, R., and Su, T. (2010). Geostatistical inference under preferential sampling. *Applied Statistics 58*, 191–232.

Diggle, P. and Ribeiro, P. (2007). *Model-based Geostatistics*. Springer.

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.

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.

Durbin, J. and Koopman, S. (2001). *Time Series Analysis by State Space Methods*. Oxford Statistical Science Series. Oxford University Press.

EIONET (2013). European Environment Information and Observation Network. http://www.eionet.europa.eu/.

European Environment Agency (2012). European environment agency. http://www.eea.europa.eu/.

European Parliament Council (2002, July). Sixth Environmental Action Programme.

Gardner, G., Haney, A., and Philips, G. (1980). Exact Maximum Likelihood Estimation of Autoregressive-Moving Average Models Average Models by Means of Kalman Filtering. *Applied Statistics 29*, 311–322.

Hsu, A., Reuben, A., Shindell, D., de Sherbinin, A., and Levy, M. (2013). Toward the next generation of air quality monitoring indicators. *Atmospheric Environment 80*, 651–570.

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

Kyrkilis, G., Chaloulakou, A., and Kassomenos, P. (2007). Development of an aggregate Air Quality Index for an urban Mediterranean agglomeration: Relation to potential health effects. *Environment International 33*, 670–6.

Lee, D., Ferguson, C., and Scott, M. (2011). Constructing representative air quality indicators with measures of uncertainty. *Journal of the Royal Statistical Society 174*, 109–126.

Ljung, G. and Box, G. (1978). On a Measure of a Lack of Fit in Time Series Models. *Biometrika 65*, 297–303.

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

Office of the Chief Statistician (2009). Scottish Index of Multiple Deprivation 2009 Technical Report. http://www.scotland.gov.uk/Publications/2009/10/28104046/0.

Patterson, D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika 58*, 545–554.

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.

Ribeiro, P. and Diggle, P. (2013). Likelihood Based Parameter Estimation for Gaussian Random Fields. http://hosho.ees.hokudai.ac.jp/~kubo/Rdoc/library/geoR/html/likfit.html.

Richardson, E., Mitchell, R., Shortt, N., Pearce, J., and Dawson, T. (2010). Developing summary measures of health-related multiple physical environmental deprivation for epidemiological research. *Environment and Planning A 42*, 1650–1668.

Scottish Air Quality (2012a). Air Quality in Scotland. http://www.scottishairquality.co.uk/.

Scottish Air Quality (2012b). Effects of air pollution. http://uk-air.defra.gov.uk/air-pollution/effects.

Sowlat, M., Gharibi, H., Yunesian, M., Mahmoudi, M., and Lotfi, S. (2011). A novel, fuzzy-based air quality index (FAQI) for air quality assessment. *Atmospheric Environment 45*, 2050–9.

Stieb, D., Burnett, R., Smith-Doiron, M., Brion, O., Shin, H., and Economou, V. (2008). A new multipollutant, no-threshold air quality health index based on short-term associations observed in daily time-series analyses. *Journal of the Air and Waste Management Association 58*, 435–50.

Tarantola, S. and Saltelli, A. (2008). Composite indicators: the art of mixing apples and oranges. http://publications.jrc.ec.europa.eu/repository/handle/111111111/11612.

The Oxford English Dictionary (2012). *The Oxford English Dictionary*, Volume Eleventh. Great Clarendon Street, Oxford, OX2 6DP: Oxford University Press.

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_Judgment.pdf.

Weather Underground Network (2012). http://www.wunderground.com/.

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

Yusof, N. F. F. M., Ghazali, N. A., Ramli, N. A., Yahaya, A. S., Sansuddin, N., and Madhoun, W. A. (2008). Correlation of PM10 Concentration and Weather Parameters in Conjunction with Haze Event in Seberang Perai, Penang. *International Conference on Construction and Building Technology 2008*, 211–220.

Zujic, A., Radak, B., Filipovic, A., and Markovic, D. (2009). Extending the use of air quality indices to reflect effective population exposure. *Environmental Monitoring and Assessment 156*, 539–49.