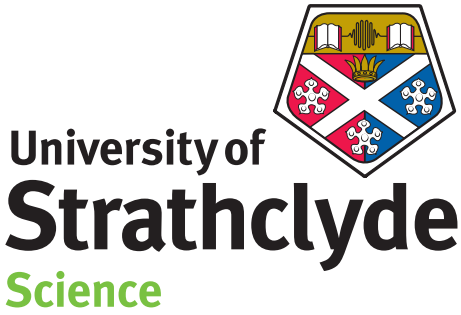



# Modelling Air Pollution in Scotland

Oyebamiji Oluwole Kehinde

Requirements for the degree of
Master of Philosophy

Department of Mathematics and Statistics
October 2010

© The copyright of this thesis belongs to the author under the terms of the United Kingdom Copyright Acts as qualified by University of Strathclyde Regulation 3.49. Due acknowledgement must always be made of the use of any material contained in, or derived from, this thesis.

# Acknowledgements

I am most grateful to my supervisors Dr. Alison Gray and Prof. Chris Robertson for their guidance. They provided an invaluable research and written directions that made this thesis exciting to write. They are highly acknowledged.
I am also indebted to the entire staff of Mathematics and Statistics Department especially services provided by the office staff members. I want to say thank you. Special mention should be made of the unquantifiable support and assistance rendered by my friends Akinyele Kazeem, Orelesi Emmanuel, Franck Kalala Mutombo, Dehinwa Adebayo, Ilesanmi Oluwaseyi, Benson Tinuola, Fatai Musibau, Adeosun Adebowale. I appreciate the forum created by our friendship and the benefit I have derived from it. I love you all.
Furthermore, my appreciation will remain incomplete if I do not mention my darling Oluwakemi Abigael, your deep love, tender care, cute attention and unparralled desire to make me happy during the course of this study became solid pillar upon which the success rests.
Lastly, the moral support, assistance, advice and prayer of Mr. Gbenga Williams, Pastor Festus, Pastor Olatunji Olanrewaju, Evang. Dejo, Pastor Segun, and Alfa Isah. May the lord reward you accordingly.

# Contents

# List of Tables

# List of Figures

# Abstract

The aim of the current work is to model sulphur dioxide ($S0_2$) levels in Scotland and relate it to health. The first part of the analysis involves basic descriptive characteristics of the data, where the structure of the dataset was explored, the missing data pattern and mapping of the spatial locations of the monitoring stations. The next part of the modelling involves time series modelling of the data in which several imputation techniques were explored for missing data. There are missing data and not all stations have measurements for all the years which necessitated us to use three different imputation techniques.

We examined temporal correlation, time series decomposition into long-term trend, seasonal and cyclical components as well as random fluctuations. The temporal modelling techniques of $AR$, $ARMA$ and $ARIMA$ models are also explored. We also investigated the spatial distribution and variation of $S0_2$ levels across Scotland using both the Bayesian and ordinary kriging techniques. Kriging provide optimal spatial prediction of $S0_2$ levels across Scotland.

Further modelling involved the use of spatial generalized additive models to incorporate the data attributes of both space and time. Univariate spatial smoothers are applied to the year and month of the observations, while a bivariate smoother was applied to the spatial location of the data. Gam technique is used in this thesis as both predictive and exploratory tools. We explored various basis functions and basis dimension.

Lastly, we observe that there is a variation in the $S0_2$ levels both within the year (seasonal variation) and across the years. Most of the stations are concentrated in Central Scotland. There is an evidence of temporal correlation as suggested by autocorrelation function (ACF). Bayesian model performs better than ordinary kriging by producing lower variance and better prediction than ordinary kriging which could be due to incorporation of uncertainty in the trend and covariance functions. A low spatial variation is observed in Central Scotland. The joint effect of the predictor variables from the best model explained just 65.4% of the variance of the dependent variable.

# Chapter 1

# Air pollution, its effects on health, and description of data

### 1.0.1 Overview

This thesis investigates modelling air pollution levels over Scotland which may have an impact on health. The methodology involves investigating both the temporal and spatial pattern of sulphur dioxide ($SO_2$) data separately before going on to advanced spatio-temporal modelling in which a generalized additive model is used to incorporate both the spatial and temporal attributes of the $S0_2$ data simultaneously.
The main objective is to obtain a statistical model to predict and estimate $SO_2$ concentrations across Scotland spatially and temporally simultaneously in which gam technique is adopted. Final results suggested that the there is a variation in $SO_2$ levels both within and across the years. There is also presence of both temporal and spatial correlation in $SO_2$ data. For the spatial analysis, Bayesian kriging performs better than ordinary kriging with better predictions and low kriging variance. In the spatial gam, the joint effect of the predictor variables from the best model explained more than 39.3% of the variance of the dependent variable, which corresponds to the model with REML estimation. High predictions for $SO_2$ are observed in Central Scotland, in accordance with the results of the Bayesian kriging predictions.

## 1.1 Thesis outline

The structure of this thesis is as follows. Chapter 1 presents a background on air pollution, and its effects on health as well as basic description of the data.

It further gives a general idea about the sources of environmental pollution, the location of the air pollution recording sites across Scotland focusing more on $SO_2$ data, as well as descriptive statistics about the $SO_2$ data, and preliminary diagnostic checks for constance of variance and normality assumptions.
Chapter 2 reviews some of the available literature on air pollution, its effects on health, and the various modelling techniques and methodology which have been adopted by various authors.
Chapter 3 describes methods for missing data and applies different imputation methods to the $SO_2$ data, and also compares the performance of each of the imputation techniques. Then time series modelling of $SO_2$ is used after the imputation. Various methods of parameter estimation and time series modelling techniques are discussed. Some background is given on time series decompostion as well as diagnostic checking of residuals. The results are presented in the last part of the chapter. Chapter 4 involves a model-based geostatistical interpolation of the $SO_2$ data across Scotland. It gives a general introduction to spatial analysis as well as outlining a number of topics from the theory of spatial stochastic analysis and concepts of autocorrelation, variograms and kriging, with emphasis on ordinary and Bayesian kriging. It later illustrates and identifies sources of variation in $SO_2$ levels and also estimates the pollution level at unmonitored spatial locations. Prediction outside the range of the present dataset is incorporated in the cross-validation section. The chapter concludes with sensitivity analysis and further considerations.
Chapter 5 describes the spatio-temporal analysis of $SO_2$ data using a generalized additive model. It reviews the theory of generalized additive models focusing more on the method of fitting, basis functions, basis selection and dimension criteria, and finally discusses the *mgcv* package in *R* before presenting the analysis results.
Finally, Chapter 6 gives the concluding remarks of this thesis based on the observations and statistical methods described in the first five chapters and also suggests areas for further work.

## 1.2 Introduction

Atmospheric pollution is any substance capable of altering the natural composition of air and causing harm to both humans and their environment. Air pollution has become an increasingly important focus of interest for European governments and international policy makers.

There is emerging evidence that air pollution may be strongly influenced by climate and there are several meteorological influences on the dispersion of air pollutants. This depends on wind direction, wind velocity, vertical turbulence, temperature and altitude etc. (Baumbach et al., 1996). Air pollutants are also associated with climatic change as a consequence of global warming and greenhouse effects.

There are many different air pollutants. In this thesis we focus on sulphur dioxide ($SO_2$), which is a corrosive acid gas which usually reacts with water vapour in the atmosphere to form acidic rain. It has been connected with the damage and destruction of vegetation, and the decaying of soil's fertility, building materials and watercourses are the long term impacts of $SO_2$ accumulation and deposition in the soil.

The main source of atmospheric $SO_2$ is industrial processing of materials that contain sulphur. The adverse health effects of some of the air pollutants like particulate matter ($PM$), carbon monoxide, ozone, sulphur dioxide and nitrogen dioxide etc. on daily morbidity and mortality in developed and developing countries have been confirmed (Samet et al., 2000).

There has been consistent evidence that $PM$ (particles measuring less than $10\mu m$ in diameter) is related to cardiovascular diseases and mortality. The impact of exposure to $PM_{2.5}$ concentration (particles with aerodynamic diameter less than $2.5\mu m$) is also associated with acute and chronic mortality (Laden et al., 2006). Epidemiological evidence has also linked concentrations of sulphur dioxide in the atmosphere with adverse human health effects (Brunekreef et al., 2002; Shearman, 2006; Abramson, 1991; World Health Organisation, 1992). Several studies on $SO_2$ pollutant have shown its adverse effects on human health (Touloumi et al. 1994; Schwartz et al. 1991 & 2001).

There has been much development in the assessment of the impact of $SO_2$ on health at different periods of time for exposure measurements. The interest in the $SO_2$ pollutant is focused mainly on smoke, which causes health problems, especially when it mixes with other pollutants. $SO_2$ in ambient air is related to chronic bronchitis and asthmatic diseases etc.

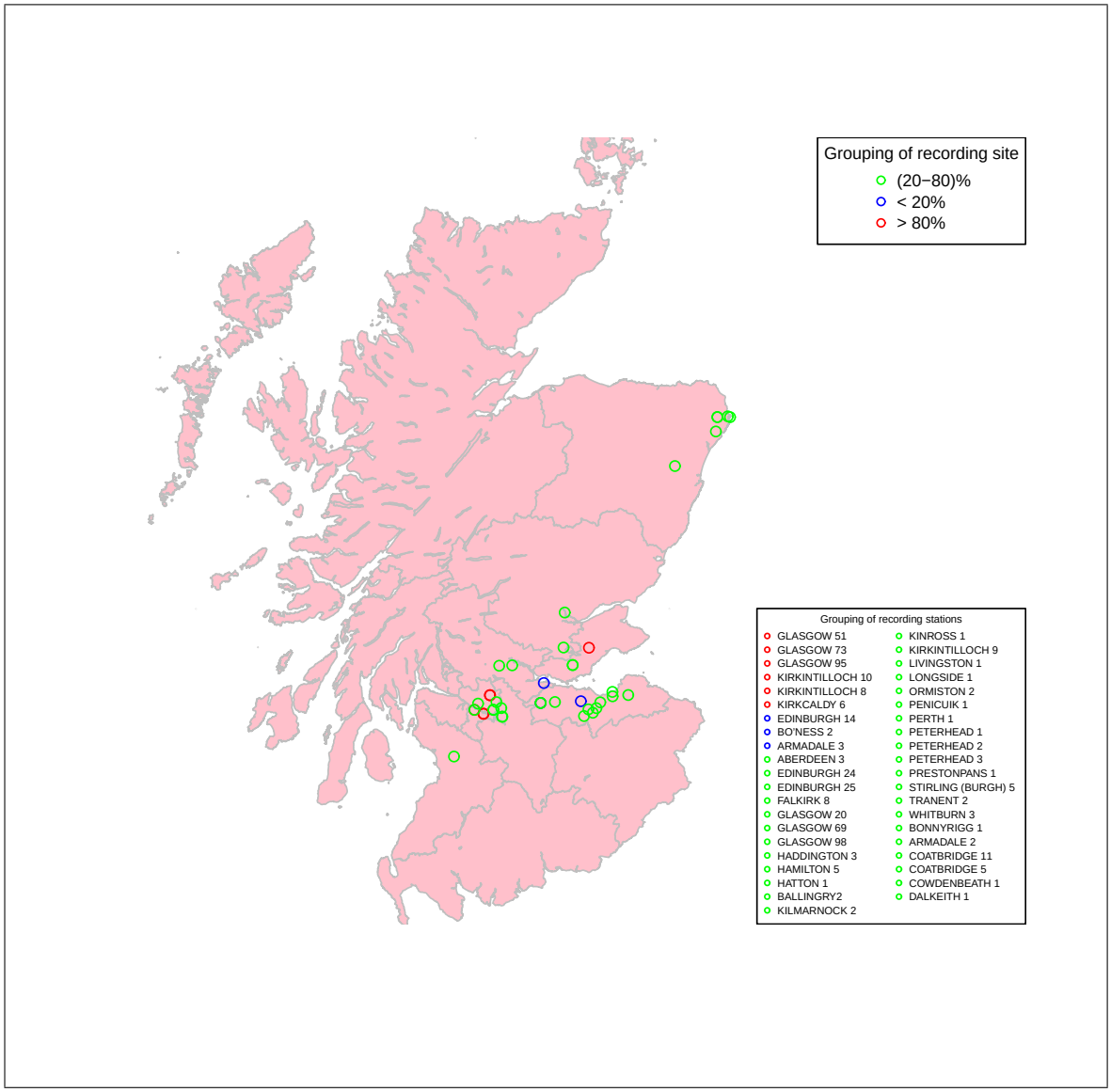
There has also been emerging evidence relating particulate matter and $SO_2$ to cardiovascular health effects (Pope et al., 1993, 1995 & 2002). Also, there is a substantive knowledge regarding interconnected pathways that relate black smoke and $SO_2$ exposure with mortality and cardiopulmonary morbidity. Governments are making rigorous policies to combat and reduce the pollutant concentration levels, especially across Europe.

Continuous measurement and assessment of the risks and potential dangers associated with air pollution is important not only because it helps to protect human health and maintain a clean environment free of hazards, but also to serve as a means to identify any areas subject to particularly high pollution levels. It may also aid government decision-making by providing daily predictions for this pollutant.
The literature review in Chapter 2 describes in more detail work which has been carried out to study air pollution and its effects on health.

## 1.3 Air pollution monitoring stations

Air quality has formed one of the UK governments key indicators of sustainable development. There are many stations across the United Kingdom which measure air pollution levels. There are mainly two different types of monitoring stations, namely automatic and non-automatic networks (see www.airquality.co.uk). In an automatic network hourly pollutant concentrations are recorded with data being collected from individual sites by modem. The detailed information for each site presently in operation is also given by the site information archive.
The Automatic Urban and Rural Network (AURN) is currently the largest automatic monitoring network in the UK. In 2007, there were 133 sites across the different countries, with England, Nothern Ireland, Scotland and Wales having 103, 3, 16 and 9 sites respectively. These recording sites provide hourly information on many pollutants to the public through online archives.
About 78 automatic stations measured $SO_2$ levels across the UK. A UV-fluorescence measurement technique is used for $SO_2$ concentration. The UK air quality strategy objective as at July 2007 for $SO_2$ to be achieved by 31st December 2004 was that "$125\mu g/m^3$ (24 hour mean) must not be exceeded more than 3 times in a year, nor should a $20\mu g/m^3$ values of $SO_2$ be recorded more than 10 times within any given year".
Non-automatic networks on the other hand measure less frequently, for daily, weekly and monthly samples. The samples are usually processed and final measurements made available for public use (www.airquality.co.uk/data).

Figure 1.1: Map showing locations of all recording stations. The colour coding indicates percentage of data available




## 1.4 The $SO_2$ data used in this thesis

This study uses data from 41 stations monitoring air pollutants widely spread over Scotland (Figure 1.1, listed by abbreviated name in Table 1.1). We obtained the data from the UK Air Quality Archive website at www.airquality.co.uk/data using the following criteria: data type (daily mean), $SO_2$ pollutant, monitoring sites across Scotland, date range between 1996-2007. The data consist of measured $S0_2$ concentration levels obtained from monitoring stations in Scotland that

measured $SO_2$ concentration as far back as 1961, but for this study we examine data from 1996-2007 for the 41 different recording stations in Figure 1.1. The choice of $SO_2$ was largely determined by the availability of data for recent years. Some of the recording sites around the Glasgow area in particular are an urban type and the main sources of pollutants are exhaust fumes from car and traffic congestion. The sites are usually surrounded with city centre businesses, offices and retail trading.

The data cover a period of 4383 days, from January 1996 to December 2007. The data represent the average of the $SO_2$ concentrations measured for one day (daily mean). We choose to use daily measurement because it gives more meaningful visualization to the seasonal pattern of the data.

We also have data on the geographical locations of the sites (Easting and Northing) obtained from the geopostcodes websites www.geopostcodes.com/index.php and www.npemap.org.uk/api/geocodes.shtml. This environmental $SO_2$ data has both attributes of space and time, as they arise from various locations in Scotland. However, there is a considerable amount of missing data and not all stations have measurements for all the years.

Table 1.1 shows the number of observations missing out of the possible 366 daily measurements for each site. "NA" indicates that there is no recorded observation for the site in that period. The stations highlighted in red have less than 20% of the data available for the years. Table 1.2 shows the full names of abbreviated stations in Table 1.1.

Forty recording stations measured $SO_2$ concentration in 1996 and 2,562 observations were missing (about 17.5% of the total data for that year), while 2003 has 16.4% of observations missing for the 16 recording stations (the lowest percentage of missing data of any of the years).

Year 2001 has the highest percentage of missing data (3,014 observations missing for only 22 recording stations that measured $SO_2$ in 2001, equivalent to about 37.5% of the data for that year). For 1997, 1998, 1999, 2000, 2002, 2004 and 2005, 34.5%, 30.1%, 32.8%, 28.5%, 26.1%, 27.2%, and 31.2% of the data are missing respectively. The minimum recorded value for all the years between 1996 and 2005 is $0\mu gm^{-3}$ and the nearest minimum value measured again is $6\mu gm^{-3}$.

In 2006 and 2007, only 4 recording stations (Aberdeen, Edinburgh St. Leonards, Glasgow Centre and Grangemouth) recorded $S0_2$ data, and 3.9% and 11.3% of the observations were missing for 2006 and 2007 respectively, as seen in Table 1.1. Only 15 stations have data in 2005, while years 2003 and 2004 have data for just 16 stations.

It can be generally observed that there are sparse observations in the later year. Years 2001 and 2002 have data from just 22 and 19 recording stations respectively, and this decreases in later years to 15 in year 2005

Only 6 out of the 41 recording stations have data available for at least 80% of the days (Figure 1.2). These are Glasgow 51, Glasgow 73, Glasgow 95, Kirkintilloch 8, Kirkintilloch 10 and Kirkcaldy 6. Three of the sites recorded less than 20% of the data possible, namely Armadale 3, Boness 2 and Edinburgh 14 (Figure 1.3). Also, because we have many observations in our data (1996-2007), most of the illustrative analysis in this thesis will focus mainly on a few years, namely on the 1996, 2000, 2005 and 2007 datasets. We assume these are representative of the whole data as these selected years represent the early, the middle and the later years in our data ranging from 1996-2007.

Table 1.1: Structure of the missing data for each station; number of missing observations out of 366 possible days for each year; "NA" indicates that there is no recorded observation for the site in that period and stations highlighted in red have less than 20% of data available for these years

| site | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | obs. Days | Total | % | | 2006 | 2007 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AB3 | 204 | 133 | 133 | 3 | 179 | 64 | 42 | 0 | 123 | 238 | 3653 | 1119 | 36.7 | AB | 5 | 132 |
| EH14 | 365 | 303 | 282 | 145 | NA | NA | NA | NA | NA | NA | 1460 | 1095 | 89.9 | ED. ST. LEC | 5 | 12 |
| EH24 | 2 | 131 | 131 | 192 | NA | NA | NA | NA | NA | NA | 1460 | 456 | 72.4 | GLA. CEN. | 38 | 14 |
| EH25 | 365 | 342 | 342 | 1 | 141 | 121 | 212 | 230 | 323 | 207 | 3653 | 2284 | 62.5 | GRANGEM | 9 | 7 |
| FK8 | 152 | 1 | 0 | 2 | 271 | NA | NA | NA | NA | NA | 1825 | 426 | 61.6 | | | |
| G20 | 2 | 4 | 4 | 263 | 8 | 161 | NA | NA | 252 | 172 | 2190 | 866 | 43.7 | | | |
| G51 | 5 | 0 | 0 | 0 | 105 | 69 | 15 | 33 | 24 | 15 | 3653 | 266 | 7.2 | | | |
| G69 | 0 | 1 | 1 | 0 | 262 | 67 | 25 | 291 | NA | NA | 2920 | 647 | 37.7 | | | |
| G73 | 1 | 1 | 1 | 0 | 1 | 77 | 62 | 129 | 219 | 192 | 3653 | 683 | 18.7 | | | |
| G95 | 1 | 6 | 6 | 168 | 0 | 65 | 77 | 28 | 46 | 23 | 3653 | 420 | 11.5 | | | |
| G98 | 1 | 275 | 276 | 11 | 1 | 165 | 157 | 45 | 131 | 202 | 3653 | 1264 | 34.6 | | | |
| EH41 | NA | 275 | 276 | 134 | NA | NA | NA | NA | NA | NA | 1095 | 685 | 78.7 | | | |
| ML3 | 93 | 19 | 18 | 13 | NA | NA | NA | NA | NA | NA | 1460 | 143 | 63.8 | | | |
| ML1 | 25 | 8 | 8 | 117 | 270 | NA | NA | NA | NA | NA | 1825 | 428 | 61.7 | | | |
| KY4 | 8 | 109 | 109 | 50 | 0 | 217 | 274 | NA | NA | NA | 2555 | 767 | 50.9 | | | |
| KA1 | 136 | 364 | NA | 17 | 103 | 365 | 274 | NA | NA | NA | 2190 | 1259 | 63.4 | | | |
| KY13 | 51 | 51 | 51 | 187 | NA | NA | NA | NA | NA | NA | 1460 | 340 | 59 | | | |
| KY2 | 111 | 45 | 0 | 77 | 109 | 70 | 24 | 44 | 45 | 282 | 3653 | 807 | 22.1 | | | |
| G66 | 12 | 1 | 1 | 188 | 16 | 6 | 26 | 0 | 9 | 0 | 3653 | 259 | 7 | | | |
| G64 | 3 | 37 | 37 | 180 | 14 | 7 | 0 | 1 | 1 | 75 | 3653 | 355 | 9.7 | | | |
| G64 | 71 | 279 | 280 | 196 | 82 | 343 | 0 | 0 | 277 | NA | 3285 | 1528 | 52 | | | |
| EH54 | 54 | 17 | 17 | 8 | NA | NA | NA | NA | NA | NA | 1460 | 96 | 62.5 | | | |
| AB43 | 16 | 275 | 276 | 103 | 168 | NA | NA | NA | NA | NA | 1825 | 838 | 72.9 | | | |
| ML3 | 61 | 19 | 18 | 168 | NA | NA | NA | NA | NA | NA | 1460 | 266 | 67.2 | | | |
| EH26 | 7 | 246 | 247 | 110 | 54 | 273 | NA | NA | NA | NA | 2190 | 937 | 54.7 | | | |
| PH2 | 4 | 2 | 1 | 188 | NA | NA | NA | NA | NA | NA | 1460 | 195 | 65.2 | | | |
| B77 | 3 | 12 | 12 | 207 | 168 | NA | NA | NA | NA | NA | 1825 | 402 | 60.9 | | | |
| AB42 | 18 | 31 | 0 | 103 | 178 | NA | NA | NA | NA | NA | 1825 | 330 | 58.9 | | | |
| AB43 | 0 | 275 | 276 | 159 | 176 | NA | NA | NA | NA | NA | 1825 | 886 | 74.2 | | | |
| EH32 | 150 | 7 | 7 | 364 | NA | NA | NA | NA | NA | NA | 1460 | 528 | 74.4 | | | |
| FK7 | 84 | 275 | 276 | 364 | 110 | 176 | 274 | NA | NA | NA | 2555 | 1559 | 72.6 | | | |
| EH33 | 112 | 48 | 48 | NA | NA | NA | NA | NA | NA | NA | 1095 | 208 | 75.6 | | | |
| EH48 | 9 | 253 | 254 | NA | 1 | 113 | 50 | 36 | 6 | 32 | 3285 | 754 | 31.4 | | | |
| EH51 | 35 | 241 | 242 | NA | 271 | NA | NA | NA | NA | NA | 1460 | 789 | 81.5 | | | |
| EH19 | 13 | 57 | 58 | NA | NA | NA | NA | NA | NA | NA | 1095 | 128 | 73.4 | | | |
| EH48 | 23 | 275 | 276 | NA | 7 | 58 | 25 | 73 | 85 | 37 | 3285 | 859 | 33.5 | | | |
| EH48 | 185 | 153 | 152 | NA | NA | NA | NA | NA | NA | NA | 1095 | 490 | 83.3 | | | |
| EH35 | 42 | 158 | 158 | NA | 105 | 215 | NA | 2 | 0 | 8 | 2920 | 688 | 38.8 | | | |
| G69 | NA | NA | NA | NA | NA | NA | NA | 44 | 41 | 167 | 1095 | 252 | 66.8 | | | |
| ML5 | 71 | 84 | 85 | NA | 81 | 109 | 274 | NA | NA | NA | 2190 | 704 | 59.2 | | | |
| KY4 | 57 | 21 | 20 | NA | NA | 0 | 2 | 0 | 7 | 61 | 2920 | 168 | 24.5 | | | |
| EH22 | 10 | 364 | 20 | NA | 37 | 273 | 0 | NA | NA | NA | 2190 | 704 | 48.2 | | | |
| **TOTAL** | **2562** | **5198** | **4399** | **3718** | **2918** | **3014** | **1813** | **956** | **1589** | **1711** | **97117** | **27878** | | | | |
| **No. of site** | **40** | **41** | **40** | **31** | **28** | **22** | **19** | **16** | **16** | **15** | | | | | | |
| **Missing %** | **17.5** | **34.5** | **30.1** | **32.8** | **28.5** | **37.5** | **26.1** | **16.4** | **27.2** | **31.2** | | | | | | |

Table 1.2: Recording stations and their abbreviations

| station | | | |
|---|---|---|---|
| ABERDEEN 3 | AB 3 | LIVINGSTON 1 | EH 54 |
| EDINBURGH 14 | EH 14 | LONGSIDE 1 | AB 43 |
| EDINBURGH 24 | EH 24 | ORMISTON 2 | ML 5 |
| EDINBURGH 25 | EH 25 | PENICUIK 1 | EH 26 |
| FALKIRK 8 | FK 8 | PERTH 1 | EH 2 |
| GLASGOW 20 | G 20 | PETERHEAD 1 | AB 77 |
| GLASGOW 51 | G 51 | PETERHEAD 2 | AB 42 |
| GLASGOW 69 | G 69 | PETERHEAD 3 | AB 43 |
| GLASGOW 73 | G 73 | PRESTONPANS 1 | EH 32 |
| GLASGOW 95 | G 95 | STIRLING (BURGH) 5 | FK 7 |
| GLASGOW 98 | G 98 | TRANENT 2 | EH 33 |
| HADDINGTON 3 | EH 41 | WHITBURN 3 | EH 48 |
| HAMILTON 5 | ML 3 | BO'NESS 2 | EH 51 |
| HATTON 1 | ML 1 | BONNYRIGG 1 | EH 19 |
| BALLINGRY2 | KY 4 | ARMADALE 2 | EH 48 |
| KILMARNOCK 2 | KA 1 | ARMADALE 3 | EH 35 |
| KINROSS 1 | KY 13 | COATBRIDGE 11 | G 69 |
| KIRKCALDY 6 | KY 2 | COATBRIDGE 5 | ML 11 |
| KIRKINTILLOCH 10 | G 66 | COWDENBEATH 1 | KY 4 |
| KIRKINTILLOCH 8 | G 64 | DALKEITH 1 | EH 22 |
| KIRKINTILLOCH 9 | G 64 | | |

## 1.5 Data description

Table 1.3 shows descriptive statistics of $SO_2$ levels for the years 1996, 2000 and 2005. Both the mean and maximum values generally decline across the years, especially for those stations that have more than 80% of data available, i.e Glasgow 51, Glasgow 73, Glasgow 95, Kirkintilloch 8, Kirkintilloch 10 and Kirkcaldy 6. The minimum and maximum recorded values are 0 and $229\mu gm^{-3}$ respectively. There is no obvious pattern in the standard deviation and median concentrations. The blanks in the table indicate that there is no recorded observation for the site in that period.

Figures 1.2-1.4 show the geographical distribution of the monitoring sites according to the percentage of available observations. They give the structure of the missing pattern according to spatial location. Most of the sites are in the Central region of Scotland. There are no monitoring stations in the North-Western region of the map because of the low population density and little industrial activities that involve pollutant emmission in the region. We also have fewer stations along the Coastal area. In Figure 1.2, the six stations that have a higher percentage of data available are located in Central Scotland. The stations are Glasgow 51, Glasgow 73, Glasgow 95, Kirkintilloch 10, Kirkintilloch 8 and Kirkcaldy 6 while the three stations that have fewest observations are located along the Coast. These are Edinburgh 14, Bo'ness 2 and Armadale 3. In Figure 1.4, the stations with 20-80% of data available are located both within the Central Scotland and North-Western regions.

Table 1.3: Descriptive statistics for $SO_2$ levels by year, for each recording station

| | min. | | | mean | | | sd | | | median | | | max. | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **station** | **1996** | **2000** | **2005** | **1996** | **2000** | **2005** | **1996** | **2000** | **2005** | **1996** | **2000** | **2005** | **1996** | **2000** | **2005** |
| ABERDEEN 3 | 0 | 0 | 30 | 15.89 | 19.93 | 48.17 | 6.8 | 11.2 | 11.6 | 14 | 19 | 50.5 | 40 | 56 | 64 |
| EDINBURGH 14 | | | | | | | | | | | | | | | |
| EDINBURGH 24 | 0 | | | 20.25 | | | 10 | | | 20 | | | 57 | | |
| EDINBURGH 25 | | 0 | 0 | | 24.2 | 4.709 | | 10.7 | 5.5 | | 21 | 0 | | 50 | 18 |
| FALKIRK 8 | 0 | | | 17.13 | | | 12.8 | | | 14 | | | 78 | | |
| GLASGOW 20 | 0 | 0 | 0 | 27.15 | 24.54 | 18.51 | 13.6 | 7.8 | 10.2 | 26 | 25 | 19 | 79 | 57 | 53 |
| GLASGOW 51 | 0 | 0 | 0 | 22.23 | 22.12 | 18.81 | 11.7 | 7.3 | 8 | 20 | 19 | 20 | 58 | 50 | 60 |
| GLASGOW 69 | 0 | 0 | | 25.59 | 22.05 | | 13.4 | 10.1 | | 26 | 24 | | | 41 | |
| GLASGOW 73 | 0 | 0 | 0 | 19.23 | 16.44 | 13.72 | 12.5 | 7.8 | 8.4 | 19 | 19 | 13 | 104 | 43 | 39 |
| GLASGOW 95 | 0 | 0 | 0 | 26.51 | 16.67 | 34.26 | 14.7 | 7.9 | 17.4 | 26 | 18 | 32 | 96 | 49 | 106 |
| GLASGOW 98 | 0 | 0 | 0 | 20.85 | 18.77 | 11.71 | 13.2 | 7 | 5.6 | 20 | 20 | 13 | 68 | 46 | 41 |
| HADDINGTON 3 | | | | | | | | | | | | | | | |
| HAMILTON 5 | 6 | | | 26.53 | | | 7.7 | | | 25 | | | 51 | | |
| HATTON 1 | 0 | 6 | | 5.153 | 7.109 | | 2.9 | 2.3 | | 6 | 6 | | 13 | 12 | |
| BALLINGRY2 | 6 | 6 | | 16.5 | 14.05 | | 5.3 | 5 | | 14 | 13 | | 34 | 32 | |
| KILMARNOCK 2 | 6 | 6 | | 9.83 | 8.866 | | 3.2 | 4.4 | | 12 | 6 | | 14 | 19 | |
| KINROSS 1 | 6 | | | 10.98 | | | 4.7 | | | 13 | | | 26 | | |
| KIRKCALDY 6 | 6 | 6 | 6 | 17.4 | 10.18 | 7.301 | 5.9 | 4.6 | 1.5 | 19 | 12 | 7 | 39 | 30 | 14 |
| KIRKINTILLOCH 10 | 7 | 7 | 0 | 25.17 | 19.53 | 18.05 | 10.2 | 7.9 | 9 | 24 | 20 | 18 | 97 | 74 | 49 |
| KIRKINTILLOCH 8 | 6 | 0 | 0 | 14.84 | 10.16 | 12.85 | 7.5 | 7 | 7.3 | 13 | 7 | 12 | 69 | 76 | 42 |
| KIRKINTILLOCH 9 | 0 | 6 | | 19.3 | 13.49 | | 9.6 | 6 | | 18 | 13 | | 102 | 74 | |
| LIVINGSTON 1 | 0 | | | 9.695 | | | 6.9 | | | 12 | | | 32 | | |
| LONGSIDE 1 | 0 | 0 | | 12.4 | 5.077 | | 5.2 | 2.5 | | 13 | 6 | | 47 | 12 | |
| ORMISTON 2 | 6 | | | 15.46 | | | 8.4 | | | 12 | | | 52 | | |
| PENICUIK 1 | 6 | 6 | | 21.03 | 17.7 | | 6.8 | 8.2 | | 19 | 15.5 | | 43 | 44 | |
| PERTH 1 | 6 | | | 18.21 | | | 6.3 | | | 18 | | | 42 | | |
| PETERHEAD 1 | 0 | 6 | | 12.39 | 8.7 | | 4.5 | 3.4 | | 12 | 6 | | 43 | 13 | |
| PETERHEAD 2 | 0 | 0 | | 7.331 | 5.707 | | 3.2 | 2.1 | | 6 | 6 | | 18 | 13 | |
| PETERHEAD 3 | 0 | 0 | | 7.997 | 8.2 | | 4.2 | 3.5 | | 6 | 6 | | 50 | 13 | |
| PRESTONPANS 1 | 6 | | | 11.03 | | | 4.7 | | | 12 | | | 38 | | |
| STIRLING (BURGH) 5 | 0 | 6 | | 16.73 | 11.6 | | 5.9 | 5.8 | | 18 | 13 | | 43 | 38 | |
| TRANENT 2 | 6 | | | 6.953 | | | 2.5 | | | 6 | | | 18 | | |
| WHITBURN 3 | 0 | 7 | 0 | 10.04 | 20.18 | 25.42 | 8.1 | 5.8 | 7.6 | 13 | 19 | 25 | 39 | 39 | 50 |
| BO'NESS 2 | 6 | | | 18.53 | | | 8.7 | | | 19 | | | 27 | | |
| BONNYRIGG 1 | 0 | | | 15.43 | | | 6.6 | | | 13 | | | 31 | | |
| ARMADALE 2 | 0 | 13 | 7 | 44.62 | 34.02 | 40.11 | 27.7 | 15.8 | 13.6 | 40.5 | 28 | 39 | 229 | 116 | 99 |
| ARMADALE 3 | 0 | | | 16.91 | | | 12.8 | | | 18 | | | 60 | | |
| COATBRIDGE 11 | 0 | 6 | 0 | 35.52 | 43.43 | 20.48 | 16.6 | | 10.8 | 35 | 26 | 19 | 129 | 118 | 61 |
| COATBRIDGE 12 | | | 0 | | | 13.93 | | | 10 | | | 13 | | | 40 |
| COATBRIDGE 5 | 0 | 0 | | 30.39 | 21.69 | | 14.4 | 8.2 | | 31 | 23 | | 126 | 70 | |
| COWDENBEATH 1 | 0 | 6 | 7 | 15.81 | 14.43 | 13.22 | 16.6 | 4.7 | 5.3 | 14 | 14 | 13 | 34 | 27 | 27 |
| DALKEITH 1 | 0 | 6 | | 24.97 | 24.55 | | 8.1 | 9.1 | | 24 | 24 | | 55 | 49 | |

Figure 1.2: Map showing the 6 recording stations with more than 80% of data available

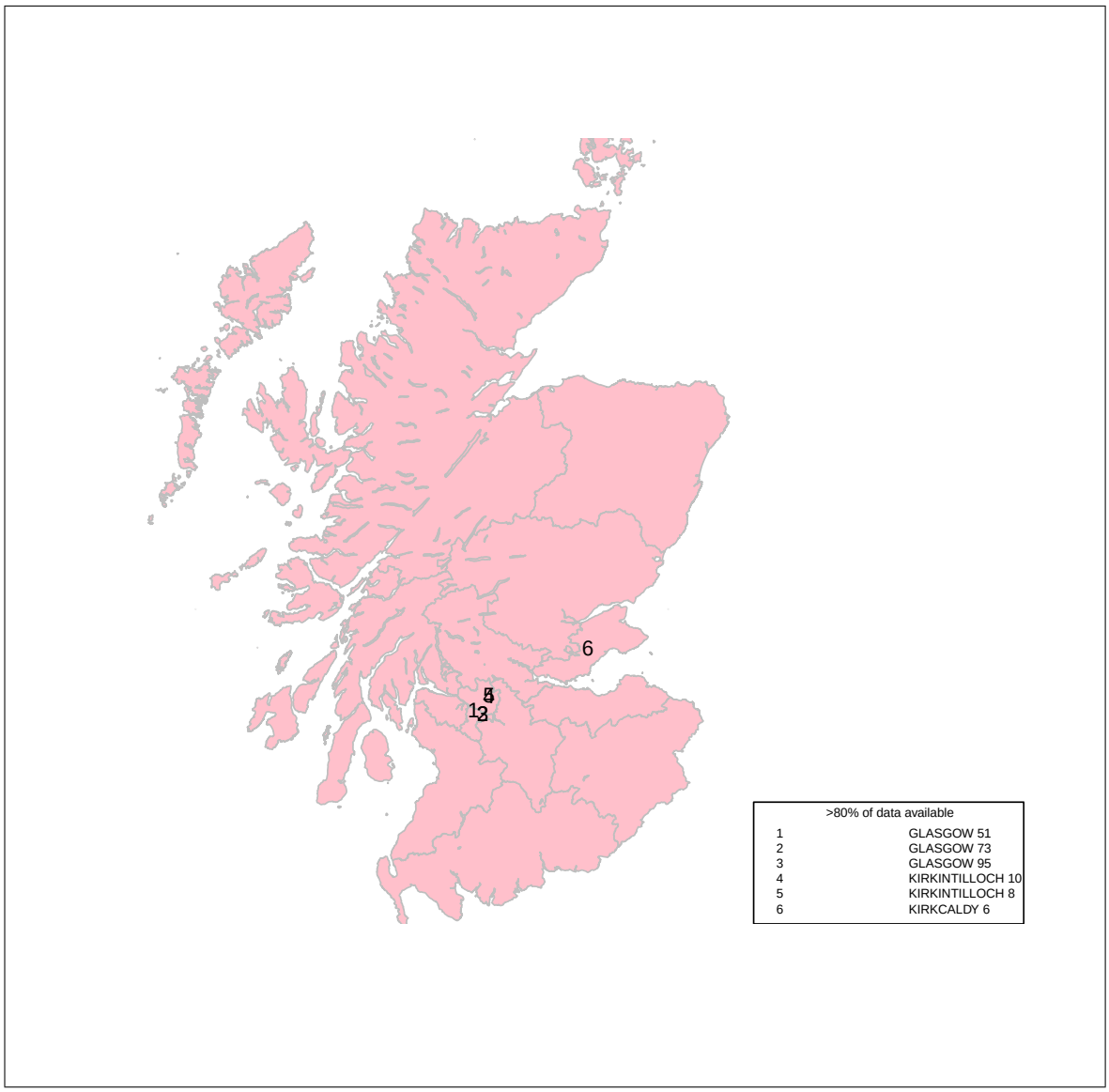


Figure 1.3: Map showing the 3 stations with less than 20% of data available

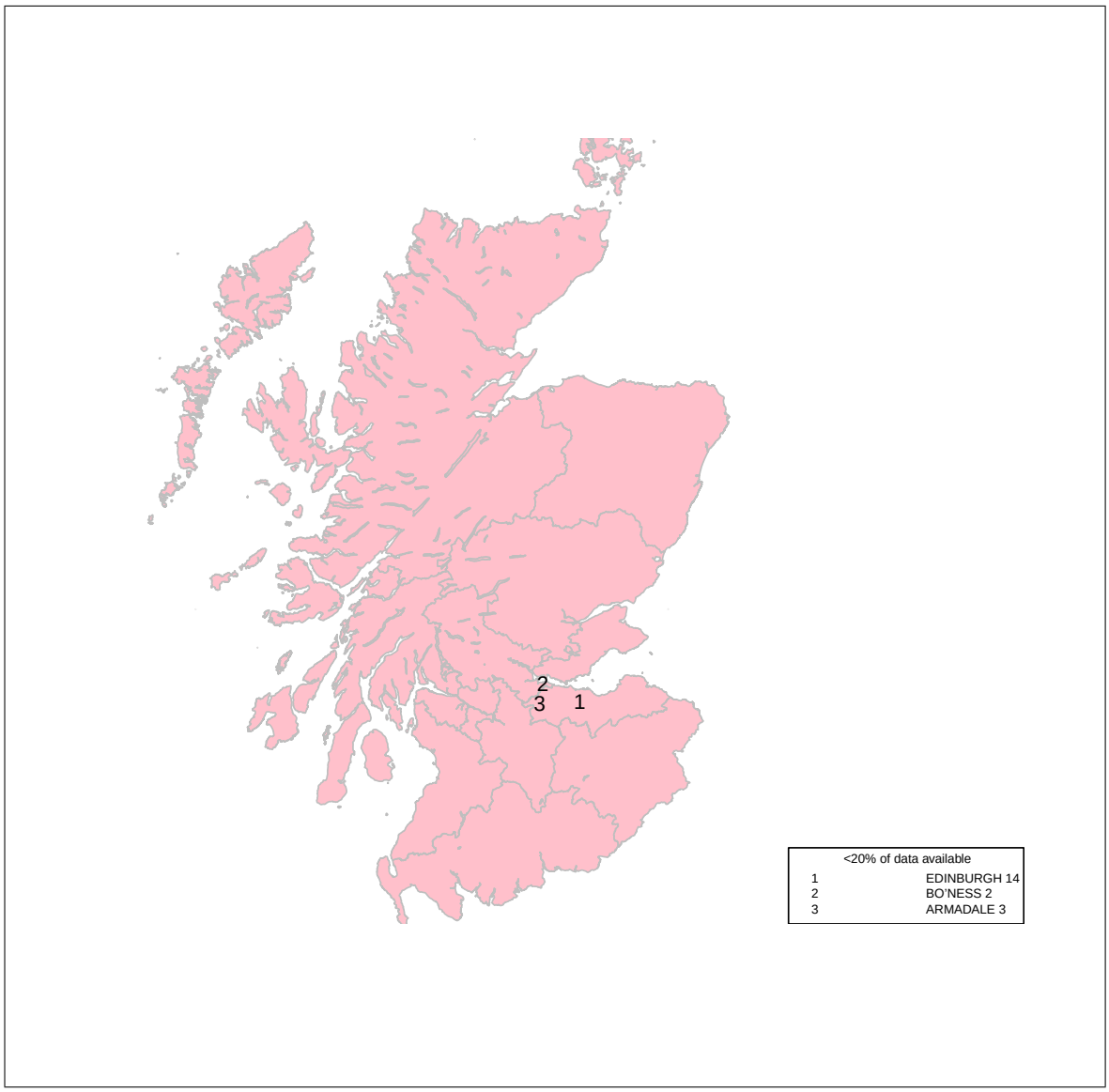
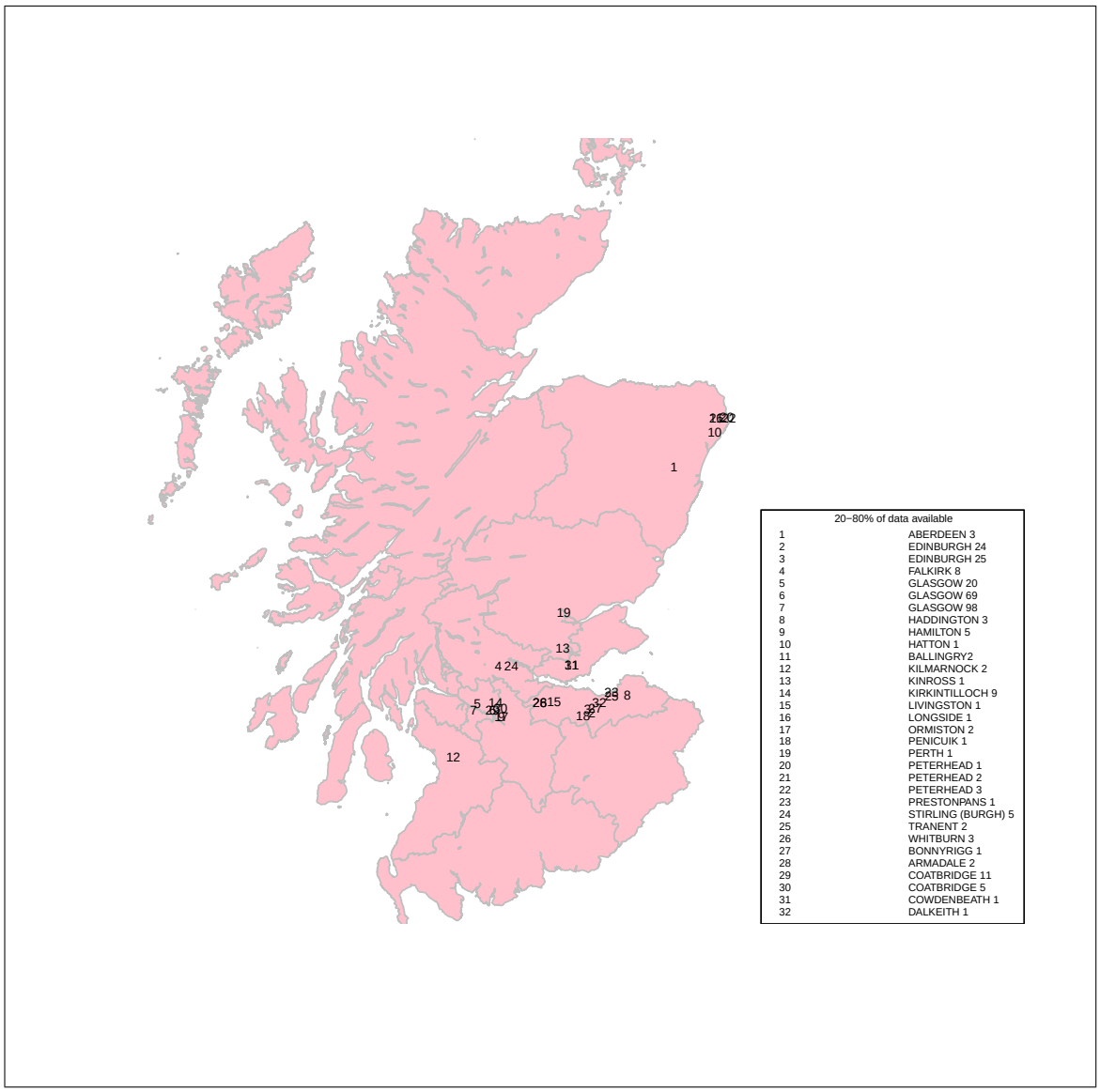


Figure 1.4: Map of recording stations with 20-80% of data available



Figures 1.5 shows time series plots of daily mean $SO_2$ for some of the recording stations, for us to see the pattern of the data in 1996. Three of these stations (Glasgow 51, Glasgow 73 and Glasgow 95) have more than 80% of observations available, and the remaining stations have $20-80\%$ of data available (Glasgow 20, Glasgow 69, Glasgow 98 and Falkirk 8). The chosen stations have fewer missing observations which will enable us to clearly see the variation in pattern of $SO_2$ levels across the stations and time period.

We observe that there is a considerable variation in mean concentration of $SO_2$. We see that the results for Glasgow 69 hover around (20-60) $\mu g/m^3$ in January and come down to around 10 $\mu g/m^3$ in February, thereafter the level increases to around 30 and there is a periodic rise and fall between (days 100-180). There is relatively little variation apart from a period around day 180-200 (July) with greater volatility. In Glasgow 51 there is much more variability than in Glasgow 69 though the levels are lower. Glasgow 20 also has similar variation in pattern to Glasgow 51 for the first 2 months (days 1-60) with decreasing levels within this interval, and there is no obvious pattern for the rest of the period except a prominent peak around days 180-200, and days 270-280.

Falkirk 8 has a quite different pattern of variation than the other 3 stations with prominent gaps in the series which are due to missing observations. It has relatively low mean levels with less than $20\mu g/m^3$ in January also, with a spike between days 140-160.

There is a relatively low mean level for Glasgow 73, with little fluctuation but has a prominent peak around day 230 (August) which is about 100 $\mu g/m^3$ of concentration. Continuous variation in mean levels for Glasgow 95 is also obvious in the plot with a spike of about 90 $\mu g/m^3$ which occurs around day 50 (February). The Glasgow 98 pattern is similar to Glasgow 51 and Glasgow 20 for the first 2 months with a prominent peak around day 300 (October). Generally, each of the 7 stations we consider in 1996 has its own unique pattern. There is no common pattern in all these figures. Specifically the spike in Glasgow 69 around day 190 is not obvious in the other series.

Some stations we consider in 1996 (Glasgow 51, Glasgow 20 and Falkirk 8) are not repeated again in the next plots because they have no or few recorded observations in 2000 and 2005. In Figures 1.6, Glasgow 69 only operates towards the end of the year (from September) with a gap between November and December which could be due to missing observations. There is a relatively low levels varies within 0-40 $\mu g/m^3$. For Glasgow 73, there is a gradual increase in mean levels between days 1-90 (January-March), and a relatively constant mean level

between days 90-270 with a little reduction in levels thereafter. This pattern is similar in Glasgow 95 except for 2 spikes between days 120-160.
Glasgow 98 also has a relatively low mean level which is similar to Glasgow 73 and Glasgow 95. Kirkintilloch 8, Kirkintilloch 9 and Kirkintilloch 10 have very similar pattern to each other, which could be due to proximity of the stations. They generally have very low mean level (below 20 $\mu g/m^3$) throughout the period. Few gaps are obvious in all three series. There is a prominent peak around day 120 (April) common to all three sites.
In Figures 1.7, gaps are getting more obvious in Glasgow 20, Glasgow 73, Coatbridge 12 and Cowdenbeath 1. These stations are characterized by unusual isolated fluctuations, also with relatively lower mean levels, as compared to stations in 1996 and 2000. Glasgow 73 recorded observations for only a very few days between days 1-200. Coatbridge 12 does not record any data in the early part of the year (days 1-90) while Cowdenbeath 1 on the other hand has no recorded observations in the later part of the year (November-December). Glasgow 51 has little variation in levels throughout the period with a spike around day 110, and Coatbridge 11 has similar low variation in pattern.
Generally, Figures 1.5-1.7 indicate unique variation in pattern for the stations, though a few stations have similar pattern for some period of time. Mean levels of $SO_2$ generally decrease with the years, in which 1996 has the highest mean levels for most stations (Glasgow 73 spike), and the least concentration in 2005 (Glasgow 73 recorded far below 10 $\mu g/m^3$ around day 40) in Figure 1.7. We next use the long term trend across years in Figure 1.8 to justify the fall in mean level pattern we observe here.
Figure 1.8 shows the long term trend across the years of daily mean concentration for two of the stations with more than 80% of data recorded (Glasgow 73 and Glasgow 95). The essence of this is to further observe the changes in variational pattern of $SO_2$ levels across the years, especially for those stations with fewer missing observations. This will also enable us to compare the trend pattern on a yearly basis. Again there is considerable fluctuation in the levels across time. For Glasgow 73, there is a gradual increase in level between 1996-1998 and then a decrease thereafter till 2002, with few high fluctuations between this period and 2005. Glasgow 95 has relatively constant mean levels between 1996 and 2002 after an initial high level, and an increase in levels characterized the 2002-2005 period.

Figure 1.5: Time series plot of daily mean $SO_2$ concentrations for some stations in 1996

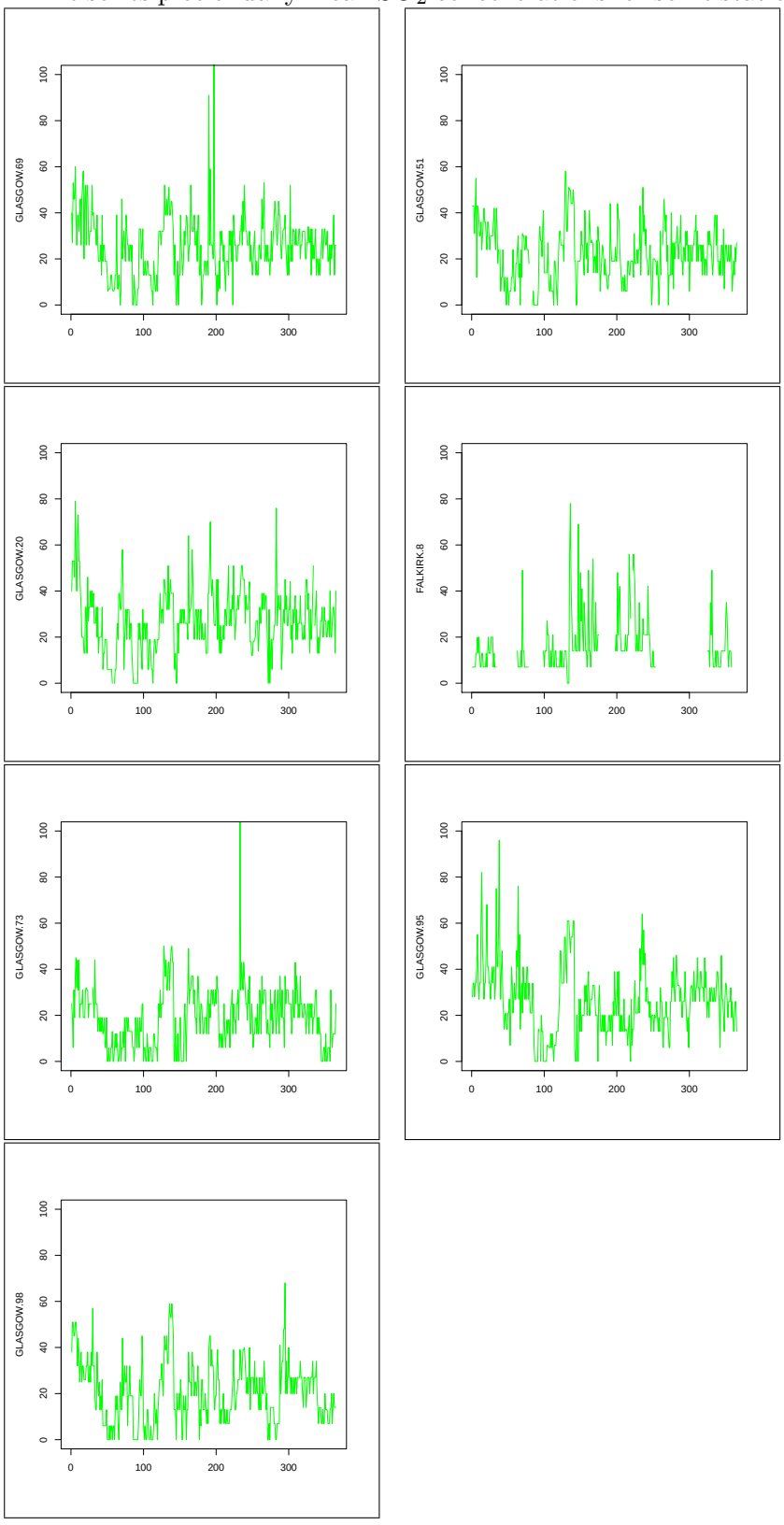


Figure 1.6: Time series plot of daily mean $SO_2$ concentrations for some stations in 2000

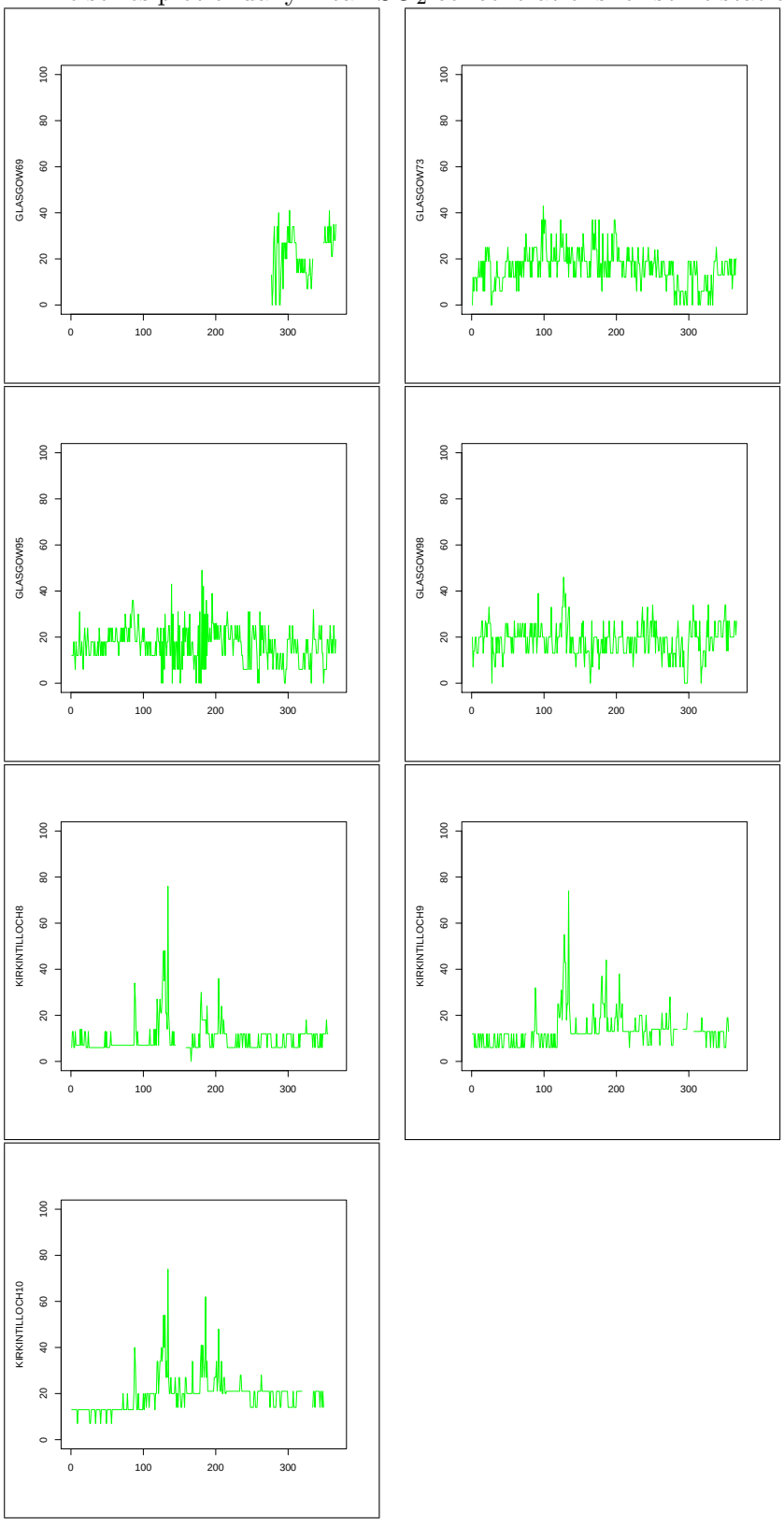
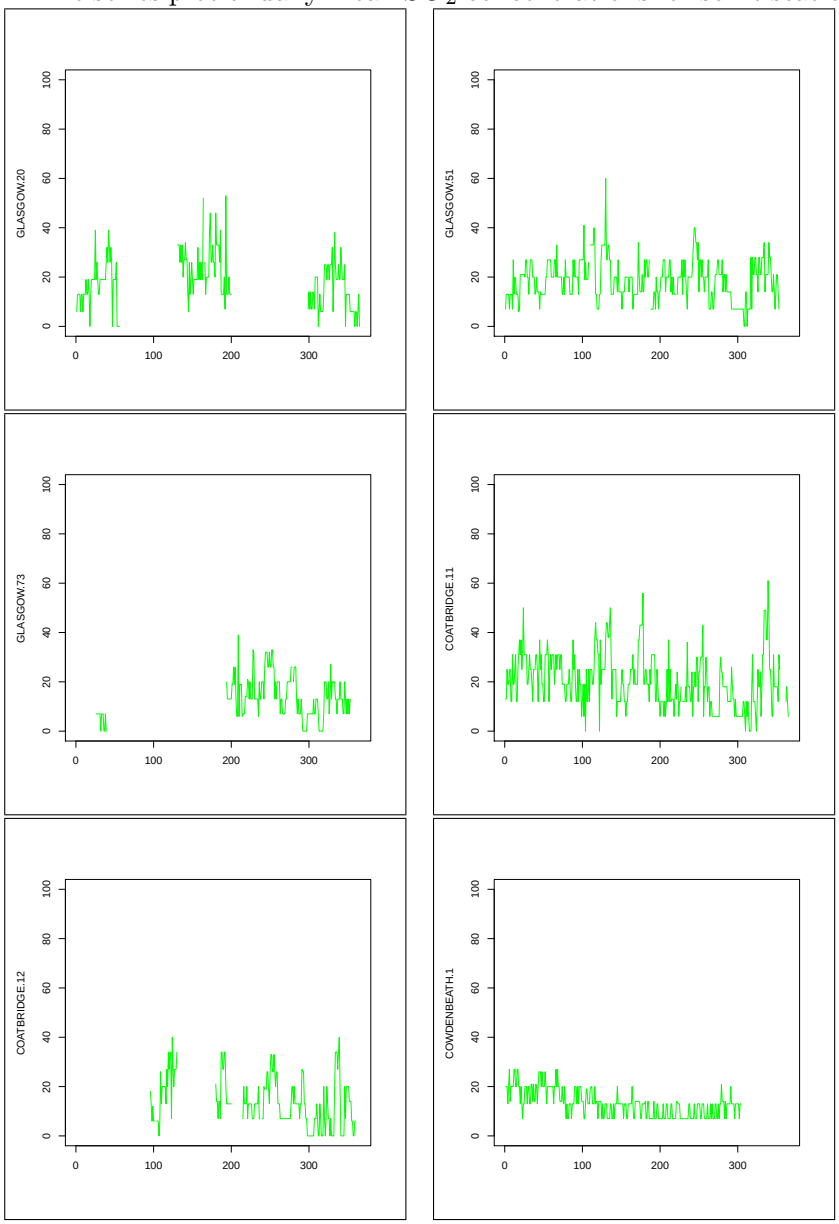


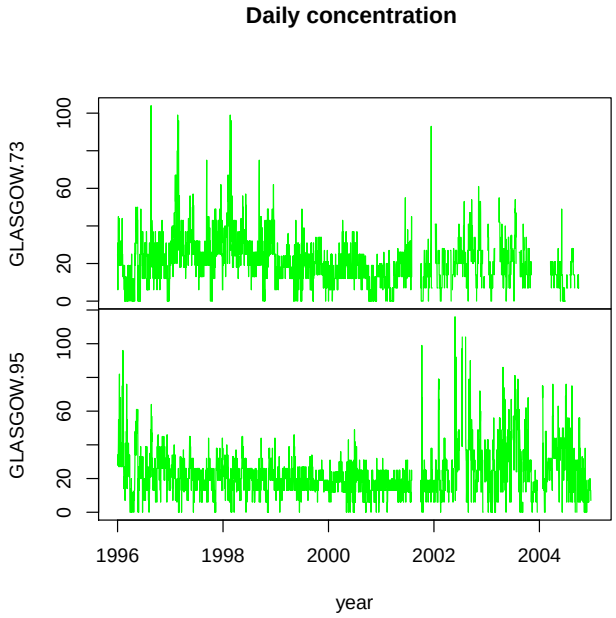
Figure 1.7: Time series plot of daily mean $SO_2$ concentrations for some stations in 2005




We consider the monthly averages rather than daily mean levels to investigate the seasonal effect more closely. Figures 1.9-1.11 show plots of monthly average $SO_2$ concentration for some sites in 1996, 2000 and 2005 respectively. In Figure 1.9, Falkirk 8 has a gradual increase in mean level between January and August. August has the highest concentration of about 25 $\mu g/m^3$ then a decrease in level to about 15 $\mu g/m^3$ between August and September. There is no recorded observation for this station between October and middle of November. The December level is relatively low.

Glasgow 20 has a decrease in level between January and February with gentle rise

Figure 1.8: Long term trend of daily $SO_2$ concentrations for Glasgow.73 and Glasgow.95




and fall between February and April and thereafter increases till the end of the year. It also has a summer peak in August similar to Falkirk 8. Glasgow 51 and Glasgow 69 have similar decreasing patterns between January and April, similar to Glasgow 20, with a prominent peak in May, a sudden fall in level between May and July, and relatively constant levels for the rest of the year (July-December). In Figure 1.10, Glasgow 20 and Glasgow 51 have increase in level between January and August with a low level in June. Glasgow 20 has a peak in August similar to Falkirk 8 and Glasgow 20 in Figure 1.9. Glasgow 51 has dual peaks, one in July, the other around September, and thereafter a rapid decrease then increase. Glasgow 69 only recorded observations for the last quarter of the year similar to what we observe in Figure 1.6. There is no obvious pattern for Glasgow 73, though it has relatively constant mean level between April and August.
In Figure 1.11, Kirkintilloch 8 and Kirkintilloch 10 have similar patterns, with both exhibiting May and September peaks. Whitburn 3 and Armadale 2 also show similar patterns with Armadale 2 having a prominent peak in July. Generally in Figures 1.9-1.11 fluctuations in levels are seen in all of these plots, although these are different for each station and there is a little variation in seasonal pattern with peak concentrations occurring in summer months.

Figure 1.9: Time series plot of monthly mean $SO_2$ for some stations in 1996

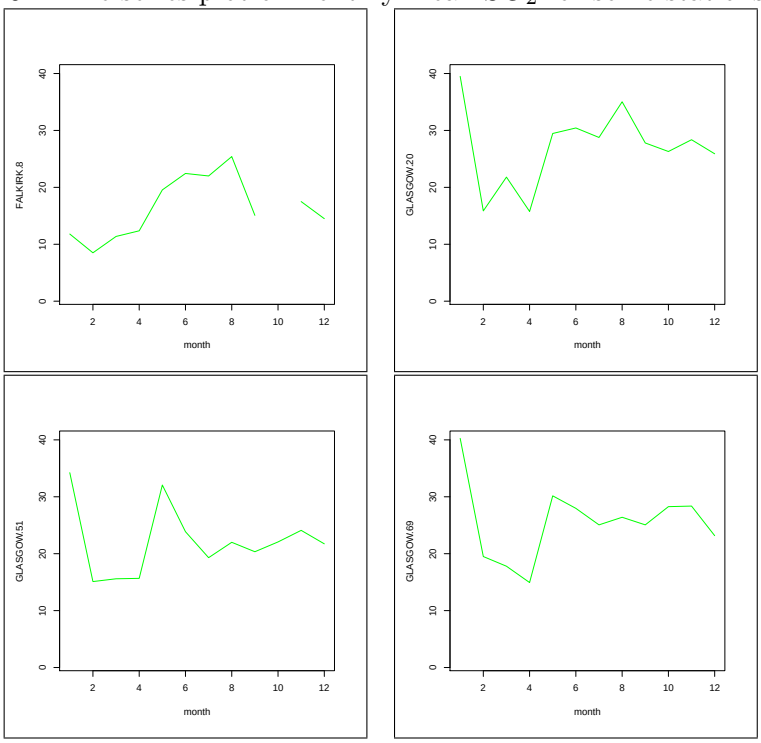



Figure 1.10: Time series plot of monthly mean $SO_2$ for some stations in 2000

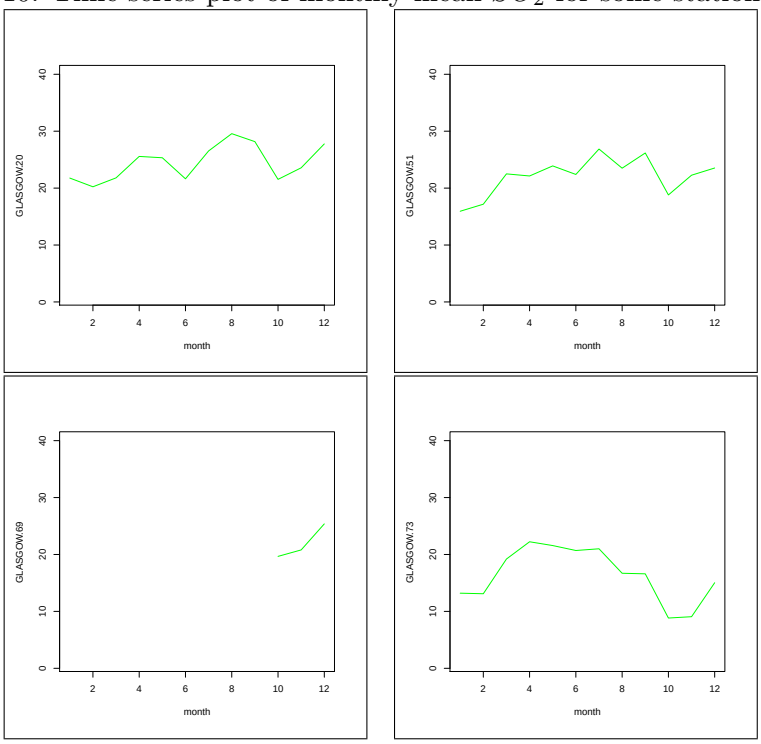
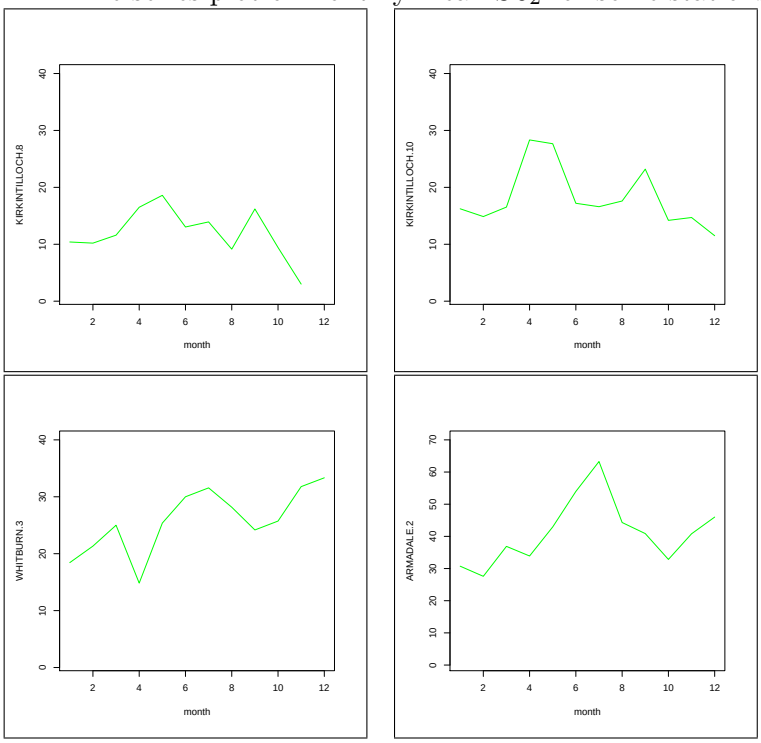


Figure 1.11: Time series plot of monthly mean $S0_2$ for some stations in 2005



Figures 1.12-1.14 now show histograms of the daily $SO_2$ concentration for some stations in 1996, 2000 and 2005. In Figure 1.12 most of the levels are within the 40 $\mu g/m^3$ limit and the data points are clustered together. None of the plots is symmetric. There is a general tendency for skewness to the right, reflecting occasional high levels of $SO_2$.

In Figure 1.13, Aberdeen 3, Edinburgh 25 and Glasgow 20 histograms are similar with clustered data points. Concentration levels are very high, Aberdeen 3 is also right skewed but Glasgow 51, Glasgow 69 and Glasgow 73 have very scanty data with generally low concentration levels.

In Figure 1.14, Aberdeen 3 has most of its data within 10-20 $\mu g/m^3$ while Edinburgh 25 has fewer recorded observations, with low mean level, and most data are within 6-7 $\mu g/m^3$. Also, Glasgow 20, Glasgow 51, Glasgow 73 and Glasgow 95 seem more symmetric with clustered data points. Generally, the histograms in Figures 1.12-1.14 show that most of the data are closer together (within the same range) for 1996 and 2000 than 2005, so the year 2005 histograms have more dispersed data.

Figure 1.12: Histograms of daily $SO_2$ concentration in 1996

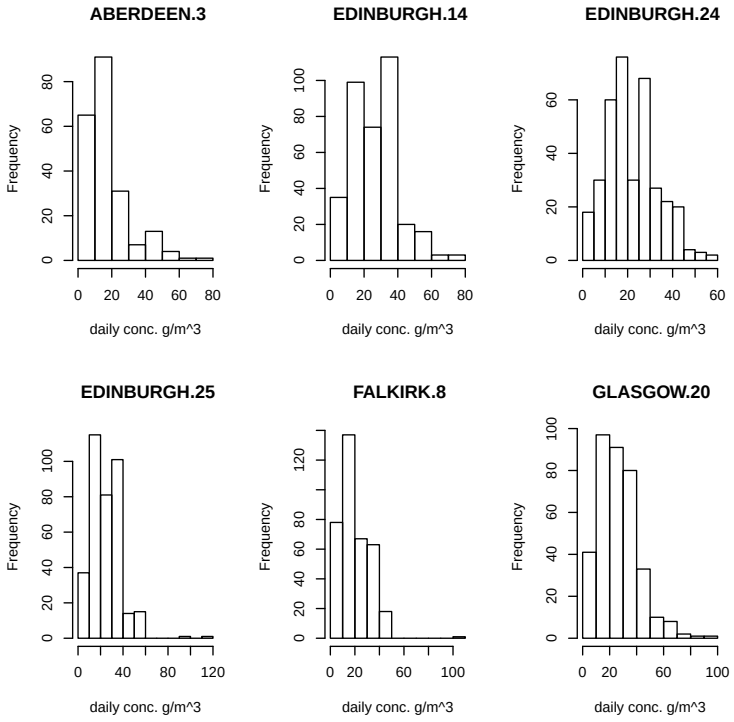


Figure 1.13: Histograms of daily $S0_2$ concentration in 2000

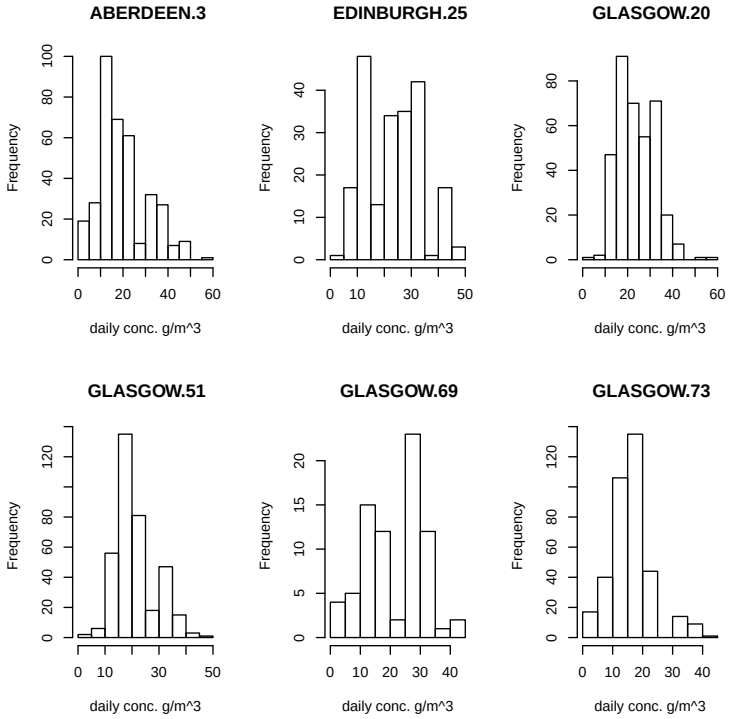
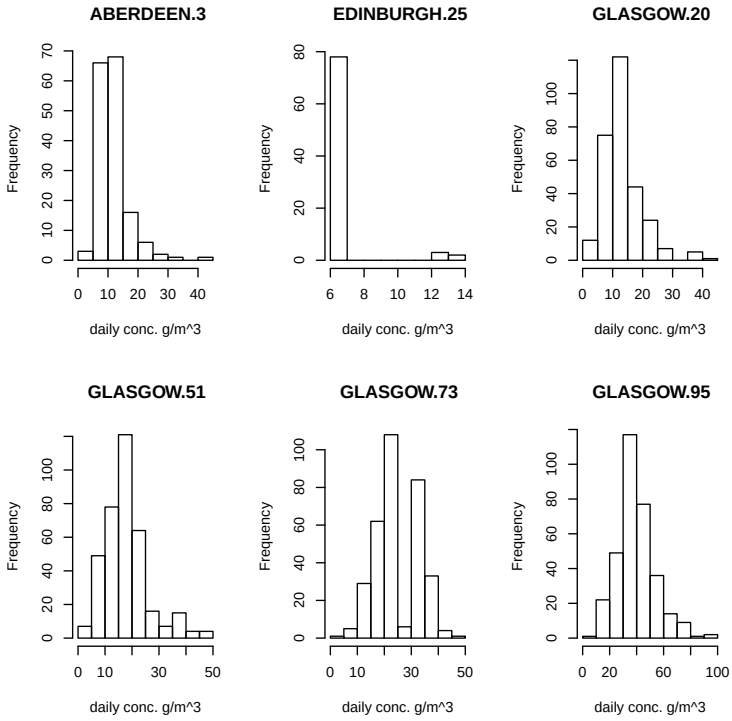



Figure 1.14: Histograms of daily $SO_2$ concentration in 2005



Considering trend within year, Figures 1.15-1.17 show boxplots of monthly $SO_2$ concentration over all sites combined in years 1996, 2000 and 2005 respectively. The boxplots for 1996 in Figure 1.15 have a steady increase in median level between February and July and decline between August and December. The median level is relatively very low with little variation from month to month. Summer months (May-August) are higher than the rest of the year. Outliers (unusual observations) are also very obvious which is an indication of presence of extreme observations.

In Figure 1.16, there is almost a constant median level of $SO_2$ between January and April 2000 though the relative location of the median is quite different from that of 1996 (where it was almost in the centre of the box). The median levels decrease between May and August, with a little rise in September before a jump down to a very low level in November. May has the highest median level.

The boxplots in 2000 are more generally widely varying in levels than those of 1996, which could be as a result of scanty observations. There are still outliers but not as many as in 1996. In Figure 1.17, the median level also rises from January to April and is almost constant between May and June before a rise again in August. August has the highest median level, in accordance with some of the earlier results that emphasize a summer peak. Thereafter, there is no obvious pattern for September-December. Generally, a cyclical pattern (periodic rise and fall in median levels) characterized $SO_2$ median level in 2005. Year 2005 has the least outliers as compared to other years, which can be attributed to very scanty observations with fewer recording stations operating.

In summary, there is a general variation in $SO_2$ median levels both within the year, in which summer months have a prominent peak, and across the years, in which data are more clustered together in 1996 than 2005. We have fewer observations in the later year (2005) as indicated by wider irregular fluctuations in median level in 2005 than 1996. Year 1996 has more prominent outliers than the other years because of clustering of data points and more extreme data. The patterns of fluctuation in levels for 2000 and 2005 are not as clear as that of 1996.

Figure 1.18 shows variance versus mean plots for daily $SO_2$ levels for all stations in 1996, 2000, 2002 and 2005. This allows inspection of the variance across stations. The top-left box represents the 1996 dataset in which we have 41 different recording stations. In this plot the data points are very clustered with a gradual rise in variance level as the mean increases. The top-right plot represents the year 2000 with 28 different recording sites and variance also increases with increase in mean level. The bottom-left and bottom-right plots correspond to 2002 and 2005

with 19 and 15 recording sites respectively. Variances also increase with increase in mean levels. Stations here are more widely scattered in the plots than in 1996 and 2000. The highest variance levels for 1996, 2000, 2002 and 2005 are about $800\mu g^2/m^3$, $1000\mu g^2/m^3$, $450\mu g^2/m^3$ and $300\mu g^2/m^3$ respectively.
In all cases, it can be seen that variance clearly increases with the mean level, so that a variance stabilising transformation is desirable and necessary, either through a square root or logarithmic transformation of $SO_2$ data before further modelling. This will be considered in the subsequent section before time series modelling and spatial analysis of this data.

Figure 1.15: Boxplots of monthly $SO_2$ concentration for all sites in 1996

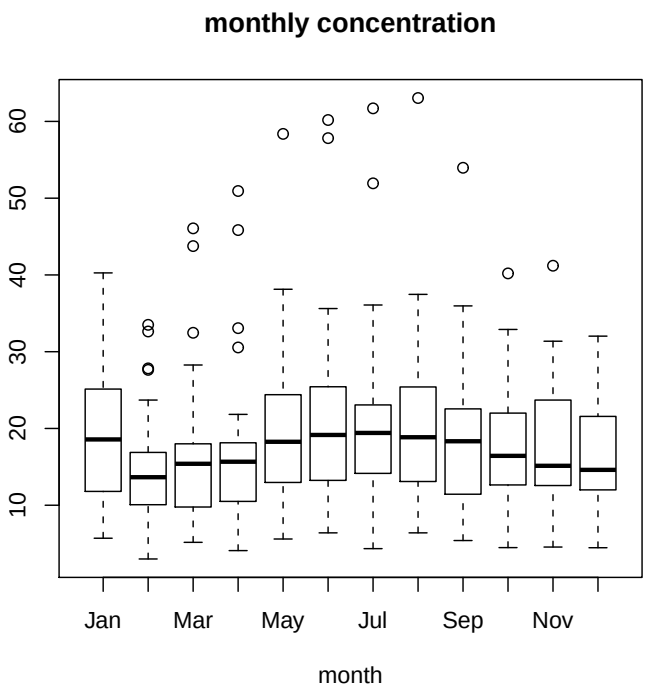


Figure 1.16: Boxplots of monthly $SO_2$ concentration for all sites in 2000

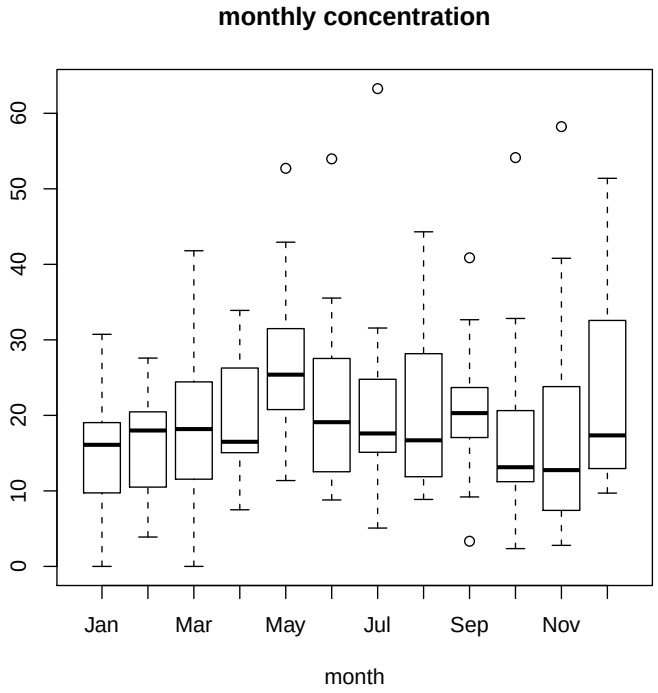



Figure 1.17: Boxplots of monthly $SO_2$ concentration for all sites in 2005

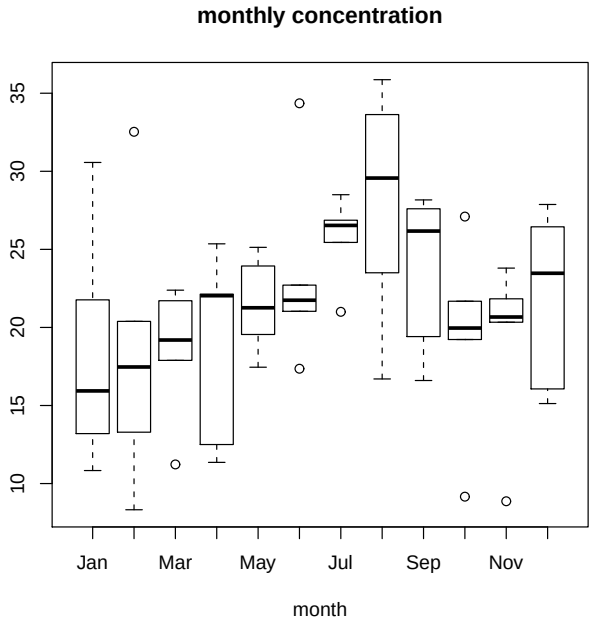
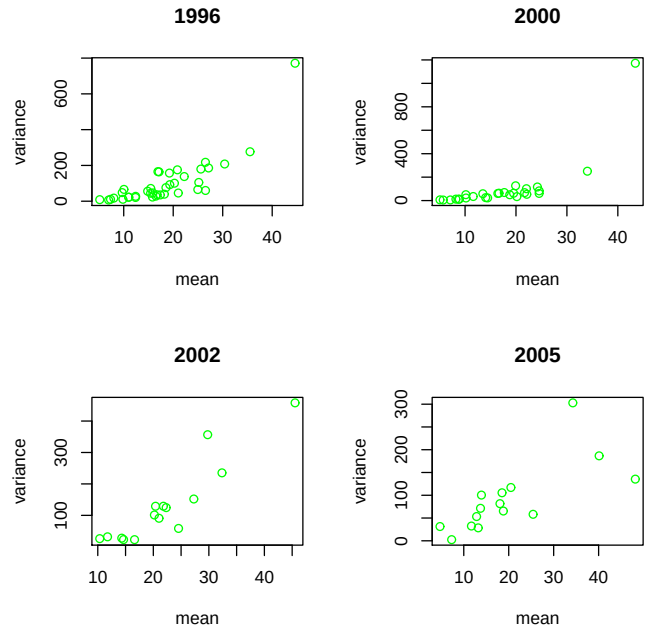


Figure 1.18: Plots of variance versus mean daily $SO_2$ levels over all station in a given year




## 1.6 Box-Cox transformation

Many spatial observations show a markedly non-Gaussian behaviour. The Box-Cox procedure to identify a suitable transformation is a standard method that is mostly used when a dataset contains outliers or when the dataset is not normally distributed. Many datasets in real life are skewed and sometimes have a heavy right tail, and in either situation the data is usually transformed to be approximately Gaussian distributed. Transformation of data is designed to achieve a specified purpose such as stability of variance, additivity of effects and symmetry of the density.

Let $Y$ be a random variable on the positive half-line, The Box-Cox transformation of $Y$ with power parameter $\lambda$ is defined as

$$Y^{\lambda} = \begin{cases} \frac{Y^{\lambda}-1}{\lambda} & : \quad \lambda \neq 0 \\ \log(Y) & : \quad \lambda = 0. \end{cases}$$

In the *geoR* package of R, the numerical value of the Box-Cox transformation parameter $\lambda = 1$ corresponds to no transformation. The parameter $\lambda$ is always regarded as fixed and data transformation is performed before the subsequent analysis. Prediction results are back-transformed and returned on the same scale

as for the original data. We do not use Box-Cox explicitly but the logarithmic transformation is considered later for our data in Chapters 3, 4 and 5, which corresponds to $\lambda = 0$ (Pengfei Li, 2005; De Oliveira and Ecker, 2002; De Oliveira et al. 1997; Box and Cox, 1964).

## 1.7 Conclusion

We generally observe that there is a variation in the levels of $SO_2$ both within the year (seasonal variation), i.e the concentration levels vary on a monthly basis with summer months usually having the highest concentration levels, and across the years, as seen in the boxplots of Figures 1.15-1.17 and time series plots in Figures 1.5-1.11. The variation is not consistent over the years as each year gives different variational patterns. Also long-term fluctuation is visible among the years in Figure 1.8.

The mean-variance plots in Figure 1.19 indicate that the variance increases with mean levels, thus will necessitate appropriate transformation in our modelling in subsequent chapters. Some of the histograms are right-skewed, which implies that the $SO_2$ data is not normally distributed in general.

Most of the stations are concentrated in Central Scotland, and there are no stations with available data in the North-Western part of Scotland where the population density is very low and there are low industrial activity, so there is low risk of high $SO_2$ concentration in this region. The years 1996 and 2003 have the lowest percentage of missing observations. Edinburgh 14 has the highest percentage of missing observations, while Glasgow 66 has the fewest number of observations missing (in Table 1.1), closely followed by Glasgow 51 and Glasgow 64.

Chapter 2 now discusses previous work on air pollution data and the sorts of models used to describe it.

# Chapter 2

# Literature review

## 2.1 Effects of air pollution, and its relationship with health

Chapter 1 discusses the general background on air pollution, its effects and relationship to health, while focusing on $SO_2$ data. Basic description of the $SO_2$ data with respect to its availability, monitoring locations and the missing structure of the data was also given.
This chapter reviews some of the previous work on adverse impacts of environmental pollutants on human health in Section 2.1, before discussing some of the various models used for air pollution data in Section 2.2.

### 2.1.1 Effects of air pollution on health

Numerous studies have investigated the health impact of environmental pollution and the adverse effects of airborne pollutants upon human health have been well established (Katsouyanni, 1997; Dockery, 1993 and 2006; Martuzzi, 2006; Baccarelli, 2008). One of the findings of the epidemiological studies was that there is a short term effect of pollutants on health, with emphasis on death and hospital admission (Brunekreef, 2002), and that based on several pollution studies, findings have also suggested that a temporal correlation exists between particulate matter and sulphur dioxide with acute increase in mortality.
$SO_2$ has been related with many adverse health impacts (Brunekreef and Holgate, 2002; Koren, 1995), including high mortality risk, diabetes related deaths, sudden infant death, heart diseases and bronchitis etc. (Biggeri et al., 2005).
Katsouyanni et al. (1997) assessed the relative risk of death and found that "in Western European cities an increase of $50\mu g/m^3$ in sulphur dioxide or black

smoke levels was associated with a 3% increase in daily mortality (95% confidence interval 2% to 4%), and was 2% (1% to 3%) for $PM_{10}$ (particles which have aerodynamic diameter less than $10\mu m$), while in Central and Eastern European cities the increase in mortality associated with a $50\mu g/m^3$ change in sulphur dioxide was 0.8% (-0.1% to 2.4%) and in black smoke 0.6% (0.1% to 1.1%)". They also observed that the pollutant effects were usually stronger during the summer months.

Burnett et al. (1995) found that a "$13\mu g/m^3$ increase in sulphates recorded on a day before a patient is admitted in hospital ($95^{th}$ percentile) was associated with a 3.7% ($p \leq 0.0001$) and a 2.8% ($p \leq 0.0001$) increase in both the respiratory and cardiac admissions respectively".

Many papers also concentrate on the short-term health impacts of environmental pollution (Schwartz, 2001; Samet et al., 2000 and Biggeri et al., 2004) rather than the long-term effects studies such as Pope et al. (2002), Dockery et al. (1996), Yap et al. (2006), Skalpe (1964) and Beeson et al. (1998).

It was estimated that 3400-5700 people in the Netherlands died prematurely through short term exposure to air pollution in 2003, of which 1/3 is caused by ozone and 2/3 by particulate matter (PM) (Fischer et al., (2004) in Stein and Dekkers et al., (2006)).

Dockery et al. (2006) confirmed that there have been recent developments in the assessment of health effects of pollutants at different times and scales of exposure in various locations, and that there has also been new evidence of inter-relationship between PM and health effects and emerging evidence of a general relationship between PM exposure and mortality and cardiopulmonary morbidity.

Sunyer et al. (1996) reported on relationship between daily air pollution levels and emergency admissions for chronic obstructive pulmonary and asthmatic diseases in Barcelona. They were able to assess the inter-relationship of environmental pollutant and mortality using autoregressive Poisson regression models that incorporate temperature, relative humidity and variables relating to temporal and autoregressive patterns. Their findings were that black smoke and $SO_2$ were related to mortality and that there is a growing link between $SO_2$ and mortality during summer.

Beeson et al. (1998) examined the long-term interconnection between atmospheric pollutants and risk of lung cancer in non-smokers, and assessment of risk of incident cancer was also analysed for ozone, $PM_{10}$ and $SO_2$ pollutants. The findings established a substantial link between high risk of lung cancer and $PM_{10}$

and $SO_2$.
From these studies we conclude that increase in $SO_2$ levels is related to increase in ill health. We next consider and discuss several air pollution models adopted by various authors.

## 2.2 Models used for air pollution data

From the previous section we can see that sulphur dioxide is one of the major pollutants that has negative impacts on health. Most environmental data are usually characterized by missing observations, and it makes data analysis more difficult. Thus, there is no generally acceptable or universal approach to adopt while analysing data of this nature.
Many authors utilize different methods and various statistical approaches. Different models have been used to analyse pollution data, from geostatistical kriging for spatial interpolation, Bayesian modelling, generalized additive models, spatio-temporal modelling etc. Some of the available literature on air pollution modelling techniques are reported below.

### 2.2.1 Kriging models

Geostatistical interpolation, otherwise known as kriging, has been applied in many spatial applications (Cressie, 1993; Chiles and Delfiner, 1999). There has been a substantial increase in application of kriging to air quality data, usually to interpolate measurements at unknown locations. There is a considerable emphasis on spatio-temporal interpolation (Wikle et al., 1998; Kyriakidis and Journel, 1999; Huerta et al., 2004). We used kriging in Chapter 4 of this thesis for spatial interpolation of $SO_2$ levels in Scotland.
Armstrong and Jabin (1981) described the method of estimating and modelling of variograms and also discussed certain mathematical conditions which a variogram or a covariance function must satisfy. One of their findings was that the shape of the variogram model for shorter distances is most important for usage in kriging. Thus in the sample variogram modelling, the values of the sample variogram for longer lag distances are not necessarily needed once a reasonable shape at shorter distance is obtained. They evaluated the sample variogram graphically to visually display closeness of experimental variograms to the theoretical model for the same lag distance, and model performance in terms of kriging was examined.
Cressie (1993) discussed general approaches and techniques for modelling spatial data based on varying assumptions about which parts of the model could be ref-

ered to as a fixed trend or come from spatial covariance with random effects and other random parameters, while Hobert et al. (1997) used a similar approach but focused on the modelling of $SO_2$. Handock and Stein (1993) worked on Bayesian kriging by predicting a Gaussian random field in a way that includes the uncertainty parameter in the covariance function. The analysis was based on the best linear unbiased prediction procedure using a Bayesian network. The formulated model was later applied on topographical data.

De Oliveira et al. (1997) on the other hand extended the findings of Handcock and Stein (1993) by using transformed Gaussian random fields with a parametric family of monotone transformations model for prediction. The Bayesian transformed Gaussian method is another form of trans-Gaussian kriging incorporating major sources of uncertainty in the predictive density estimation. Unlike the trans-Gaussian approach, it adopted a more robust predictive approach and employed the median as the optimal predictor because the predictive mean distribution does not exist for most common transformations.

Pilz et al. (1997) worked on a prediction problem by deriving a minimax approach which accounted for the uncertainty in variogram selection and computation rather than concentrating on a simple estimated variogram, while Griffith and Layne (1999) presented graphical, numerical and empirical findings that help to show the association between geostatistics and spatial autoregression.

Holland et al. (2000) compared the performance of generalized additive models with the Bayesian method. The trends and their standard errors are estimated in a two way procedure. Firstly, a generalized additive model is fitted to $SO_2$ data to compute the magnitude of the site-specific trend by incorporating the meteorology and season in the analysis. A stationary normal random variable was also adopted for the site dependent measurement error from the estimated trend. Kriging methodology was later utilized in the constructon of spatially smoothed estimates of the true trend. The last part of the analysis involves the use of Bayesian analysis with the Markov Chain Monte Carlo (MCMC) method, to estimate the regional trends and their standard errors.

Wikle and Royle (2002) utilized the mixed model framework method for characterizing spatial statistics. They focused on the classical geostatistical approach, kriging, in which they adopted a linear mixed model. Modification was also made to the generalized linear mixed model to incorporate non-Gaussian spatial processes. They later used the mixed model framework to describe multivariate spatial models and many spatio-temporal models.

Diggle et al. (2003) gave an introduction to model-based geo-statistics by for-

mulating the general modelling network for geo-statistical problems. Spatial prediction was also considered within the framework. Kammann and Wand (2003) also used the kriging method for spatial prediction. Ruppert et al. (2003) also considered this approach widely.
Pilz et al. (2005) utilized a non-parametric variogram function that gives a minimum mean square error prediction. They adopted another Bayesian procedure to predict the model uncertainty by means of reliable posterior probability distribution for the parameters. Matern covariance functions were used in the variogram computation. They later examined various parameter estimation methods including weighted least squares, generalized least squares, maximum-likelihood, REML and Bayesian techniques. The method was later applied on log-transformed normally distributed data.
Pilz and Spock (2007) derived another form of Bayesian kriging which incorporates uncertainty in covariance estimation and efficient normal transformation of the data. The method utilized the sample estimate of covariance and Box-Cox transformation. They also constructed spatial sampling design and optimal planning of recording stations in order to get better and reliable spatial results, by approximating the spatial process using a linear regression model with uncorrelated errors.

### 2.2.2 Bayesian models

Xia and Carlin (1998) used a mixture of two approaches by combining the methods of spatial-temporal mapping with covariate error treatment in a Bayesian hierarchical model. The posterior distribution of the model parameters was estimated using the MCMC method. The model was applied to data on lung cancer rates.
Host (1999) presented a statistical method for estimating national emissions of air pollutants across European countries by formulating models which used $SO_2$ in a spatial linear regression model with the measured deposition as a response and the emissions as independent variables within a Bayesian framework .
Little and Rubin (2002) discussed widely on classes of missing data and also considered a Bayesian MCMC procedure for missing data models. This procedure was used in fitting a Bayesian logistic regression to model and analyse environmental pollution data with missing observations. They also observed that the procedure performed very well, especially for modelling data with relatively few parameters, and that the number of parameters which can be specified are unlimited for this method, even though the MCMC procedure may be computationally

costly for missing data and might also take a longer period of time for its computation.

MacNab (2004) proposed a Bayesian spatial regression model that was used to estimate the rate of accident and injury. The results demonstrated how Bayesian modelling techniques can be used in a risk assessment. She later presented a general modelling framework that enables in depth investigations into relationship between injury rates and associated variables. The method was applied to $SO_2$ data. The model result was effective in verification of national compliance with emissions standards, and the method can also be extended to a wide range of other pollutants rather than only $SO_2$.

Riccio et al. (2006) presented a hierarchical Bayesian methodology. The model was applied to ozone data and also used to validate the CAMx (Comprehensive Air Quality Model with Extensions) result. Fasso et al. (2007), on the other hand, suggested a hierarchical model as a suitable solution to the dynamics of spatio-temporal modelling of environmental data, again using the EM algorithm. They illustrated how the algorithm can be combined with a parametric bootstrap for the evaluation of the parameter estimation through computation of standard errors and confidence intervals.

### 2.2.3 Spatio-temporal models

Handcock and Wallis (1994) presented a spatio-temporal model of winter temperature data by making use of separate spatial analysis in each year and a Gaussian random field, while Li et al. (1999) worked on spatial-temporal models for ambient hourly $PM_{10}$, and found that there was a constant temporal pattern across the monitoring sites for $PM_{10}$ observed in the Vancouver area, by adopting a common temporal correlation structure for all the recording stations in the model.

Also, Smith (2003) analysed and decomposed spatio-temporal data into deterministic non-parametric factors of time and space and applied the result for spatial interpolation. The EM algorithm was used for missing data imputation. The result was used to predict $PM_{2.5}$ levels.

In a similar study, Huerta et al. (2004) worked on the spatio-temporal modelling of ozone levels in Mexico City, using a dynamic linear model framework to calculate prediction values for ozone levels. The model incorporated spatial covariance functions for the observations and harmonic component parameters. Romanowicz et al. (2006) presented spatial-temporal interpolation of environmental pollution data by using time series analysis, and also showed how a non-stationary time series procedure could be used to analyse the data. The methodology was applied

to nitrogen oxide concentration.

Yap et al. (2006) worked on the risk assessment of long term exposure to air pollution on mortality by developing a spatio-temporal model for the estimation of cohort exposure to air pollution. Several methodologies were used, ranging from multiple imputation for missing data, multilevel and mixed effect modelling.

Fan et al. (2008) worked on spatio-temporal modelling of ambient $SO_2$ concentrations and proposed a modification to the kernel mixing method using a dynamic linear spatio-temporal model, which improves model flexibility in handling temporal and spatial data. Spatial and temporal autocorrelation are analysed by fitting semivariograms to determine the *range*, the *sill* and the *nugget*. These are three parameters that define the semivariogram. The nugget is a measurement error and is represented by the intercept of the variogram, the range is a constant that determines the degree of correlation between data, always represented as a distance, while the sill is the value of the semivariance as lag $k$ tends to infinity (Cressie, 1993); these are also discussed in Chapter 4. For each monthly dataset, the map of interpolated regions was also provided in the study.

Several other studies reported modelling of pollutants using other methods. Schwartz et al. (2001) assessed the short term health impacts of pollutants on daily mortality. Fixed and random effect models were utilized to obtain the combined individual city regression coefficients. Sensitivity analysis was also invesigated to examine the effect of the chosen statistical models. The study indicated that "an increase in $SO_2$ concentration of $50\mu g/m^3$ was associated with 2.2% increase in mortality when analysis was applied to data for days with $S0_2$ concentration of less than $200\mu g/m^3$". This method was described as a more rigid approach than generalized additive models.

Fuentes et al. (2006) quantified the effects of $PM$ exposure on mortality. They used the best available spatial $PM_{2.5}$ information from monitoring networks to better estimate the increase in rate of mortality as $PM$ level increases. A generalized Poisson regression (GPR) model was used. The GPR model discards the standard Poisson model but allows for both under- and over-dispersion to model mortality data. They also carried out a spatial-temporal analysis to identify the main constituents of the $PM$ mixture that are the most significant in causing mortality. The GPR model was not considered in this thesis because we do not have any observation related to number of counts or rate for which GPR is most appropriate.

Yanosky et al. (2008) assessed chronic $PM_{10}$ exposure by fitting a monthly smooth spatial term and smooth regression term of GIS-derived and meteorolog-

ical predictors, using a cross-validation approach and other useful pre-specified selection criteria. Recently, Sabah et al. (2008) used an artificial neural network to predict the ground levels of $SO_2$ for determination of meteorological factors that can affect $SO_2$ concentration.

### 2.2.4 Other spatio-temporal models

Wood and Augustin (2002) presented the basic mathematical and numerical approaches to generalized additive models implemented in the *mgcv* package of *R*, and they demonstrated the methods by illustrating with two different environmental data examples.

Bowman and Azzalini (2002) gave general theoretical ideas about the computational aspects of non-parametric smoothing with illustrations using the *sm* library in *R*. This is basically a computationally efficient procedure for analysis of a very large dataset with many evaluation points or multivariate data. They provided an efficient matrix formulation of non-parametric smoothing and also derived a modified binning approach for the data when the sample is very large.

Roca-Padinas et al. (2004) predicted binary time series data of $SO_2$ concentration by using a generalized additive model that has an unknown link function. The methodology involves incorporating a non-parametric estimation of the link function by a local scoring algorithm. The non-parametric estimation stage involves using a local linear kernel smoother, and because of high computational cost, the binning technique was also adopted to hasten the computation. One of the findings was that this model performs better than those with transformed binary regression.

Roca-Padinas et al. (2005) also considered a generalized additive model which involves a second-order interaction term by using a local scoring algorithm in combination with local linear kernel smoothers for parameter estimation. This technique is similar to the approach in Roca-Padinas et al. (2004). A bootstrap procedure was utilized for estimation rather than the backfitting method, which is difficult because of high computational cost, and a binning procedure was also incorporated to speed up computation. The method was applied on $SO_2$ data.

Mentzakis and Delfino (2005) examined the effects of pollution and climate parameters on human health, and sensitivity of their results was tested under different model specifications, which involve linear models and generalized linear models as well as generalized additive models.

Also, Bowman et al. (2009) considered spatio-temporal modelling of $SO_2$ using an additive model procedure. Basic description of the correlation pattern of the

dataset was carried out and a binning technique was derived and adopted (as in Bowman and Azzalini (2002)) to provide an effective computation of the backfitting algorithm, because of the large sample size. A three-dimensional smoothing technique was also developed and applied to $SO_2$ data collected over European countries to model the interaction among the space, time and seasonal effects. Along the same lines, Terzi and Cengiz (2009) utilized a generalized additive model procedure in multiple Poisson regression for modelling the relationship between air pollution and increases in hospital admissions for respiratory diseases. The model was demonstrated on an $SO_2$ dataset.

The following authors also adopted generalized additive models in different areas to analyse their data Azadeh and Salibian-Barrera (2009); Yee and Mitchell (1991); Hastie and Tibshirani (1986); Kauermann and Opsomer (2000); Olsson and Oard (2008); O'Brien and Rago (1996); Lopez-Moreno and Nogues-Bravo (2005); Jackson et al. (2008); Murase et al. (2009); Wang et al. (2009).

## 2.3 Conclusion

Having reviewed the literature on health impact of air pollution and various models that could be used to analyse air pollution data, we conclude that air pollution really has adverse effects on human health. We are going to use kriging and spatial GAM models to investigate spatial heterogeneity, as our data consists of measurements at 41 locations. With both of these methods we will be able to interpolate $SO_2$ in regions where there is no station and provide an $SO_2$ map of Scotland. We will not pursue any links to health data in this thesis though it is an obvious extension to this work.

We now base our subsequent analysis, which involves the imputation of missing data and time series analysis of $S0_2$ data, on the ideas and knowledge we obtained from this review. The next chapter will introduce imputation methods and time series analysis of $SO_2$ data, before kriging and spatial GAM modelling are considered in Chapters 4 and 5 respectively. We adoted a gam spatial method in our analysis because it has been used to analyse similar $SO_2$ level over European countries by Bowman et al. (2009) and the kriging which is also an efficient method for obtaining optimal spatial prediction for our $SO_2$ was also used by Hobert et al (1997) and Handock and Stein (1993) to also analyse similar data.

# Chapter 3

# Imputation methods and time series analysis of $SO_2$ data

## 3.1 Missing value and multiple imputation

Chapter 2 reviews previous work on air pollution and its effects on human health, before discussing some of the models used for air pollution data. This chapter focuses on methods for imputation of missing data, then applies these to the $SO_2$ data using the time series methodology. The chapter also describes some theoretical background, and some elements of data description software in $R$ and $SPSS$.

Section 3.1 discusses the missing value and multiple imputation techniques, focusing on missing data classifications, EM, regression and MICE procedures.

Section 3.2 applies different imputation methods to the $SO_2$ data, and also compares the performance of each of the imputation techniques, using time series plots and box plots. Section 3.3 describes time series data, correlograms and parameter estimation. Section 3.4 describes different time series modelling techniques such as $AR$, $MA$, $ARMA$, $ARIMA$ and time series decomposition as well as diagnostic checking of residual models. Section 3.5 presents the model results and Section 3.6 gives the conclusion.

### 3.1.1 Missing value analysis

Missing data is usually a source of problems in statistical modelling. Proper handling of missing values is important in all analyses, for instance in time series analysis of environmental data which is usually characterized by missing observations. The missing data can be attributed to equipment failure or malfunctioning

and an inability of equipment to measure below certain threshold values and uncontrollable external factors, and administrative decisions to stop recording for a period of time. The $SO_2$ data we are considering in this thesis have many missing values.
Data with missing observations usually causes a serious challenge to data analysis. Many modelling techniques remove completely observations with missing values from the analysis. The two methods for analysing missing observations in $SPSS$ are missing value analysis and multiple imputation (MI) procedures. The central aim of MI is to obtain several possible values for missing data ($SPSS$, 17.0).

### 3.1.2 Missing data classifications

Let $Z$ be a matrix of observed value, and $Y$ a matrix of missingness indicator response with $Y_i = 1$ if the $i^{th}$ element of $Z$ is missing, and 0 otherwise, with parameters $\theta$. The missing completely at random ($MCAR$) assumes that the missing pattern is not associated with any variables either known or unknown, i.e data are said to be MCAR when the missing pattern does not depend on observed or unobservable quantities and is defined as

$$P(Y|Z) = P(Y|Z^{obs}, Z^{mis}, \theta) = P(Y|\theta). \tag{3.1}$$

Data are said to be missing at random ($MAR$) when the missing pattern does not depend on the unobserved data ( it may depend on observed data). Little's MCAR test can be used to determine if dataset are actually missing at random or not (see section 3.1.5). The assumption of MAR is that

$$P(Y|Z) = P(Y|Z^{obs}, \theta) = P(Y|\theta). \tag{3.2}$$

(Little & Rubin, (2002); Horton & Lipsitz, (2001); www.washington.edu/uware/spss/).

### 3.1.3 Methods for multiple imputation in SPSS and R

Here, we examine multiple imputation techniques and assumptions that it requires, as well as review some software packages that implement this procedure. MI involves imputing missing values several times using a correct model specifications, by incorporating random variation in a dataset. MI introduces appropriate random error into the imputation process to make it possible to obtain approximately unbiased estimates of all parameters and repeated imputation often allows

one to get better estimates of the standard errors. The various imputation models that can be used include *predictive mean matching*, *logistic regression* or *propensity* and $MCMC$ (Markov Chain Monte Carlo).
The *predictive mean matching* and $MCMC$ procedures require assumptions of normality for the data to be imputed. The inference is still robust to little deviations from this assumption (Schafer, 1997; Rubin, 1996 & 1977). *Predictive mean matching* (*pmm*) uses a linear regression model assumption for the distribution of incomplete data conditional on other variables. Let variable $Y_j$ have missing values. Then a model can be fitted using completed observations for $Y_1, \ldots, Y_{j-1}$, as

$$E[Y_j|\varphi] = \varphi_0 + \varphi_1 Y_1 + \varphi_2 Y_2 + \ldots + \varphi_{j-1} Y_{j-1}. \tag{3.3}$$

On the $i^{th}$ iteration, the parameters $\varphi^i$ are sampled from a normal distribution for the parameters and the missing values are then replaced by

$$Y_j^i = \varphi_0^i + \varphi_1^i Y_1 + \varphi_2^i Y_2 + \ldots + \varphi_{j-1}^i Y_{j-1} + \sigma^i \varepsilon, \tag{3.4}$$

where $\sigma^i$ is the variance estimate from the model, and $\varepsilon$ is a simulated $N(0,1)$ random variate.
The *pmm* is based on assumption of a linear regression model, while the *propensity* method utilizes a logistic regression model for the missing pattern indicators. The $MCMC$ method, on the other hand, simulates samples from the posterior density $f(Y^{mis}|Y^{obs})$. This can be computed by the method proposed by Schafer (1997) in which at the $j^{th}$ iteration the imputation step draws $Y^{mis,(j+1)}$ from $f(Y|Y^{obs}, \varphi^{(j)})$ and the parameter estimation step draws $\varphi^{(j+1)}$ from $f(\varphi|Y^{obs}, Y^{mis,(j+1)})$
The Markov chain consists of

$$\left(\left\{Y^{(1)}, \varphi^{(1)}\right\}, \left\{Y^{(2)}, \varphi^{(2)}\right\}, \ldots, \left\{Y^{(j+1)}, \varphi^{(i+1)}\right\}, \ldots\right). \tag{3.5}$$

The $MCMC$ method handles data with arbitrary missing patterns but is based on normality assumptions, and it is complicated as well as computationally expensive (Horton & Lipsitz, 2001).

### 3.1.4 Missing value analysis

Missing value analysis as available in $SPSS$ describes the pattern of missing data, and investigates if data have values missing in multiple cases or are missing randomly. Missing value techniques include listwise, pairwise, regression (described

above) and Expectation-Maximization methods. We focus mainly on regression and EM described below.

Missing value analysis addresses several problems caused by incomplete data. Missing data reduces the precision and accuracy of calculated statistics because important information may be lost in the missing data, and many statistical assumptions are based on complete observations rather than missing data, which complicates the required theory.

Missing value analysis is mostly based on the assumption that the pattern of missing values does not depend on the data values. If the data are not missing completely at random, then an EM estimation technique may be most suitable (Little and Rubin, 1987; *SPSS* 17.0).

### 3.1.5 Expectation Maximization method in SPSS

The EM estimation method is based on the assumption that the pattern of missing data is dependent on observed data only, i.e missing at random. The method assumes a distribution for the incomplete data and bases inferences on the likelihood under that distribution. Each iteration consists of $E$ and $M$ steps. The $E$ step computes the conditional expectation of the missing data, given the observed values and current estimates of the parameters. These expectations are then substituted for the missing value. In the $M$ stage, maximum likelihood estimates of the parameters are calculated with the assumption that the missing data have been completed.

A missing completely at random (MCAR) test is usually computed by Little's chi-square statistic. The null hypothesis is that the data are missing completely at random. The test statistic is

$$d^2 = \sum_{i=1}^{N} d_i^2 = \sum_{i}^{N} M_i(\bar{Y}_i - \hat{\mu}_i)\hat{\Sigma}^{-1}(\bar{Y}_i - \hat{\mu}_i)^T, \tag{3.6}$$

which has an asymptotic $\chi^2$ distribution with $\sum_i^N (P_i - P)$ degrees of freedom, where $P$ is the number of variables, $N$ is the number of patterns of missing values from a possible $2^P$, while $\hat{\mu}$ and $\hat{\Sigma}$ are the maximum likelihood estimates of the parameters of a $p$-dimensional multivariate normal distribution based on the available data, $M_i$ is the the number of observations with the $i^{th}$ missingness pattern, $P_i$ is the number of complete observations in pattern $i$, $\hat{\mu}$ and $\hat{\Sigma}^{-1}$ are the subsets of the parameters corresponding to complete observations for pattern $i$ (they are vector of length $P_i$ and a matrix of dimension $P_i X P_i$, respectively)

and $\bar{Y}_i$ is the $P_i$- dimensional sample average of the observed data in pattern $i$. If the null hypothesis is rejected (when $d^2$ is large), then either the data is missing at random (MAR) or not missing at random (NMAR) but is not MCAR (Little and Rubin, 1988; Hesterberg, 1999).

## Regression Method in SPSS

The method calculates multiple linear regression estimates by adjusting the estimates with random components. For each predicted value in the iteration procedure a residual from a randomly selected complete variable can be added to the results or added from a random normal deviate or from a random deviate (scaled by the square root of residual mean square) from the t-distribution. This adds an optional random part to a regression estimate of a missing value ($SPSS$ 17.0).

## MICE in R

Another method for complex incomplete data is Multivariate Imputation by Chained Equations, provided in the MICE package of R. This generates multiple imputation, analyses imputed data and pools analysis results (Rubin, 1987 & 1996).

The main features of this package are columnwise specification of the imputation model, the ability to deal with arbitrary patterns of missing data, passive imputation, subset selection of predictors, support of arbitrary complete data methods, support of pooling various types of statistics, diagnostics of imputations and a callable user-written imputation function package.

The methods in MICE include predictive mean matching (pmm), Bayesian linear regression, linear regression, unconditional mean imputation, two-level logistic regression (logreg) and unordered polytomous regression (polyreg).

The method of predictive mean matching is related to the regression method, except that for each missing observation an observed value is imputed with a closest value to the predicted value from the simulated regression model (Rubin, 1987). The method assumes that imputed data are more reliable and may also perform better than the regression method when there is little departure from the normality assumption (Horton & Lipsitz, 2001, Van Buuren et al, 2006).

## 3.2 Comparison of imputation modelling performance on $SO_2$ data

Before we proceed with time series modelling of the $SO_2$ data it is imperative for us to explore the appropriate imputation technique for the analysis. In the basic description of the data in Chapter 1, Table 1.1 describes the missing data structure, while Figures 1.4-1.13 show the time-series patterns.

We found out that our data has a considerable number of missing observations in some stations and time periods, thereby necessitating the use of an appropriate imputation technique. In order to provide a common and universal set of data for modelling of $SO_2$, the missing data was first imputed. This procedure was carried out year by year on the whole of the $SO_2$ pollution dataset as some of the time series modelling techniques to be considered (for instance, the autoregressive modelling technique (AR) and autoregressive integrated moving average (ARIMA)) are not able to handle missing data.

The replacement was based on a missing value analysis using both regression and EM imputation in SPSS as well as MICE in R (explained in section 3.1.5). We first log-transformed the data before imputing the missing observations, this is done to prevent negative imputed values (from results of EM and regression methods) which are not possible for $SO_2$ in reality, as the values must be non-negative.

We next compare the performances of each imputation method using the descriptive summary of the imputed data, time series plots, boxplots, and various time series models using different datasets from each method of imputation. In the MICE procedure, 5 multiple imputations were used.

The MCAR test results in Table 3.1 are significant for the datasets in 1996-2005, and this suggests that the data are not missing completely at random for these years, therefore, the EM method is more appropriate to impute the missing data in this category. The test is not significant for the datasets in 2006 and 2007, which may be attributed to the lower number of recording stations and less missing data in these two years which give their test much lower degrees of freedom compared to the other years.

Table 3.1: MCAR test output for $SO_2$ datasets from 1996-2007

| year | Little's MCAR test: |
|---|---|
| 1996 | Chi-Square = 5594.621, DF = 3989, Sig. = .000 |
| 1997 | Chi-Square = 5098.684, DF = 4799, Sig. = .001 |
| 1998 | Chi-Square = 4082.835, DF = 3099, Sig. = .000 |
| 1999 | Chi-Square = 1464.308, DF = 852, Sig. = .000 |
| 2000 | Chi-Square = 1679.393, DF = 849, Sig. = .000 |
| 2001 | Chi-Square = 1679.393, DF = 849, Sig. = .000 |
| 2002 | Chi-Square = 1036.448, DF = 611, Sig. = .000 |
| 2003 | Chi-Square = 1091.382, DF = 621, Sig. = .000 |
| 2004 | Chi-Square = 1192.956, DF = 619, Sig. = .000 |
| 2005 | Chi-Square = 1134.128, DF = 451, Sig. = .000 |
| 2006 | Chi-Square = 9.094, DF = 12, Sig. = .695 |
| 2007 | Chi-Square = 17.140, DF = 18, Sig. = .513 |

Descriptive statistics for the 1996, 2000 and 2005 datasets after imputation by the EM, regression and MICE methods are shown in Tables 3.2-3.4. In Table 3.2, the minimum imputed data for all the three years is 3. The maximum value for EM is 234, which occurs at Armadale 2 (before imputation, the maximum recorded concentration is $229g/m^3$ in Table 1.3 in Chapter 1). In Tables 3.3 and 3.4, the maximum value for both regression and MICE is 190 and occurs in Kirkintilloch 9 for both methods. The regression and MICE methods are similar but EM is very different.

The minimum recorded observations is 3 for all three methods. The corresponding maximum values are 120, 118 and 118 for EM, regression and MICE respectively, and these occur at Coatbridge 11 for the three of them. The mean pattern for the EM method is generally higher than for the other two methods, while that of regression is also higher than for MICE. Similarly to the 1996 dataset, there is no obvious pattern of the standard deviations.

Maximum values are 169, 108 and 106 for EM, regression and MICE respectively,

which occur in Glasgow 98 for the EM method and Glasgow 95 for both regression and MICE methods. We observe that the MICE method produce higher mean and standard deviation than the other two methods except for Aberdeen 3 and Glasgow 98.

In summary, the three methods show similar results, especially regression and MICE. The mean level generally decreases across the year in accordance with our conclusion from Chapter 1 that $SO_2$ mean level decreases with year.

Having imputed the missing values, our subsequent analysis involves comparison of imputation methods using both daily time series plots and box plots for each method. We consider the 1996, 2000, and 2005 datasets. The stations in the subsequent analysis are randomly selected for each of these years.

Table 3.2: Comparison of descriptive statistics for EM imputed data

| | EM | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1996 | 2000 | 2005 | 1996 | 2000 | 2005 | 1996 | 2000 | 2005 | 1996 | 2000 | 2005 |
| station | Min | | | Max | | | Mean | | | Sd | | |
| ABERDEEN 3 | 3 | 3 | 10 | 45 | 57.5 | 68 | 17.76 | 26.57 | 44.38 | 6.72 | 11.3 | 6.85 |
| EDINBURGH 14 | 3 | | | 5 | | | 0.4 | | | 0.7 | | |
| EDINBURGH 24 | 3 | | | 62 | | | 20.63 | | | 1.7 | | |
| EDINBURGH 25 | 3 | 3 | 3 | 5 | 51.5 | 41 | 0.4 | 26.54 | 2.21 | 0.7 | 8.95 | 3.91 |
| FALKIRK 8 | 12 | | | 83 | | | 15.85 | | | 11.84 | | |
| GLASGOW 20 | 3 | 3 | 3 | 84 | 58.5 | 78 | 27.51 | 24.11 | 18.08 | 14.32 | 7.93 | 9.47 |
| GLASGOW 51 | 3 | 3 | 3 | 63 | 51.5 | 55 | 22.6 | 23.91 | 16.09 | 12.38 | 7.42 | 7.15 |
| GLASGOW 69 | 3 | 3 | | 117 | 52.5 | | 25.99 | 18.44 | | 14.1 | 8.25 | |
| GLASGOW 73 | 3 | 3 | 3 | 109 | 44.5 | 57 | 19.66 | 18.67 | 10.46 | 13.23 | 7.9 | 7.23 |
| GLASGOW 95 | 3 | 3 | 3 | 101 | 50.5 | 101 | 26.91 | 20.77 | 31.71 | 15.42 | 8.04 | 16.02 |
| GLASGOW 98 | 3 | 3 | 5 | 73 | 47.5 | 169 | 21.2 | 3.749 | 9.49 | 13.95 | 7.16 | 24.65 |
| HAMILTON 5 | 11 | | | 56 | | | 26.58 | | | 7.54 | | |
| HATTON 1 | 3 | 17 | | 18 | 8.46 | | 5.59 | 16.05 | | 3.58 | 1.1 | |
| BALLINGRY2 | 9 | 9 | | 39 | 33.5 | | 16.94 | 11.16 | | 5.99 | 5.11 | |
| KILMARNOCK 2 | 9 | 7 | | 19 | 20.5 | | 10.27 | 11.58 | | 3.4 | 2.74 | |
| KINROSS 1 | 6 | | | 31 | | | 11.39 | | | 5.18 | | |
| KIRKCALDY 6 | 9 | 5 | 6 | 44 | 31.5 | 9 | 17.83 | 12.1 | 5.27 | 6.01 | 4.21 | 0.21 |
| KIRKINTILLOCH 8 | 9 | 3 | 3 | 74 | 77.5 | 37 | 15.52 | 15.28 | 9.73 | 8.61 | 6.97 | 6.15 |
| KIRKINTILLOCH 9 | 3 | 9 | | 107 | 75.5 | | 19.7 | 21.46 | | 10.28 | 7.47 | |
| KIRKINTILLOCH 10 | 10 | 10 | 3 | 102 | 75.5 | 44 | 25.45 | 7.02 | 15.55 | 10.18 | 7.72 | 8.13 |
| LIVINGSTON 1 | 3 | | | 37 | | | 10.2 | | | 7.19 | | |
| LONGSIDE 1 | 3 | 3 | | 52 | 13.5 | | 12.9 | 19.83 | | 5.85 | 1.75 | |
| ORMISTON 2 | 9 | | | 57 | | | 15.94 | | | 8.57 | | |
| PENICUIK 1 | 9 | 9 | | 48 | 45.5 | | 21.46 | 10.55 | | 7.42 | 7.01 | |
| PERTH 1 | 9 | | | 47 | | | 18.62 | | | 6.92 | | |
| PETERHEAD 1 | 3 | 6 | | 48 | 14.5 | | 12.8 | 7.61 | | 5.25 | 2.42 | |
| PETERHEAD 2 | 3 | 3 | | 23 | 14.5 | | 7.73 | 10.3 | | 3.88 | 1.63 | |
| PETERHEAD 3 | 3 | 3 | | 55 | 16.5 | | 8.4 | 13.7 | | 4.86 | 2.68 | |
| PRESTONPANS 1 | 7 | | | 43 | | | 11.31 | | | 4.58 | | |
| STIRLING (BURGH) 5 | 3 | 9 | | 48 | 39.5 | | 16.77 | 22.17 | | 6.13 | 5.67 | |
| TRANENT 2 | 7 | | | 23 | | | 7.36 | | | 2.93 | | |
| WHITBURN 3 | 3 | 10 | 3 | 44 | 40.5 | 45 | 10.44 | 2 | 23.14 | 8.75 | 5.92 | 6.5 |
| BO'NESS 2 | 9 | 3 | | 82 | 1.5 | | 18.86 | 36.05 | | 9.07 | 0.1 | |
| BONNYRIGG 1 | 3 | | | 36 | | | 15.84 | | | 7.27 | | |
| ARMADALE 2 | 3 | 16 | 10 | 234 | 118 | 94 | 45.8 | 23.59 | 37.92 | 27.91 | 15.6 | 12.14 |
| ARMADALE 3 | 15 | | | 65 | | | 15.36 | | | 12.01 | | |
| COATBRIDGE 5 | 3 | 3 | 3 | 131 | 71.5 | 56 | 29.78 | 38.98 | 17.92 | 14.67 | 7.33 | 9.81 |
| COATBRIDGE 11 | 3 | 5 | 6 | 134 | 120 | 105 | 36.6 | 16.43 | 10.86 | 16.26 | 30.5 | 26.4 |
| COWDENBEATH 1 | 3 | 9 | 7 | 39 | 28.5 | 22 | 16.15 | 27.2 | 10.36 | 5.23 | 4.89 | 4.16 |
| DALKEITH 1 | 3 | 9 | | 60 | 50.5 | | 25.41 | 18.4 | | 8.68 | 7.46 | |

Table 3.3: Comparison of descriptive statistics for regression imputed data

| | REG | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1996 | 2000 | 2005 | 1996 | 2000 | 2005 | 1996 | 2000 | 2005 | 1996 | 2000 | 2005 |
| station | Min | | | Max | | | Mean | | | Sd | | |
| ABERDEEN 3 | 4 | 3 | 20 | 38 | 56 | 78 | 16.37 | 20.4 | 22 | 6.926 | 11 | 9.6 |
| EDINBURGH 14 | 3 | | | 3 | | | 16.34 | | | 0.54 | | |
| EDINBURGH 24 | 5 | | | 59 | | | 20.66 | | | 10.262 | | |
| EDINBURGH 25 | 4 | 3 | 5 | 3 | 55 | 23 | 12.34 | 24.9 | 2.97 | 0.49 | 10 | 4.47 |
| FALKIRK 8 | 13 | | | 80 | | | 16.22 | | | 12.337 | | |
| GLASGOW 20 | 3 | 3 | 7 | 81 | 56 | 55 | 27.44 | 25 | 17.22 | 13.825 | 7.8 | 8.817 |
| GLASGOW 51 | 4 | 3 | 3 | 60 | 53 | 62 | 22.45 | 22.6 | 17.14 | 11.918 | 7.4 | 6.904 |
| GLASGOW 69 | 3 | 3 | | 114 | 56 | | 25.93 | 22.2 | | 13.615 | 10 | |
| GLASGOW 73 | 3 | 3 | 3 | 106 | 43 | 43 | 19.59 | 16.9 | 11.51 | 12.735 | 7.8 | 7.387 |
| GLASGOW 95 | 3 | 3 | 3 | 98 | 49 | 108 | 26.81 | 17.2 | 32.6 | 14.957 | 7.9 | 16.031 |
| GLASGOW 98 | 5 | 3 | 8 | 70 | 45.1 | 43 | 21.17 | 19.3 | 9.42 | 13.437 | 7.1 | 4.866 |
| HAMILTON 5 | 9 | | | 53 | | | 26.8 | | | 7.94 | | |
| HATTON 1 | 3 | 3 | | 56 | 38 | | 5.52 | 7.57 | | 3.171 | 2.3 | |
| BALLINGRY2 | 9 | 3 | | 36 | 36 | | 16.87 | 14.6 | | 5.516 | 5 | |
| KILMARNOCK 2 | 6 | 3 | | 19 | 19 | | 10.25 | 9.48 | | 3.341 | 3.5 | |
| KINROSS 1 | 8 | | | 28 | | | 11.4 | | | 4.903 | | |
| KIRKCALDY 6 | 9 | 3 | 5 | 141 | 30 | 16 | 17.62 | 10.4 | 5.79 | 5.789 | 4.6 | 0.363 |
| KIRKINTILLOCH 8 | 9 | 3 | 3 | 71 | 49 | 44 | 15.36 | 10.8 | 10.48 | 7.919 | 7 | 6.114 |
| KIRKINTILLOCH 9 | 3 | 3 | | 190 | 74 | | 19.65 | 13.8 | | 9.818 | 7.5 | |
| KIRKINTILLOCH 10 | 10 | 3 | 9 | 99 | 75 | 51 | 25.32 | 20 | 16.45 | 9.803 | 7.7 | 7.832 |
| LIVINGSTON 1 | 8 | | | 34 | | | 10.18 | | | 7.243 | | |
| LONGSIDE 1 | 3 | 3 | | 49 | 13 | | 12.83 | 5.73 | | 5.403 | 2.5 | |
| ORMISTON 2 | 6 | | | 54 | | | 15.71 | | | 8.334 | | |
| PENICUIK 1 | 9 | 3 | | 45 | 44 | | 21.36 | 18.8 | | 6.995 | 8.1 | |
| PERTH 1 | 9 | | | 44 | | | 18.54 | | | 6.445 | | |
| PETERHEAD 1 | 3 | 3 | | 45 | 17 | | 12.71 | 9.14 | | 4.767 | 3.3 | |
| PETERHEAD 2 | 7 | 3 | | 20 | 13.4 | | 7.69 | 6.22 | | 3.439 | 2.2 | |
| PETERHEAD 3 | 3 | 3 | | 52 | 19 | | 18.34 | 8.73 | | 4.377 | 3.4 | |
| PRESTONPANS 1 | 5 | | | 40 | | | 11.5 | | | 4.767 | | |
| STIRLING (BURGH) 5 | 3 | 3 | | 45 | 38 | | 17.07 | 12.2 | | 6.123 | 5.8 | |
| TRANENT 2 | 4 | | | 20 | | | 7.36 | | | 2.723 | | |
| WHITBURN 3 | 3 | 3 | 3 | 41 | 39 | 52 | 10.33 | 20.7 | 24.15 | 8.262 | 5.8 | 6.469 |
| BO'NESS 2 | 3 | 3 | | 79 | 3 | | 18.84 | 3.5 | | 8.938 | 1 | |
| BONNYRIGG 1 | 5 | | | 33 | | | 15.87 | | | 6.823 | | |
| ARMADALE 2 | 3 | 3 | 3 | 178 | 112 | 101 | 45.43 | 34.3 | 38.61 | 27.702 | 16 | 12.551 |
| ARMADALE 3 | 3 | | | 62 | | | 17.42 | | | 12.282 | | |
| COATBRIDGE 5 | 4 | 3 | 5 | 128 | 70 | 63 | 30.65 | 22.4 | 18.68 | 14.678 | 8.2 | 9.693 |
| COATBRIDGE 11 | 6 | 3 | 5 | 131 | 118 | 48 | 36.49 | 38.6 | 13.35 | 17.097 | 33 | 8.784 |
| COWDENBEATH 1 | 3 | 3 | 8 | 36 | 27 | 29 | 15.93 | 14.9 | 11.25 | 5.013 | 4.8 | 4.129 |
| DALKEITH 1 | 7 | 3 | | 34 | 48 | | 25.2 | 25.3 | | 8.228 | 8.4 | |

Table 3.4: Comparison of descriptive statistics for MICE imputed data

| | MICE | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1996 | 2000 | 2005 | 1996 | 2000 | 2005 | 1996 | 2000 | 2005 | 1996 | 2000 | 2005 |
| station | Min | | | Max | | | Mean | | | Sd | | |
| ABERDEEN 3 | 4 | 3 | 20 | 38 | 52 | 76 | 16.03 | 19.9 | 23.6 | 6.716 | 10.9 | 10.8 |
| EDINBURGH 14 | 3 | | | 6 | | | 16 | | | 0.33 | | |
| EDINBURGH 24 | 5 | | | 59 | | | 20.32 | | | 10.052 | | |
| EDINBURGH 25 | 4 | 3 | 5 | 8 | 55 | 21 | 12 | 24.4 | 4.57 | 0.28 | 9.75 | 5.67 |
| FALKIRK 8 | 13 | | | 80 | | | 15.88 | | | 12.127 | | |
| GLASGOW 20 | 3 | 3 | 7 | 81 | 57 | 53 | 27.1 | 24.5 | 18.82 | 13.615 | 7.53 | 10.017 |
| GLASGOW 51 | 4 | 3 | 3 | 60 | 50 | 60 | 22.11 | 22.1 | 18.74 | 11.708 | 7.06 | 8.104 |
| GLASGOW 69 | 3 | 8 | | 114 | 56 | | 25.59 | 21.7 | | 13.405 | 10 | |
| GLASGOW 73 | 3 | 3 | 3 | 106 | 48 | 41 | 19.25 | 16.4 | 13.11 | 12.525 | 7.5 | 8.587 |
| GLASGOW 95 | 3 | 3 | 3 | 98 | 49 | 106 | 26.47 | 16.7 | 34.2 | 14.747 | 7.64 | 17.231 |
| GLASGOW 98 | 5 | 3 | 8 | 70 | 46 | 41 | 20.83 | 18.8 | 11.02 | 13.227 | 6.76 | 6.066 |
| HAMILTON 5 | 9 | | | 53 | | | 26.46 | | | 7.73 | | |
| HATTON 1 | 3 | 3 | | 56 | 38 | | 5.18 | 7.07 | | 2.961 | 1.97 | |
| BALLINGRY2 | 9 | 3 | | 36 | 32 | | 16.53 | 14.1 | | 5.306 | 4.71 | |
| KILMARNOCK 2 | 6 | 3 | | 19 | 19 | | 9.91 | 8.98 | | 3.131 | 3.17 | |
| KINROSS 1 | 8 | | | 28 | | | 11.06 | | | 4.693 | | |
| KIRKCALDY 6 | 9 | 3 | 5 | 141 | 30 | 14 | 17.28 | 9.93 | 7.39 | 5.579 | 4.26 | 1.563 |
| KIRKINTILLOCH 8 | 9 | 3 | 3 | 71 | 76 | 42 | 15.02 | 10.3 | 12.08 | 7.709 | 6.67 | 7.314 |
| KIRKINTILLOCH 9 | 3 | 3 | | 190 | 74 | | 19.31 | 13.3 | | 9.608 | 7.17 | |
| KIRKINTILLOCH 10 | 10 | 8 | 9 | 99 | 74 | 49 | 24.98 | 19.5 | 18.05 | 9.593 | 7.39 | 9.032 |
| LIVINGSTON 1 | 8 | | | 34 | | | 9.84 | | | 7.033 | | |
| LONGSIDE 1 | 3 | 3 | | 49 | 13 | | 12.49 | 5.23 | | 5.193 | 2.22 | |
| ORMISTON 2 | 6 | | | 54 | | | 15.37 | | | 8.124 | | |
| PENICUIK 1 | 9 | 10 | | 45 | 44 | | 21.02 | 18.3 | | 6.785 | 7.8 | |
| PERTH 1 | 9 | | | 44 | | | 18.2 | | | 6.235 | | |
| PETERHEAD 1 | 3 | 3 | | 45 | 17 | | 12.37 | 8.64 | | 4.557 | 3.03 | |
| PETERHEAD 2 | 7 | 3 | | 20 | 13 | | 7.35 | 5.72 | | 3.229 | 1.94 | |
| PETERHEAD 3 | 3 | 3 | | 52 | 17 | | 18 | 8.23 | | 4.167 | 3.15 | |
| PRESTONPANS 1 | 5 | | | 40 | | | 11.16 | | | 4.557 | | |
| STIRLING (BURGH) 5 | 3 | 3.5 | | 45 | 38 | | 16.73 | 11.7 | | 5.913 | 5.53 | |
| TRANENT 2 | 4 | | | 20 | | | 7.02 | | | 2.513 | | |
| WHITBURN 3 | 3 | 3 | 3 | 41 | 39 | 50 | 9.99 | 20.2 | 25.75 | 8.052 | 5.52 | 7.669 |
| BO'NESS 2 | 3 | 3 | | 79 | 4 | | 18.5 | 3 | | 8.728 | 0.3 | |
| BONNYRIGG 1 | 5 | | | 33 | | | 15.53 | | | 6.613 | | |
| ARMADALE 2 | 3 | 3 | 3 | 178 | 116 | 99 | 45.09 | 33.8 | 40.21 | 27.492 | 15.7 | 13.751 |
| ARMADALE 3 | 3 | | | 62 | | | 17.08 | | | 12.072 | | |
| COATBRIDGE 5 | 4 | 3 | 5 | 128 | 70 | 61 | 30.31 | 21.9 | 20.28 | 14.468 | 7.88 | 10.893 |
| COATBRIDGE 11 | 6 | 3 | 5 | 131 | 118 | 46 | 36.15 | 38.1 | 14.95 | 16.887 | 32.5 | 9.984 |
| COWDENBEATH 1 | 3 | 6 | 8 | 36 | 27 | 27 | 15.59 | 14.4 | 12.85 | 4.803 | 4.49 | 5.329 |
| DALKEITH 1 | 7 | 7 | | 34 | 51 | | 24.86 | 24.8 | | 8.018 | 8.11 | |

The time series plots in Figures 3.1-3.3 show the comparison of each imputed data for the daily mean $SO_2$ concentrations for stations in 1996, 2000 and 2005 respectively. We chose station that have fewer missing observations in each year. For each box, the upper panel represents EM in green, the middle panel is regression in blue, while the bottom is the MICE method in red.

In the 1996 plot in Figure 3.1, Falkirk 8 has the same pattern for all three imputation methods. Mice has has a double spike around days 130-140 which is not present in both regression and EM plots in Figure 3.1. Glasgow 20, Glasgow 51 and Glasgow 73 also have similar patterns for each of the three imputation methods respectively. Glasgow 20 around day 280. Glasgow 69 also has double spike around days 190-200. There is no prominet peak pattern observed in Glasgow 51. The patterns we observe here are similar to what we earlier observed in Figure 1.5 from Chapter 1, except that the gaps have been replaced with imputed data. Also, each station has a unique pattern, different from other stations.

In Figure 3.2, Glasgow 69 has the same pattern for both regression and EM methods, while the MICE imputation pattern is clearly different from the other two. Glasgow 73, Glasgow 95 and Glasgow 98 have similar patterns for each of the three techniques.

Lastly, in Figure 3.3, the EM method for Glasgow 20 is quite different from the other two methods, with a double spike around day 140 which is not present in the other methods. Glasgow 51 has a similar pattern for both EM and regression methods with a prominent peak around day 140, but MICE is different. Glasgow 73 has different patterns for each of the methods and EM has a double spike between days 120-130, and neither of the other two methods has this feature. The regression and MICE methods also show different patterns from each other. We observe that there is more prominent variation in levels after imputation in all of Figures 3.1-3.3. We conclude that the various imputation methods give different patterns (though some similarity are also observed).

Figure 3.1: Comparison of different imputation methods for the daily mean $SO_2$ concentrations for stations in 1996. The upper panel represents EM (green), the middle panel is regression (blue), while the bottom panel is MICE (red) imputation for each station

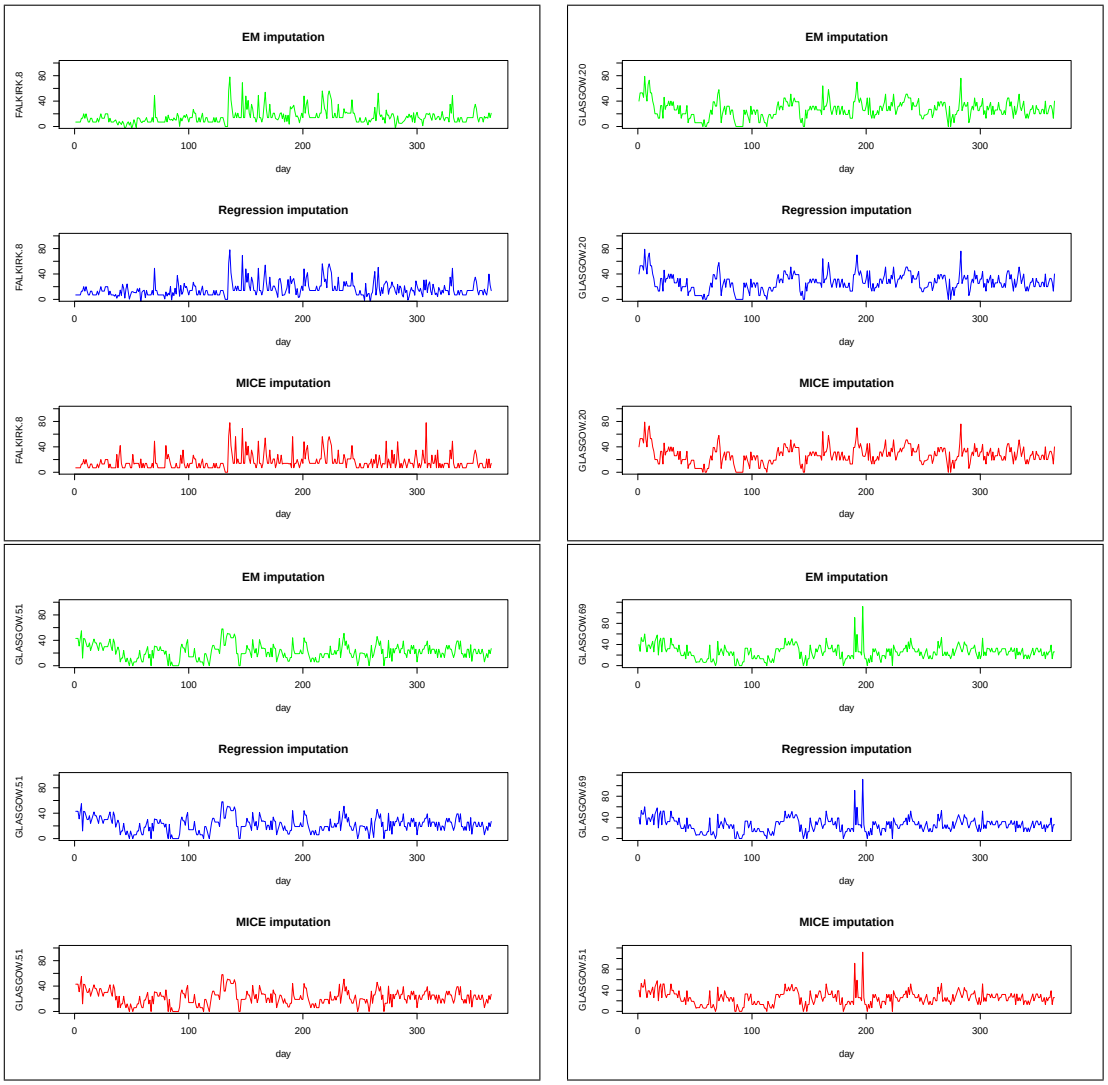


Figure 3.2: Comparison of different imputation methods for the daily mean $SO_2$ concentrations for stations in 2000. The upper panel represents EM (green), the middle panel is regression (blue), while the bottom panel is MICE (red) imputation for each station

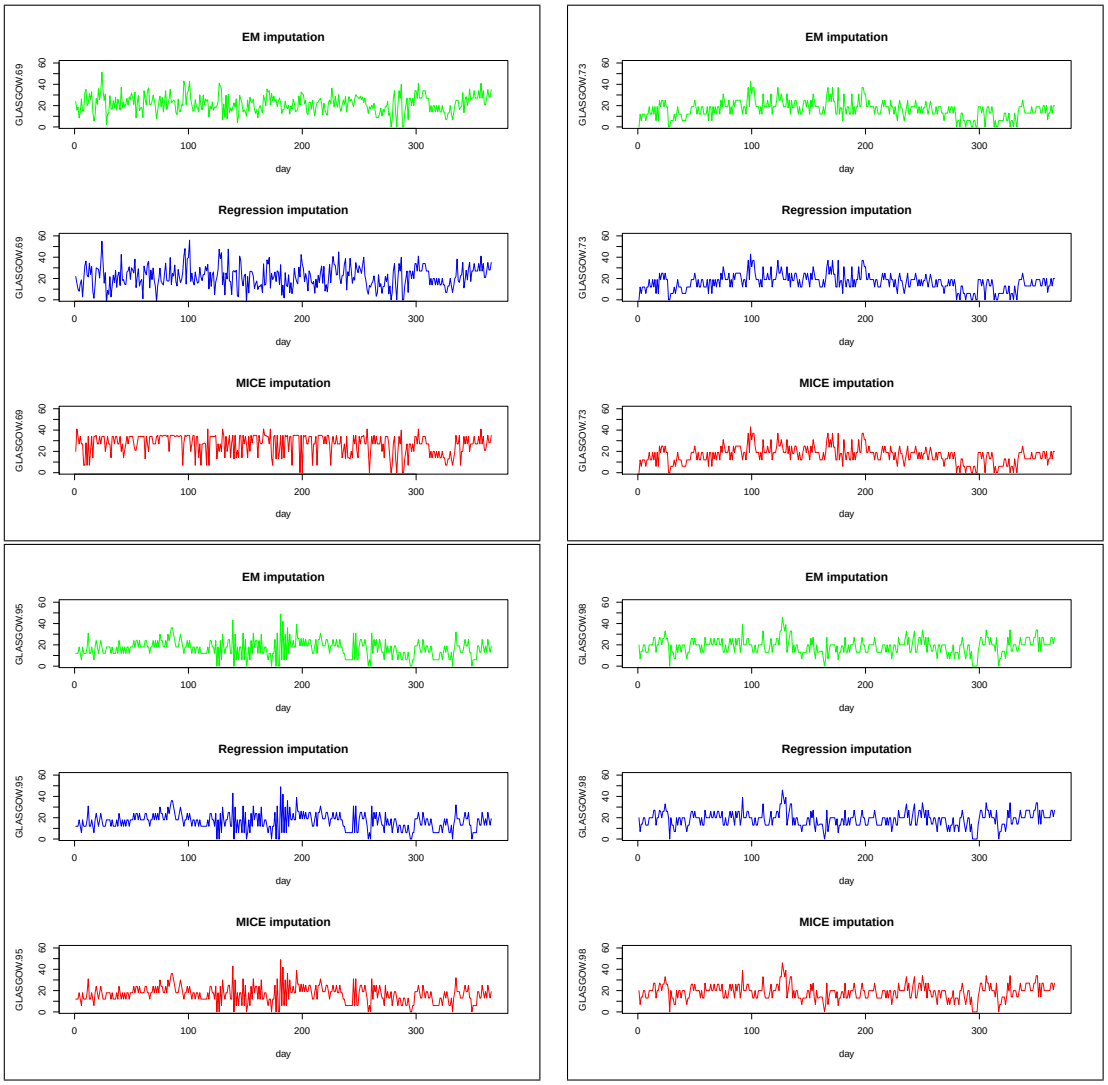
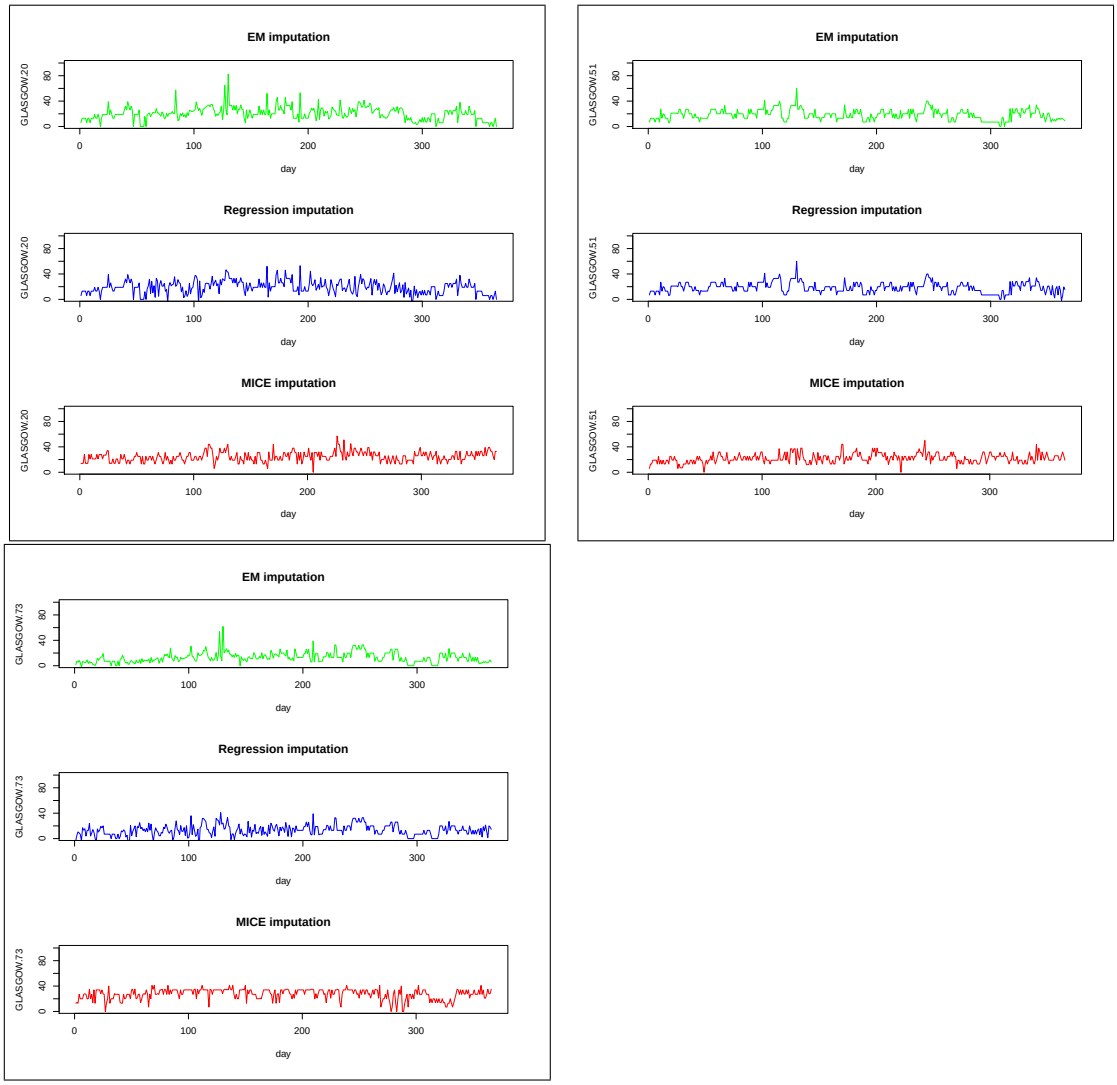


Figure 3.3: Comparison of different imputation methods for the daily mean $SO_2$ concentrations for stations in 2005. The upper panel represents EM (green), the middle panel is regression (blue), while the bottom panel is MICE (red) imputation for each station



The corresponding boxplots for comparison of different imputation methods for the logarithm of daily mean $SO_2$ concentration for those stations considered previously are shown in Figures 3.4-3.6 for some stations in 1996, 2000, and 2005 respectively. In each row, the first panel represents EM, the middle panel is regression while the third panel is MICE imputation.

For 1996 in Figure 3.4, Falkirk 8 has the same median level for the three methods. Glasgow 20, Glasgow 51 and Glasgow 69 have similar patterns for each method. Outliers are very prominent in Glasgow 20, Glasgow 51 and Glasgow 69 for all three methods, unlike Falkirk 8. The three methods give almost similar results.

For 2000 in Figure 3.5, outliers are still very prominent and all three methods also give similar results. MICE for Glasgow 69 has the highest median level about 3.5. The median is very close to the upper quartile in Glasgow 73, Glasgow 95 and Glasgow 98 for all the three methods.

For 2005 in Figure 3.6, MICE gives the highest median level for Glasgow 20. The median level is very close to the upper quartile for both EM and regression and to the lower quartile for MICE in Glasgow 51. In Glasgow 73, median level is very closer to the upper quartile for both EM and MICE than for regression. Outliers are more prominent for MICE in Glasgow 51 than any other stations. Overall, the median level here is generally lower than for 1996 and 2000. The three methods still give similar patterns.

In summary, we observe that outliers are more prominent in 2005 than in other years. Mostly, each of the stations has a varying pattern, and each of the methods gives rise to a different pattern, though some are similar. Generally, we see that there is a degree of similarity in the effect of the various types of imputation techniques. We still consider all three methods in our subsequent time series analysis in section 3.3.

Figure 3.4: Boxplots of comparison of different imputation methods for the logarithm of daily mean $SO_2$ concentrations for stations in 1996. In each row the first panel represents EM, the middle panel is regression, while the third panel is MICE imputation for each station

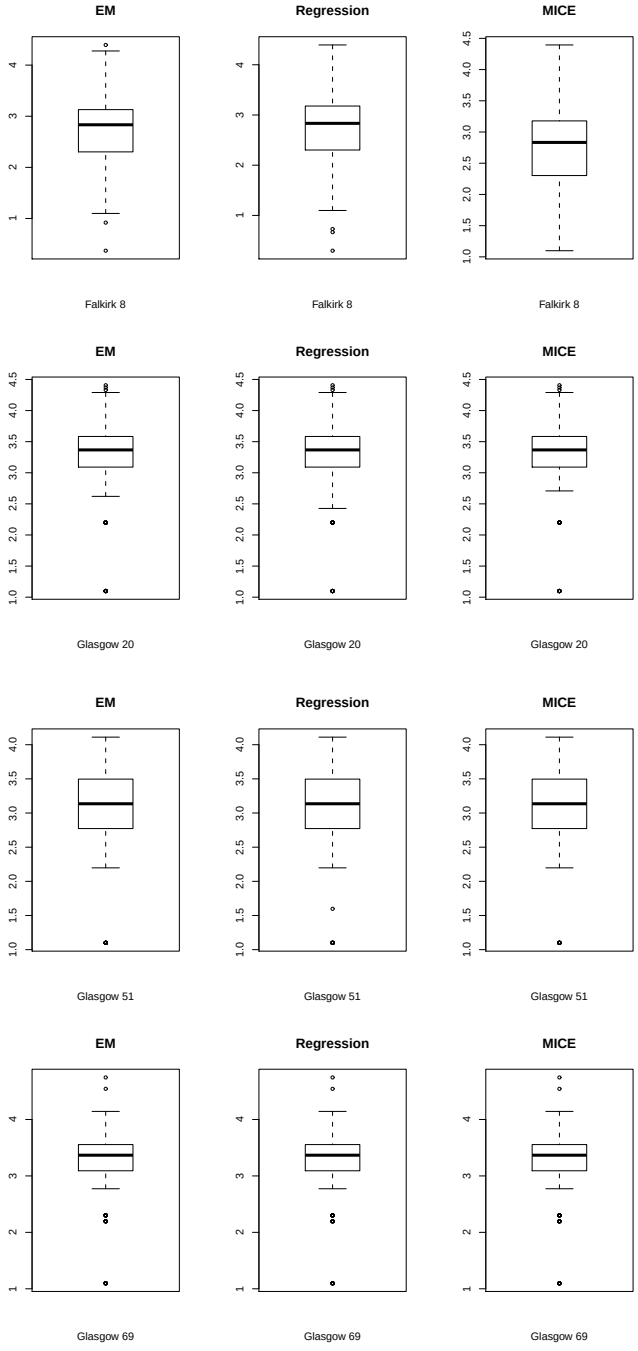


Figure 3.5: Boxplots of comparison of different imputation methods for the logarithm of daily mean $SO_2$ concentrations for stations in 2000. In each row the first panel represents EM, the middle panel is regression, while the third panel is MICE imputation for each station

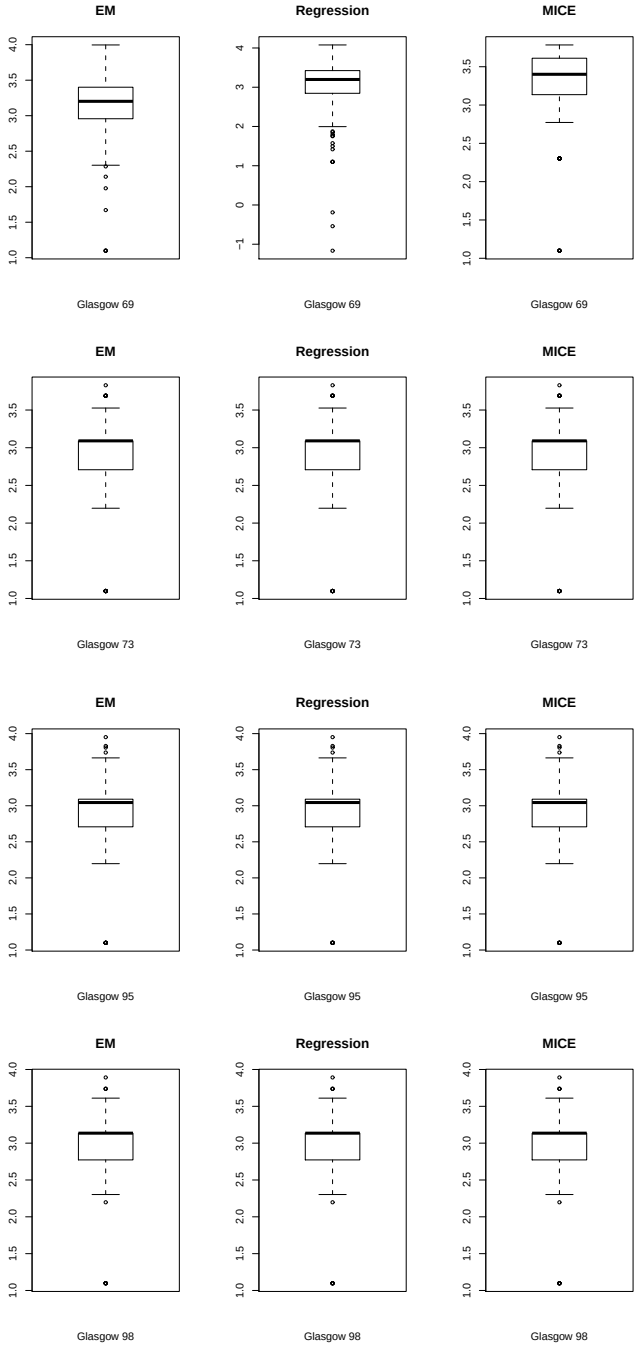
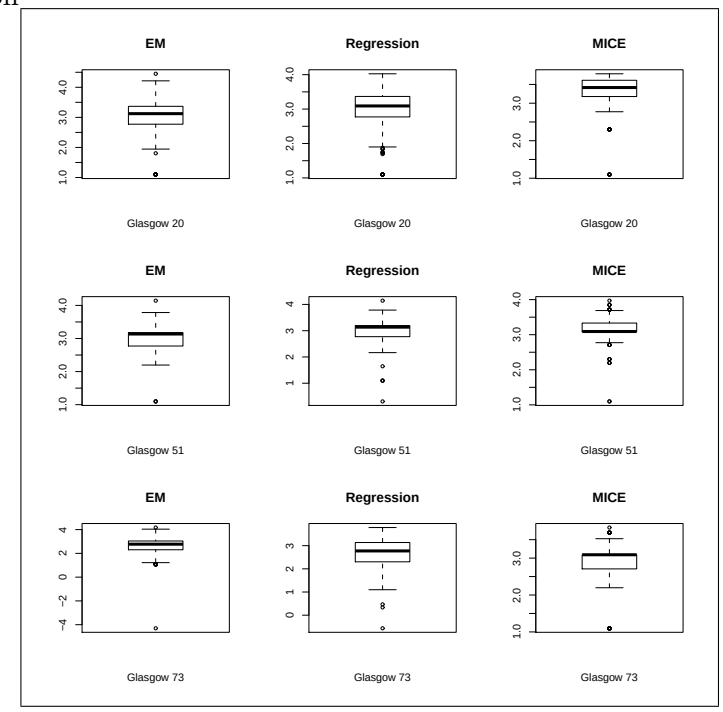


Figure 3.6: Boxplots of comparison of different imputation methods for the logarithm of daily mean $SO_2$ concentrations for stations in 2005. In each row the first panel represents EM, the middle panel is regression, while the third panel is MICE imputation for each station



## 3.3 Background on time series modelling techniques

The formulation of relevant and appropriate statistical models for temporal data is very important, relative to the exploration of the structure and shape of the data in Chapter 1. This section will build on the previous descriptive and exploratory analysis. We will consider data from 1996, 2000 and 2007 for few analysis. We will also combine the aggregated data from 1996-2007 for the last part of the analysis rather than individual year (for ARIMA models). We introduce year 2007 in this analysis rather than 2005 (considered in previous analysis) in order to make the extrapolation and forecasting more tractable.
The histogram of the raw data is right skewed, as seen from Figures 1.14-1.16, and Figure 1.20 in Chapter 1 also indicated that the variance increases with an increase in the mean level, thus variance stabilising transformation is necessary. Yap et al. (2006) observed a similar pattern for black smoke pollutants and used a logarithmic transformation, while Li et al. (1999) also observed the same pattern for $PM_{10}$ and also used a logarithmic transformation. Guttorp et al. (1994) and Carroll et al. (1997) used a square root transformation for ozone. We will use time series modelling of $log(SO_2)$ data in section 3.5, after describing the methods.

### 3.3.1 Time series data

A time series is an ordered sequence of observations with respect to time. Time series analysis often deals with how a variable such as daily $SO_2$ concentration varies over time, that is ordered sequences of measurements from a non-random pattern, unlike the analyses of random samples of observations that are discussed in the context of most other statistics.
We might be interested in how a measured variable changes on a daily, monthly, seasonal and yearly basis. One of the principal aims of time series is the investigation of the data generating process by inferring from what is observed to the underlying structure.
Two basic components of any time series data are trend and seasonality. Trend can be described as a systematic increasing or decreasing component that changes with time and does not usually repeat itself, while data with seasonality has a similar pattern that repeats itself in systematic intervals over time, that is exhibits periodic fluctuation. Some data also exhibit cyclical behaviour.
We are interested in identifying the nature of the phenomenon represented by

the sequence of $SO_2$ observations, checking and removing temporal correlation, estimating parameters of multivariate autoregressive (AR) and Autoregressive integrated moving average (ARIMA) models, diagnostic checking of fitted models and predicting future observations.

Box Jenkins (1976) described a basic approach for time series analysis which involves three different stages, namely identification, estimation and diagnostic checking. The identification stage is where a time series is visually inspected for stationarity. If it is not stationary, a series transformation may be necessary either by removing deterministic trend or taking first differences with respect to time. Variance stability may also be achieved by logarithmic transformation or other suitable transformation.

The autocorrelation function (ACF) and partial autocorrelation function (PACF) can be inspected to form a temporary $AR(p)$ or $ARMA(p, q)$ model. The estimation stage involves checking model stability and significance of the parameters, while the diagnostic stage involves examining the residuals for correlation and normality or evidence of (over- or under-) fitting of the model (Pfaff, 2008).

### 3.3.2 Correlograms

Time series analysis usually involves examination of the sample autocorrelation and partial autocorrelation functions in order to observe the functional form of the data, and to also check independence of the observations. The autocorrelation function (ACF) is obtained by computing autocorrelations for the observations at different time lags. The autocorrelation is the cross-correlation of the series with itself.

It is a statistical tool for finding repeating patterns, such as the presence of a periodic pattern. To plot the autocorrelation and partial autocorrelation functions, we used the *ts* package of R. The sample autocorrelation at lag $k$ is given as

$$\hat{\rho_k} = \frac{C_k}{C_0} \tag{3.7}$$

where $C_k$ is the sample autocovariance function at lag $k$,

$$C_k = \frac{1}{N} \sum_{i=1}^{N-k} (X_i - \bar{X})(X_{i+k} - \bar{X}), \tag{3.8}$$

which estimates the actual autocovariance function $\gamma_k = Cov(X_i, X_{i+k})$, assuming that the series is stationary. The sample size is represented by $N$, and $C_0$ is

the variance function

$$C_0 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2, \tag{3.9}$$

and $-1 \leq \gamma_k \leq +1$. The autocorrelation function is

$$\rho_k = \frac{\gamma_k}{\gamma_0}, \tag{3.10}$$

where $\gamma_0 = Var(X_i)$.

The partial autocorrelation at lag $k$ is the autocorrelation between $X_t$ and $X_{t-k}$ that is not accounted for by lags 1 through $k-1$. The pacf() function in R is used to plot the PACF which is another method for identifying serial dependencies and the order of autoregressive model.
The sample autocorrelation plot is used to determine if an autoregressive ($AR$) model will be appropriate for time series modelling, then a sample partial autocorrelation plot is examined to help identify the order. We look for the point on the plot where the partial autocorrelations essentially become zero. The significance of the test is determined by the lag that falls outside the 95% CI line (dotted line on the plot) (Chatfield, 1989; Box et al., 1994), see section 3.4.1.

### 3.3.3 Estimating model parameters

The Maximum Likelihood and Least squares methods are two common techniques for parameter estimation. The maximum likelihood method is used when the distribution of residual terms (white noise) is known.
The least squares method employs a regression equation, and assumes that the error variance in the measurement of each case is identical. The residual variance around the regression line is the same across all values of the independent variables (Gebhard et al., 2008). The least squares method will be used for subsequent $AR$, modelling while the maximum likelihood method will be employed for $ARIMA$ modelling of the $S0_2$ data in this chapter.

#### Maximum likelihood method

For any independent and identically distributed random variables $X_1, \ldots, X_P$ with probability density function $f(x_t, \theta)$, $t = 1, \ldots, P$ and parameter vector $\theta$, the joint density function is the product of the marginal densities,

$$f(x, \theta) = f(x_1, \ldots, x_P, \theta) = \Pi_{t=1}^{P} f(x_t, \theta). \tag{3.11}$$

Viewed as a function of the parameter(s) $\theta$ given data $x$, this is referred to as the likelihood function, and

$$L(\theta|x) = L(\theta|x_1, \ldots, x_t) = \Pi_{t=1}^{P} f(x_t, \theta). \tag{3.12}$$

The log-likelihood function is

$$\log L(\theta|x) = \sum_{t=1}^{P} \log f(x_t, \theta). \tag{3.13}$$

The method of maximum likelihood estimates $\theta$ by obtaining the value of $\theta$ that maximizes $L(\theta|x)$. This is the MLE of $\theta$

$$\hat{\theta}_{\text{mle}} = \arg\max_{\theta \in \Theta} \ell(\theta \,|\, x_1, \ldots, x_P). \tag{3.14}$$

### Least squares method

If we are fitting a model to data involving errors or deviations $\varepsilon$ from the model, it may be that $var(\varepsilon) = \sigma^2 I$, but it is also possible that the residuals have non-constant variance or are correlated. In that case let $var(\varepsilon) = \sigma^2 \Sigma$ where $\sigma^2$ is unknown but $\Sigma$ is known, which means that we know the correlation and relative variance between the errors. The generalized least squares method estimates the parameter $\beta$ to minimize

$$(y - x\beta)^T \Sigma^{-1} (y - x\beta), \tag{3.15}$$

giving

$$\beta = (x^T \Sigma^{-1} x)^{-1} (x^T \Sigma^{-1} y) \tag{3.16}$$

(Faraway, 2002).

## 3.4 Time series models

Here, we give basic theoretical background on time series modelling techniques, namely the $AR(p)$, $MA(q)$, $ARMA(p, q)$ and $ARIMA(p, d, q)$ models.

### 3.4.1 Models

## Autoregressive AR(p) models

An autoregressive model $AR$ is a simple linear statistical model for stationary time series data is represented by an AR($p$) model

$$X_t = \mu + \sum_{i=1}^{p} \alpha_i X_{t-i} + \varepsilon, \tag{3.17}$$

where $\alpha_i$ are the autoregressive model parameters, $\mu$ is the mean or intercept, $p$ is the order of the $AR$ process and $X_t$ is the value at time $t$ of the series under investigation. The error term or residual, represented by $\varepsilon$, is always assumed to be Gaussian white noise. A phenomenon is called white noise if it satisfies the following conditions:

$$E(\varepsilon_t) = 0, \qquad E(\varepsilon_t^2) = \sigma^2, \qquad E(\varepsilon_t \varepsilon_\tau) = 0, \quad t \neq \tau. \tag{3.18}$$

Using this model, the current term of the series can be estimated by a linear weighted sum of previous terms in the series in which the weights are the autoregression coefficients. $AR$ analysis derives the "best" values for $\alpha_i$ given a series $X_t$.

The $AR(1)$ sample autocorrelation function has an exponentially decaying appearance. However, higher-order $AR(p)$ are usually a combination of sinusoidal and exponential functions.

We can determine the order of higher order autoregressive processes by exploiting both autocorrelation and partial autocorrelation plots. The partial autocorrelation of an $AR(p)$ data is usually 0 at lag $\geq (p+1)$.

The sample partial autocorrelation function can also be examined to see if there is any departure from 0 at any lag $k$. This could be achieved by adding a 95% confidence interval to the plot of sample partial autocorrelation function, it is $\pm 2/\sqrt{N}$, where $N$ denotes the sample size (Box et al., 1994). If the order of the $AR(p)$ is unknown, AIC (Aikake's Information Criterion) can be used to determine the order and a low value indicates a better fitting model. This is defined as

$$AIC = \log \frac{1}{T} \sum_{t=1}^{T} (\hat{\mu}_t^p)^2 + n\frac{2}{T}, \tag{3.19}$$

in which $(\hat{\mu}_t^p)^2$ are the estimated residuals of the $AR(p)$ process at a time points $t = 1, \ldots, T$, and $n$ is the number of estimated parameters. We used the *ar* function in the R package to fit our data.

## Moving Average MA(q) model

The moving average is used with time series data to smooth out short-term fluctuations from the long-term trends or cycles. The threshold between short-term and long-term depends on the application. The MA($q$) process is given by

$$X_t = \mu + \sum_{i}^{q} \theta_i \varepsilon_{t-i}, \tag{3.20}$$

where $q$ is the order of the series, the $\varepsilon_i$ satisfy (3.18), and the $\theta_i$ are parameters.

## Autoregressive Moving Average ARMA($p, q$) model

It also possible to mix both time series processes ($AR$ and $MA$) together, which means that time series has been generated by a mixed autoregressive moving average process ($ARMA$). An ARMA($p, q$) model is represented by

$$X_t = \mu + \sum_{i}^{p} \alpha_i X_{t-i} + \sum_{i}^{q} \theta_i \varepsilon_{t-i}. \tag{3.21}$$

Akaike (1981) and Schwarz (1978) in Pfaff (2008) defined information criteria for ARMA($p, q$) process as

$$AIC = \log(\hat{\sigma})^2 + 2(\frac{p+q}{T}), \tag{3.22}$$

$$BIC = \log(\hat{\sigma})^2 + \log T(\frac{p+q}{T}), \tag{3.23}$$

$\hat{\sigma}^2 = \frac{\sum residuals}{T}$ as in (3.19), so AIC in (3.23) is same as in (3.19), where $\sigma^2$ is the estimated variance of an ARMA($p, q$) process. The lag order $(p, q)$ that minimizes the information criteria is selected as the best.

## Autoregressive Integrated Moving Average ARIMA(p,d,q) model

The autoregressive integrated moving average (ARIMA) model extends an autoregressive moving average (ARMA) model by incorporating an integration term. It is usually applied on non-stationary data. The model is generally referred to as an ARIMA(p,d,q), and the first component is the autoregressive term ($AR$) part, the second is the integration ($I$) part, while the third component is the moving average ($MA$) term. An ARIMA($p, d, q$) model is such that the $d^{th}$ dif-

ference of $X_t$ is a stationary ARMA$(p, q)$ process. The Box-Jenkins procedure for time-series analysis usually includes ARIMA models (Box and Jenkins, 1976).

### 3.4.2 Time series decomposition

The most common feature about time series data is that it has values that vary with time. For instance, $SO_2$ level varies on a daily basis and it is also possible to explore variation beyond day to day changes. We may decide to examine the series over a long period of time (years) or the seasonal behaviour of the series. In either of those cases it is necessary to decompose the time series into separate components for us to examine their behaviour separately rather than mix them together (Peng et al., 2008).
The main objective in any time series analysis is to model the main features in the data either by a trend or seasonal and cyclical effects, and a model formulation is usually based on these components. A simple additive decomposition model is given by

$$X_t = T_t + S_t + \varepsilon_t, \tag{3.24}$$

where $X_t$ is the observed series (e.g $S0_2$), $T_t$ is the trend which spans across the years, $S_t$ is the seasonal effect which is the within-year variation, and $\varepsilon_t$ is the residual series for short term fluctuations. This is a classical decomposition model. Classical decomposition essentially has its main advantage as a descriptive tool to enable the main components of a time series to be viewed prior to any substantial statistical analysis. If the seasonal effect tends to increase as the trend increases, a multiplicative model may be more appropriate, i.e

$$X_t = T_t.S_t + \varepsilon_t. \tag{3.25}$$

### 3.4.3 Model residual checking

A first step in diagnostic checking of fitted models is to analyze the residuals from the fit for any signs of nonrandomness. The function $tsdiag()$ in $R$ produces diagnostic plots for a fitted time series model. The BoxPierce test also examines the null hypothesis of independently distributed residuals. It is based on the assumption that the residuals of a correctly specified model are independently distributed. The function $Box.test()$ in R computes the test statistic for a given lag. The LjungBox test can be defined as follows.
$H_0$: The data are random (uncorrelated).
$H_a$: The data are not random (correlated).

The test statistic is:

$$R = N\,(N+2)\sum_{k=1}^{n}\frac{\hat{\rho}_k^2}{N-k}, \tag{3.26}$$

in which $N$ is the sample size, $\hat{\rho}_k$ is the sample autocorrelation at lag $k$, and $n$ is the number of lags being tested. At significance level $\alpha$, the hypothesis of randomness is rejected if

$$R \geq \chi^2_{1-\alpha,n}, \tag{3.27}$$

in which $\chi^2_{1-\alpha,n}$ is the upper $\alpha$-quantile of the chi-square distribution with $n$ degrees of freedom (Ljung et al., 1978; Brockwell et al. 2002).

## 3.5 Model results for $SO_2$ data

In our analysis, the grand average daily $SO_2$ concentrations are obtained as the mean of the daily average concentrations of $SO_2$ over all the sites and for each day between 1996-2007. For the ACF, PACF, $AR$ and $ARMA$ we randomly chose some stations we consider in previous analysis (those with fewer missing observations). We only consider 1996, 2000 and 2007, but for the ARIMA model we combine the whole year range from 1996-2007 to form an aggregated data.
Firstly, we examine the ACF and PACF for possible correlation using the time series dataset generated from 3 different imputation techniques (MICE, EM and regression methods). We also compare the sensitivity of the analysis results. We used a logarithmic transformation to stabilise the variance, so the log mean daily $SO_2$ levels are used.
The autocorrelation function and 95% CI for some stations in 1996, 2000 and 2007 are plotted in Figures 3.7-3.9. In Figure 3.7 for year 1996, Glasgow 51, Glasgow 73 and Glasgow 95 are significant, thus correlated, for lags 1-10 for all three methods. The correlations are approximately between 0.2-0.6. Also, Kirkcaldy 6 is correlated for lags 1-8, and lag 13 is also significant for the EM method. Correlations range between 0.2-0.4.
Kirkintilloch 8 is significant for lags 1-6 with correlation between 0.2-0.4, and no higher lags are significant. The results are the same for all three methods. Kirkintilloch 10 is significant for all the lags with only a little evidence of periodic variation (periods 8 and 16), and correlation varies between 0.2-0.6. The results are also similar for all three methods.
In Figure 3.8 for year 2000, Glasgow 51 is significant for lags 1-3, and correlation is between 0.2-0.4. The regression method is also significant at lag 11. Glasgow 73 is significant for all the lags with little evidence of periodic variation. The

patterns are the same for the three methods. Glasgow 95 is significant at lags 1, 2, 4 and 6, and the patterns are the same in all three imputation methods. Kirkcaldy 6 is significant for lags 1-3 and in some higher lags, and the correlation is as low as 0.2.

Kirkintilloch 8 is significant for lags 1-10 with correlation between 0.2-0.6. The correlation is almost linear in trend, which is an indication of non-stationarity. We observe similar patterns for each imputation method. Kirkintilloch 10 is significant for all the lags. The correlation is as high as 0.4-0.7. Each method gives a similar pattern.

In Figure 3.9 for year 2007, there is no significant lag for the regression method for Aberdeen, EM is only significant at the first lag while MICE is significant for all the lags. Each method gives rise to a different pattern. Edinburgh St Leonards is significant for only lag 1, with the same pattern for each method. Glasgow Centre is significant for the first 2 lags in both EM and MICE, while regression is significant for the first three lags. Grangemouth has similar patterns for the three methods with lags 1-4 significant.

In summary, Kirkintilloch 10 seems to be highly correlated for all the lags irrespective of the year and method of imputation. Most of the stations are correlated in the first two lags. In general, each imputation method produces similar results, especially for 1996 and 2000. The decreasing autocorrelation pattern is almost linear in trend for all the stations, with some stations showing a little evidence of periodic fluctuations (Kirkintilloch 10). A similar procedure was applied to some of the remaining stations (the output is not shown here). The results are still consistent and similar to the one above. Generally, it can be assumed tentatively that the $SO_2$ data are temporally correlated, while carrying out further analysis. (Nunnari et al. (2004) claimed that there was no temporal correlation in the $SO_2$ data they analysed). The low variability and weather factor as well as wind speed etc could be responsible for the presence of temporal correlation for $SO_2$ data.

Figures 3.10-3.12 show the corresponding PACF plots for some stations and the 95% confidence interval for the PACF. The partial autocorrelation plots for the 1996 datasets in Figure 3.10 show clear statistical significance for lags 1 and 2, except for Kirkcaldy 6 and Kirkintilloch 10 which are also significant at higher lags 13 and 7 for both EM and regression methods respectively. The results are similar for the three methods. Since the autocorrelation plots in Figure 3.8 indicate that an $AR$ model is appropriate, we can start our modelling in 1996 with AR(2).

In Figure 3.11, Glasgow 51 and Glasgow 73 are significant at lag 3 and both still

show evidence of significance at higher lag 11, while Glasgow 95 is significant at lag 2 with evidence of significance at higher lags 6 and 17. Kirkcaldy is significant at lag 3 and some higher lags thereafter. Kirkintilloch 8 and Kirkintilloch 10 are significant for lags 1-4 with evidence of significance at higher lags. The imputation techniques also give similar results for each station. We can start our modelling here with AR(3).

In Figure 3.12, Aberdeen is significant only at lag 1 for both the EM and MICE methods, but the regression method shows evidence of significance at other lags. Also, Glasgow Centre, Edinburgh St Leonards and Grangemouth are only significant at lag 1 for all three imputation methods. In summary, the analysis gives similar results. It is appropriate to start our modelling by fitting an autoregressive model of order 1 based on the ACF and PACF results. Since the plots we have observed from 3.7-3.12 are not totally consistent we decided to consider AR(2) model for the individual stations in the next analysis. We used both maximum likelihood and least squares methods. The *ar* function in R is used to achieve the results.

Figure 3.7: Comparison of autocorrelation functions for the EM, regression and MICE imputed datasets for Glasgow 51, Glasgow 73, Glasgow 95, Kirkcaldy 6, Kirkintilloch 8 and Kirkintilloch 10 in 1996

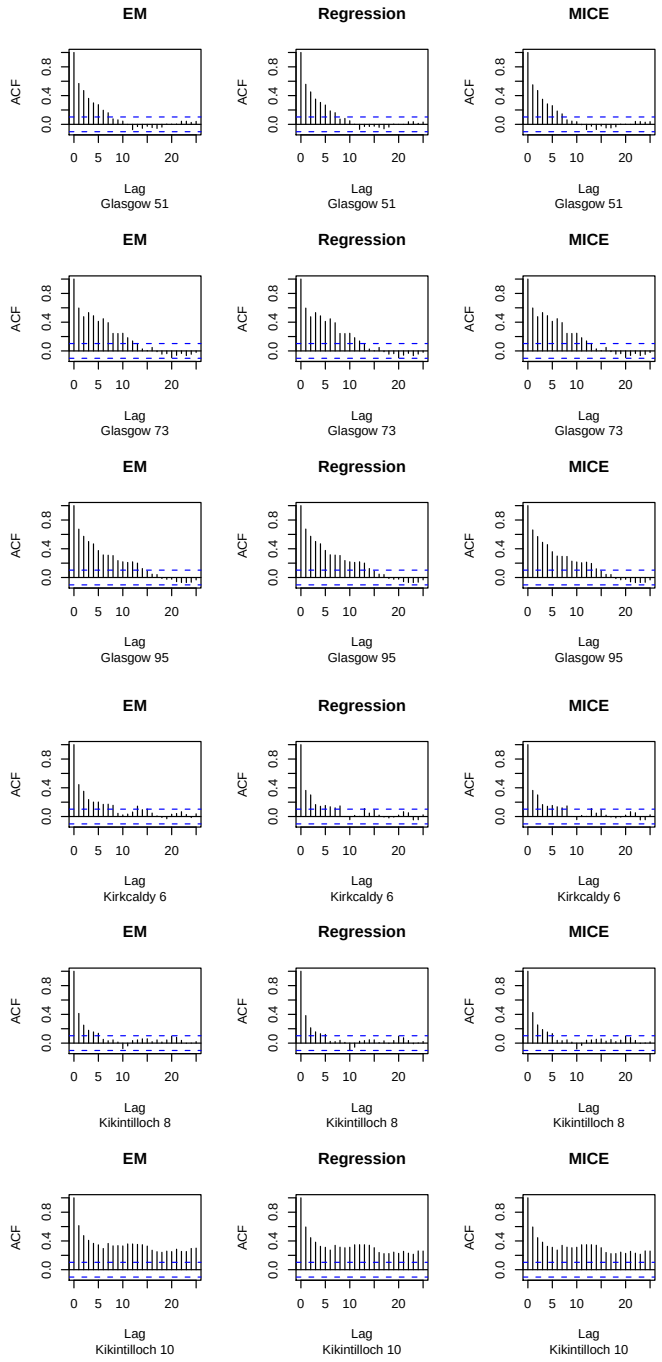


Figure 3.8: Comparison of autocorrelation functions for the EM, regression and MICE imputed datasets for Glasgow 51, Glasgow 73, Glasgow 95, Kirkcaldy 6, Kirkintilloch 8 and Kirkintilloch 10 in 2000

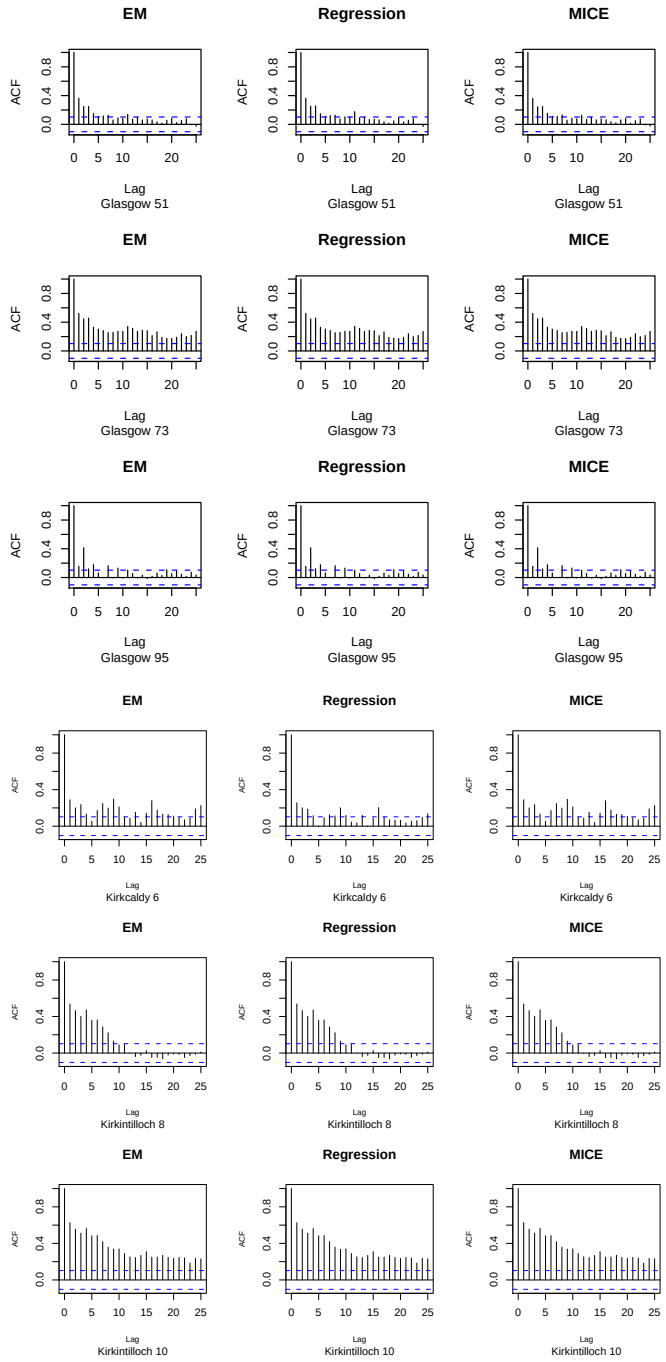


Figure 3.9: Comparison of autocorrelation functions for the EM, regression and MICE imputed datasets for Glasgow Centre, Aberdeen, Edinburgh St. Leonards and Grangemouth in 2007

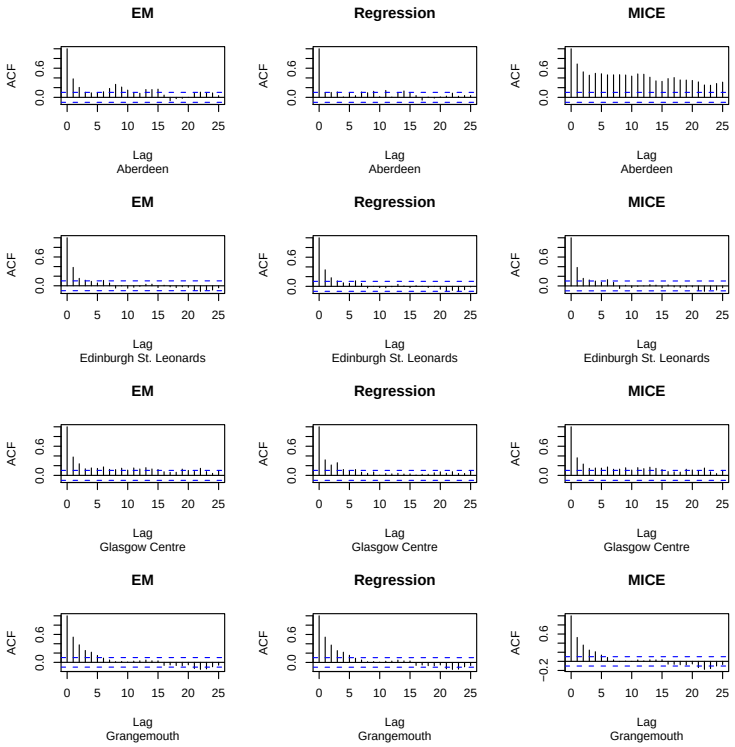


Figure 3.10: Comparison of partial autocorrelation functions for the EM, regression and MICE imputed datasets for Glasgow 51, Glasgow 73, Glasgow 95, Kirkcaldy 6, Kirkintilloch 8 and Kirkintilloch 10 in 1996

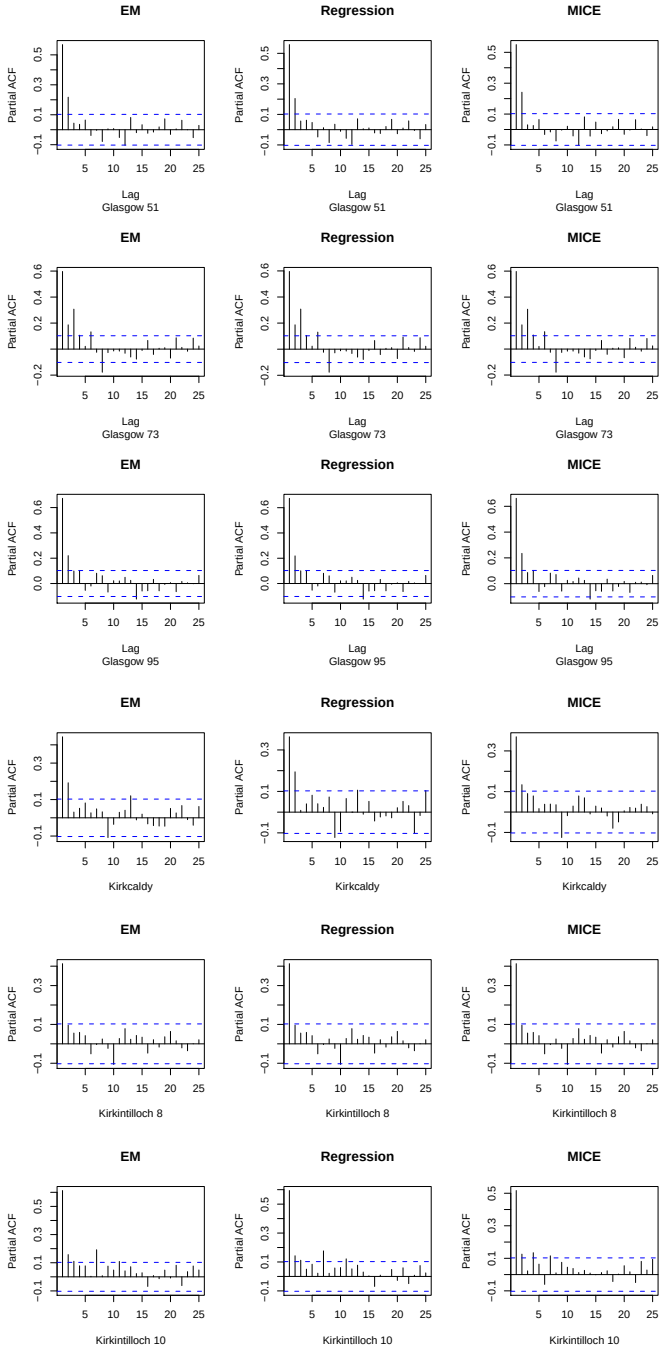


Figure 3.11: Comparison of partial autocorrelation functions for the EM, regression and MICE imputed datasets for Glasgow 51, Glasgow 73, Glasgow 95, Kirkcaldy 6, Kirkintilloch 8 and Kirkintilloch 10 in 2000

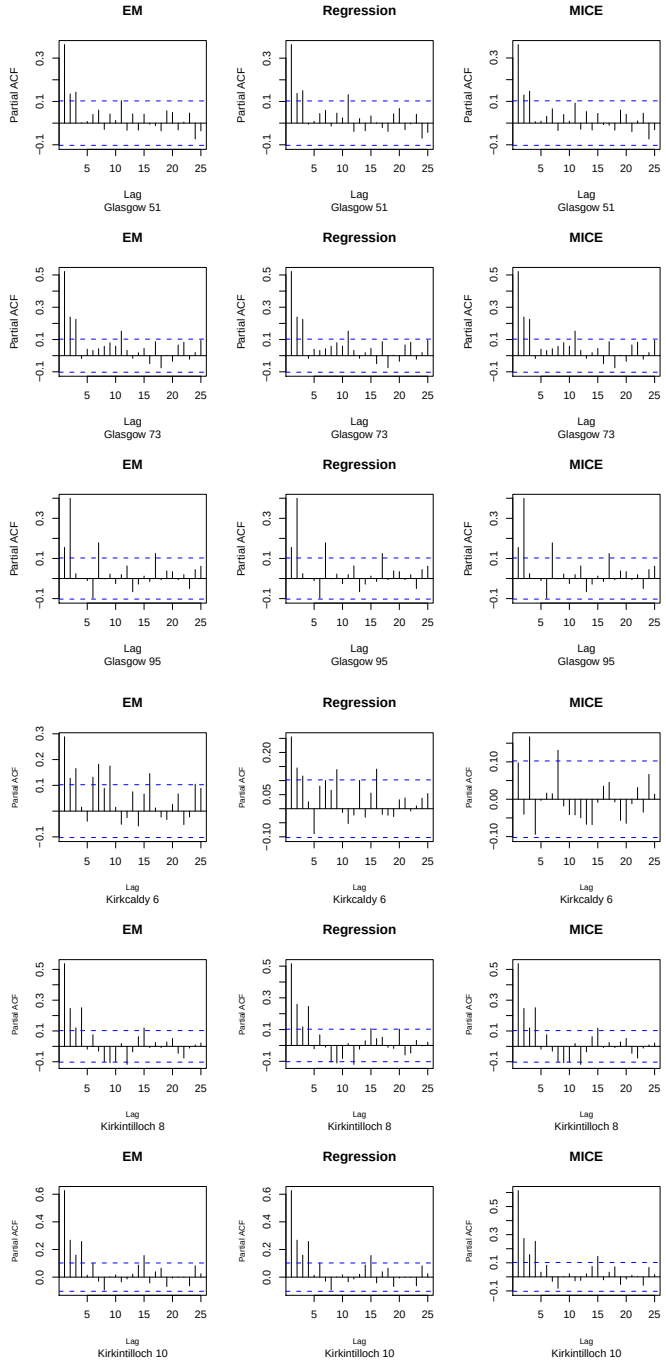
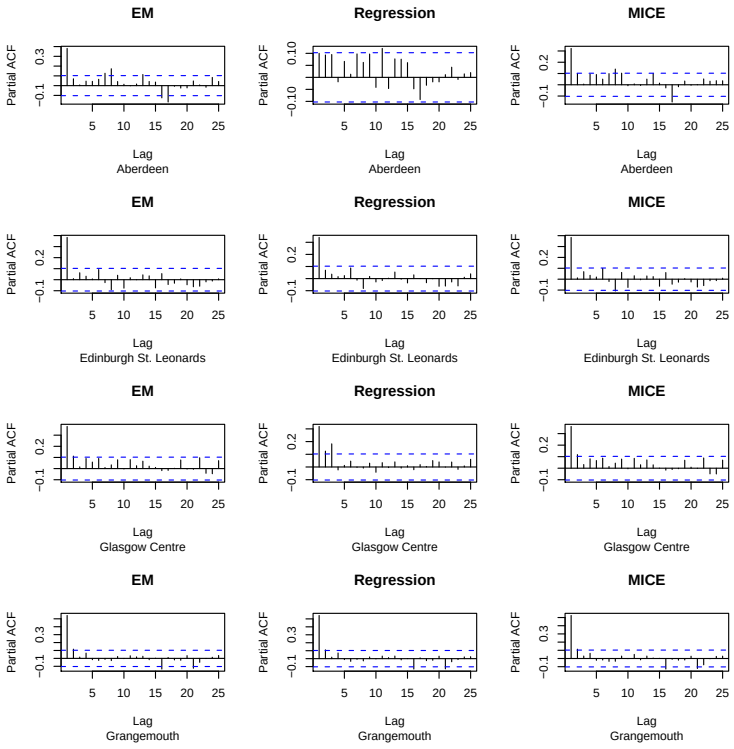


Figure 3.12: Comparison of partial autocorrelation functions for the EM, regression and MICE imputed datasets for Glasgow Centre, Aberdeen, Edinburgh St. Leonards and Grangemouth in 2007




The general time series equation of autoregressive model of order $p$ is

$$X_t = \mu + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots + \alpha_p X_{t-p} + \varepsilon. \tag{3.28}$$

If we let $log(SO_2(x,t))$ represent the log-transformed $S0_2$ level at station $x$ and time $t$, we fit autoregressive model of order 2 as

$$log(SO_2(x,t)) = \mu_t + \alpha_1 log(SO_2(x,t-1)) + \alpha_2 log(SO_2(x,t-2)) + \varepsilon(x,t). \tag{3.29}$$

We consider datasets in 1996 and 2007 to capture our data range from 1996-2007 as both year represent the early and last part of our dataset. We only consider Glasgow 51 and Glasgow 73 in 1996, and Glasgow Centre and Aberdeen in 2007. The stations are chosen randomly among stations that we have earlier considered in previous analysis to ensure consistency. The analysis is carried out on the stations using different imputation techniques, with both ordinary least squares and maximum likelihood estimation methods to fit the AR(2) model. The summary of results is shown in Table 3.5.

We observe that the values of the estimated $AR$ coefficients are significantly dif-

ferent from zero in all the chosen stations, which could imply that the assumed $2^{nd}$ order autoregressive model is sufficient for the $SO_2$ modelling.

In order to check if the coefficients are significantly different from zero we compare the estimated coefficients with their standard deviations, i.e coefficient $\pm$ twice standard error is approximately a 95% CI for the significant of parameters. In Glasgow 51 for the EM imputed data and ordinary least squares estimation method, for instance $\pm 2 * se = 2 * (0.0512) = 0.1024$ for $\alpha_1$, $(0.4421 \pm 0.1024)$, therefore $\alpha_1$ belongs to (0.3397, 0.5445) which excludes 0. The two coefficients are highly significant for this dataset (or rather since the correlogram of the true residuals which are unknown is normally distributed, $\frac{\alpha_i}{se(\alpha_i)} \approx N(0,1)$. Therefore, $\frac{0.4421}{0.0512} = 8.6347 > 1.96$ the value of test statistic falls outside the critical (rejection region), so the errors are normally distributed which implies that the coefficient is highly significant for this dataset). Also, for both the MICE and regression methods, this is significant for the two coefficients too. Also using MLE for Glasgow 51, the two coefficients are still very significant.

By continuing this same procedure for each of the stations in Table 3.5, in summary, we observe that the coefficients are significantly different from zero for all these chosen stations. It is observed that different estimation methods (ols or mle) indeed gives rise to different results. Imputation techniques (MICE, EM, regression) also produced different results.

Also, we observe that the regression method tends to produce higher variances than the MICE and EM methods. In a similar vein, the regression method also produces the least value for both AIC and the autoregressive coefficients than the other methods. The intercept estimates are similar irrespective of which imputed dataset is used. The standard errors of the coefficients are similar for both $\alpha_1$ and $\alpha_2$. The order 1 coefficient is generally higher than the order 2 coefficient. Glasgow 73 has higher variance and AIC than the other 3 stations. We observe only a fitted autoregressive model of order 1 with MICE in both Aberdeen and Glasgow Centre, which implies that AR(1) is sufficient to model these series. Maximum likelihood gives higher values of the $AR$ coefficients than least squares method in Glasgow 51, and the reverse is the case for Glasgow 73. Maximum likelihood and least squares methods produce similar $AR$ coefficients for both Aberdeen and Glasgow Centre. Least squares tends to give slightly a higher variance than the corresponding maximum likelihood method. Finally, the estimate are still very similar irrespective of the imputation methods, and both ols and mle give relatively similar results.

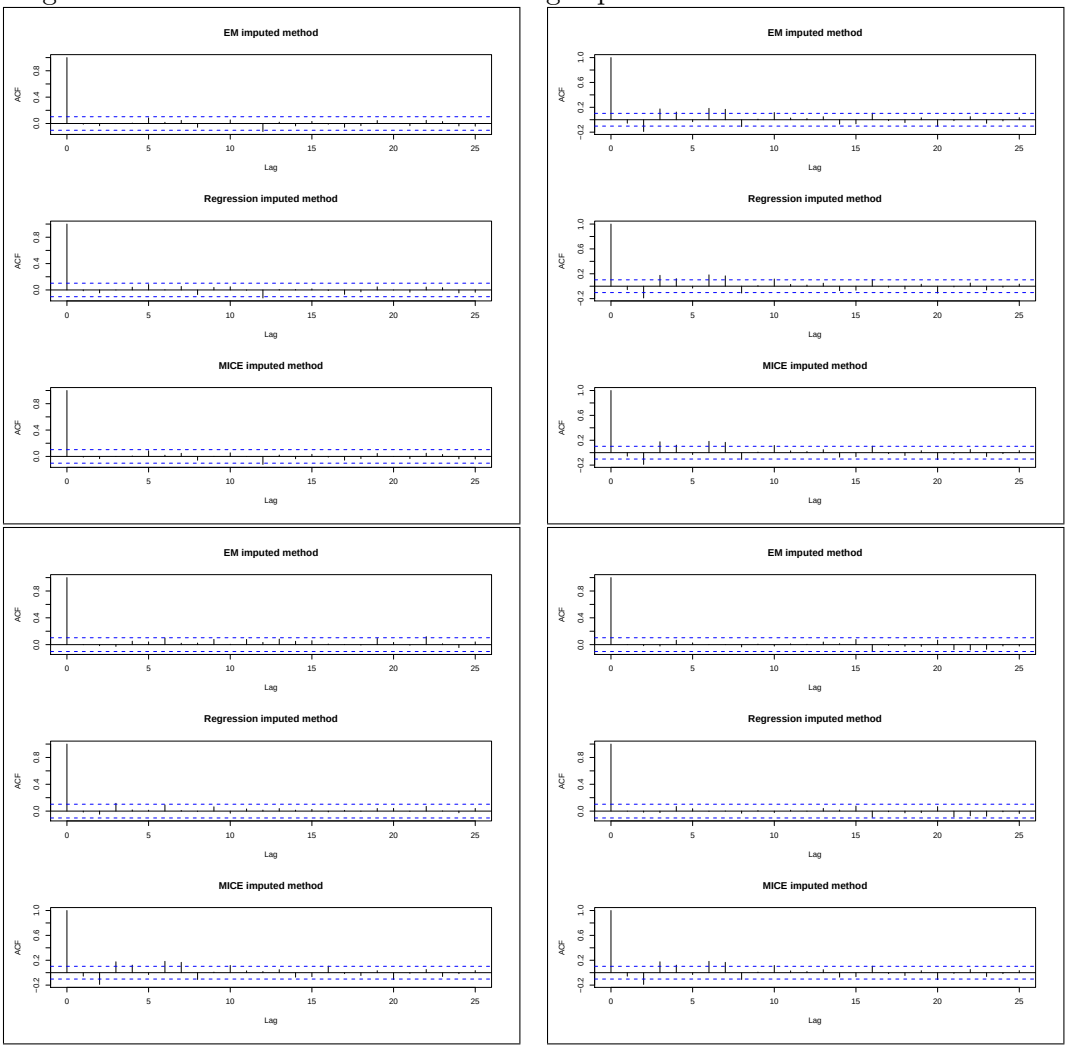
We next examine the autocorrelation function for the residuals from this fitted

model for any evidence of correlation, and the results are shown in Figure 3.13. There is no correlation in any of the plots except that Glasgow 73 in the 1996 dataset (top-right) still shows a little evidence of autocorrelation at some lags for all three imputation techniques. We have earlier seen in the PACF Figure 3.11 that AR(3) is a convenient point to start modelling data from this station. Aberdeen in 2007 (bottom-right) also indicates evidence of autocorrelation for the MICE imputed dataset only, which is also justified by the PACF in Figure 3.12. We can carry out further analysis on these stations using higher order $AR$ models or a more complex ARIMA model. Also, Glasgow Centre shows a little correlation at lags 3, 6 and 7 for MICE method. The next analysis will involve using ARIMA models for the aggregated data from 1996-2007.

Table 3.5: Comparison of autoregressive (AR) models of order 2, using both the maximum likelihood and least squares estimation methods for the three imputed dataset for Glasgow 51 and Glasgow 73 in 1996, as well as Glasgow Centre and Aberdeen in 2007. The results show the *ar* coefficients, the intercept, $\sigma^2$ and estimated AIC. G51 is Glasgow 51, G73 is Glasgow 73, Abd is Aberdeen, while Gla Cen stands for Glasgow Centre

| Site | Method | Imp | Intpt(se) | ($\alpha_1$,$\alpha_2$) | se($\alpha_1$,$\alpha_2$) | $\sigma^2$ | AIC |
|---|---|---|---|---|---|---|---|
| G 51 | ols | EM | -.0026(.0262) | (.4421,.2191) | (.0512,.0511) | .2506 | 157 |
| G 51 | ols | reg | -.0026(.0268) | (.4406, .2059) | (.0513,.0512) | .2612 | 148 |
| G 51 | ols | MICE | -.0026(.0265) | (.4194, .2410) | (.0509,.0508) | .2549 | 153 |
| G 51 | mle | EM | | (.4451,.2190) | (.0026,.0026) | .2505 | 156 |
| G 51 | mle | reg | | (.4416,.2058 ) | (.0027,.0026) | .2611 | 147 |
| G 51 | mle | MICE | | (.4225,.2410 ) | (.0025,.0024) | .2548 | 152 |
| G 73 | ols | EM | -.0004(.0311) | (.4862,.1861) | (.0516,.0516) | .3516 | 168 |
| G 73 | ols | reg | -.0004(.0311) | (.4851,.1864) | (.0516,.0516) | .3519 | 167 |
| G 73 | ols | MICE | -.0004(.0311) | (.4859,.1862) | (.0516,.0516) | .3516 | 168 |
| G 73 | mle | EM | | (.4853,.1853) | (.0026,.0026) | .3501 | 169 |
| G 73 | mle | reg | | (.4842,.1856) | (.0026,.0026) | .3503 | 168 |
| G 73 | mle | MICE | | (.4850,.1854) | (.0026,.0026) | .35 | 169 |
| Abd | ols | EM | .0008(.0139) | .3809 | .0483 | .0707 | 55 |
| Abd | ols | reg | .0006(.0162) | (.1499,.1188) | (.0521,.0520) | .09528 | 10.4 |
| Abd | ols | MICE | .0002(.0106) | .271 | .0505 | .0410 | 25 |
| Abd | mle | EM | | .3809 | .0023 | .0707 | 55 |
| Abd | mle | reg | | (.1497,.1183) | (.0027,.0027) | .0948 | 11.8 |
| Abd | mle | MICE | | (.2479,.0878) | (.0027,.0027) | .0406 | 26.6 |
| Gla Cen | ols | EM | -.0005(.0156) | (.3353,.1126 ) | (.0521,0.0521) | .0882 | 55.6 |
| Gla Cen | ols | reg | -.0005(.0149) | (.3119,.1164 ) | (.0521,.0521) | .0808 | 48.1 |
| Gla Cen | ols | MICE | -.0001(.0087) | .3802 | .0484 | .0278 | 54.3 |
| Gla Cen | mle | EM | | (.3354,.1122 ) | (.0027,.0027) | .0879 | 56.9 |
| Gla Cen | mle | reg | | (.3119,.1159 ) | (.0027,.0027) | .0805 | 49.5 |
| Gla Cen | mle | MICE | | (.3456,.0892) | (.0027,.0027) | .0347 | 55.7 |

Figure 3.13: Comparison of autocorrelation functions for the residuals using different imputation methods for the AR(2) model for Glasgow 51 and Glasgow 73 in 1996, as well as Glasgow Centre and Aberdeen in 2007; the top-left panel is Glasgow 51 in 1996, the top-right panel corresponds to Glasgow 73 in 1996, the bottom-left panel is Glasgow Centre in 2007 while the bottom right panel is Aberdeen in 2007



We then examine more complex ARIMA models here since the AR model can only be applied to a time series that is stationary (series with constant mean, variance, and autocorrelation through time) and the estimated parameters are also assumed constant throughout the series. Our datasets between 1996-2007 has relatively little variation in mean level (Figure 3.15), however the mean level drops sharply after 2005. We used the log-tranformation to stabilise the variance. The $SO_2$ data we consider in this thesis is not stationary thus justify the intoduction of ARIMA model which could be used for either stationary or non-stationary dataset. We used aggregated data in this section for us to see the general trend behaviour and pattern of $SO_2$ data from 1996-2007 and not for individual year. This will enable us to make a general comment on the over all correlation pattern and residual summary for the whole data.

We start from fitting a simple ARIMA(2,0,0) model to the chosen stations. ARIMA(2,0,0) is equivalent to AR(2) used in Table 3.5. Firstly, we computed averages of the daily mean concentration for all the stations from 1996-2007. We still consider the three imputed datasets using only the maximum likelihood method. The results are presented in Table 3.6 for each imputation method.

We observe that the order 1 coefficients are higher than the corresponding order 2 coefficients for all three methods of imputation, and both coefficients are still highly significantly different from zero for each imputation method, because 0 does not lie within the 95% CI (coefficient $\pm 2*$ *standard error*). MICE has a higher estimated value for the first order coefficient (0.6831), intercept (2.8738) and the variance (0.03007) than the other two methods. The standard errors for each coefficient are the same for each imputed dataset. The EM method has the highest log likelihood (1793.78) and the lowest AIC (-3579.55), thus looks better than the other results. In summary, each imputation method produces similar results for the coefficients. We further our analysis using the EM method only, considering ARIMA(3,0,0), ARIMA(1,0,1), ARIMA(2,1,0), ARIMA(3,1,0) and ARIMA(1,1,1) again using the joint stations dataset.

Table 3.6: ARIMA (2,0,0) model results for the MICE, EM and regression imputed datasets using maximum likelihood method, for all the stations

```
MICE imputed results
------------------------------------------------------------
Coefficients:
         ar1     ar2  intercept
      0.6831  0.2723     2.8738
s.e.  0.0145  0.0145     0.0583
sigma^2 estimated as 0.03007:
log likelihood = 1458.84, aic = -2909.68

EM imputed dataset result
------------------------------------------------------------
Coefficients:
         ar1     ar2  intercept
      0.6728  0.2849     2.8149
s.e.  0.0145  0.0145     0.0571
sigma^2 estimated as 0.0258: log likelihood = 1793.78, aic = -3579.55

Regression imputed dataset result
------------------------------------------------------------
Coefficients:
         ar1     ar2  intercept
      0.6609  0.2927     2.8085
s.e.  0.0144  0.0144     0.0546
sigma^2 estimated as 0.02849: log likelihood = 1576.83, aic = -3145.66
```

The results for the other models are shown in Table 3.7. We consider ARIMA(3,0,0) because we have earlier seen that some of the stations show evidence of autocorrelation at lag 3. For ARIMA(3,0,0), the three $AR$ coefficients are significantly different from zero. The order 1 coefficient is higher than the other two coefficients. The estimated variance is also very low (0.01054) and is lower than that of ARIMA(2,0,0). The log likelihood is 3756.47, which is higher than that of ARIMA(2,0,0), likewise the AIC (-7502.94) is lower than for the ARIMA(2,0,0) model. The standard errors of coefficients are similar and are also very low.
We decided to choose another model of lower order since both ARIMA(2,0,0) and ARIMA(3,0,0) models result indicated higher coefficient value for order 1. We incorporate an MA term of order 1 in the model to form ARIMA(1,0,1), the essence of the the MA term is to smoothen out the short fluctuations from the series. For the ARIMA(1,0,1) model, the AR(1) coefficient is 1, and the MA(1) is 0.0106, very low compared to the autoregressive coefficient, thus we can easily assume that the autoregressive parameter dominates the series. The standard

error of the $AR$ coefficient is not defined which implies that the series is not yet converged and that of the intercept is extremely very high (166.5452). The estimated variance is 0.01251, lower than that of ARIMA(3,0,0). The log likelihood and AIC are (3381.630) and (-6755.25) respectively. This model is not too reliable because of the high standard error of estimation and lack of convergence which could due to introduction of MA terms.

In order for us to overcome the little problem we observed in ARIMA(1,0,1) model, we decided to drop and replace the MA term with difference term of order 1, I(1) to form ARIMA(2,1,0) model. For ARIMA(2,1,0), the two coefficients now have negative values (-0.3917 and -0.1977) and there is no intercept parameter. This is because the series has been differenced. The standard errors of coefficients are still very low and similar for both coefficients (0.0148), and the estimated variance is now 0.01066, also similar to the model with no differencing term. The log likelihood is higher than that of ARIMA(2,0,0) and ARIMA(1,0,1) models (3731.37), but lower than for ARIMA(3,0,0).

Because problem of negative coefficients observed in ARIMA(2,1,0) model, we consider higher AR term model but still maintaining the differencomng term to form ARIMA(3,1,0) model. For ARIMA(3,1,0), the three coefficients here also have negative values (-0.4269, -0.2673, and -0.1776) and there is no intercept parameter, which is due to the introduction of a differencing term in the model. The standard errors of coefficients are still very low, and are the same for both order 1 and 3 coefficients (0.0149, 0.0157, 0.0149), and the estimated variance is now 0.01032, also similar to the model with no differencing term. The log likelihood is higher than for the four previous models (3801.64).

Lastly, we consider a model that gives a uniform order the 3 terms to form ARIMA(1,1,1). For ARIMA(1,1,1), the AR and MA coefficients are 0.3648 and -0.8360 respectively. The standard errors are different for each of them, and the estimated variance is lowest (0.009888) and log likelihood is highest (3895.6) for all the models we consider.

If we compare the 6 models in Table 3.8, we observe that based on both the minimum AIC (-7785.2), and maximum loglik (3895.6) and least estimated variance (0.00988) criteria, ARIMA(1,1,1) seems better than the other models though it uses 3 degrees of freedom. Generally, the six ARIMA models are still consistent with the autoregressive results we got earlier. The models with differencing terms ARIMA(2,1,0), ARIMA(3,1,0) and ARIMA(1,1,1) seem better than those with no differencing term, which could easily be attributed to the fact that the series now has more stable parameters after differencing of the series to achieve

stationarity.

Table 3.7: ARIMA(3,0,0), ARIMA(1,0,1), ARIMA(2,1,0), ARIMA(3,1,0) and ARIMA(1,1,1) model results using the EM imputed datasets with maximum likelihood method, for the combined stations

```
ARIMA (3,0,0) model result
-------------------------------------------------------------------
Coefficients:
         ar1     ar2     ar3  intercept
      0.5944  0.1850  0.1836     3.1917
s.e.  0.0148  0.0171  0.0149     0.0415
sigma^2 estimated as 0.01054: log likelihood = 3756.47,aic = -7502.94

ARIMA (1,0,1) model result
-------------------------------------------------------------------
Coefficients:
      ar1     ma1  intercept
        1  0.0106     3.1917
s.e.  NaN  0.0186   166.5452
sigma^2 estimated as 0.01251: log likelihood = 3381.63,aic = -6755.25

ARIMA (2,1,0) model result
-------------------------------------------------------------------
Coefficients:
          ar1      ar2
      -0.3917  -0.1977
s.e.   0.0148   0.0148
sigma^2 estimated as 0.01066: log likelihood = 3731.37,aic = -7456.74

ARIMA (3,1,0) model result
-------------------------------------------------------------------
Coefficients:
          ar1      ar2      ar3
      -0.4269  -0.2673  -0.1776
s.e.   0.0149   0.0157   0.0149
sigma^2 estimated as 0.01032: log likelihood = 3801.64,aic = -7595.27

ARIMA (1,1,1) model result
-------------------------------------------------------------------
Coefficients:
         ar1      ma1
      0.3648  -0.8360
s.e.  0.0247   0.0163
sigma^2 estimated as 0.009888: log likelihood = 3895.6,aic = -7785.2
```

Table 3.8: Comparison of ARIMA model result for the EM imputed dataset, showing the degrees of freedom, AIC, logLik and variance, using maximum likelihood method for all the stations

| | df | AIC | loglik | Variance |
|---|---|---|---|---|
| ARIMA(2,0,0) | 4 | -3579.555 | 1793.7 | 0.0258 |
| ARIMA(3,0,0) | 5 | -7502.935 | 3756.47 | 0.01054 |
| ARIMA(1,0,1) | 4 | -6755.253 | 3381.63 | 0.01251 |
| ARIMA(2,1,0) | 3 | -7456.745 | 3731.37 | 0.01066 |
| ARIMA(3,1,0) | 4 | -7595.275 | 3801.64 | 0.01032 |
| ARIMA(1,1,1) | 3 | -7785.202 | 3895.6 | 0.00988 |

Table 3.9 shows the result of the Ljung-Box test (Box.test) in R for the six ARIMA models we consider, and Figure 3.14 gives equivalent diagnostic plots for the standardized residuals, the ACF of the residuals and $p$ values for the Ljung-Box statistic for the ARIMA models. The p-values (Table 3.9) are very low compared to 5% significance, which implies that the tests are highly significant for each model. The data is assumed to be non-random (correlated) for each model.

In Figure 3.14, the standardized residual plots indicate that the residuals are small for days 1-3653, which is equivalent to years 1996-2005, and larger between days 3654-4382, which correspond to years 2006-2007 data. Within this later interval where the data are more sparse thus suggest that the model fits less well. Autocorrelation plots (middle panel in each box) also indicate that for the first 35 lags all sample autocorrelations are within the 95% confidence interval for ARIMA(2,0,0), except at lag 1 only. Also for ARIMA(3,0,0), all sample autocorrelations are within the 95% confidence interval except at lag 3. For ARIMA(1,0,1), it shows evidence of correlation at lag 1, for ARIMA(2,1,0) it shows evidence of correlation at lags 2 and 3. ARIMA(3,1,0) show evidence of correlation at lag 4 only. Lastly, for ARIMA(1,1,1) there is no evidence of correlation.

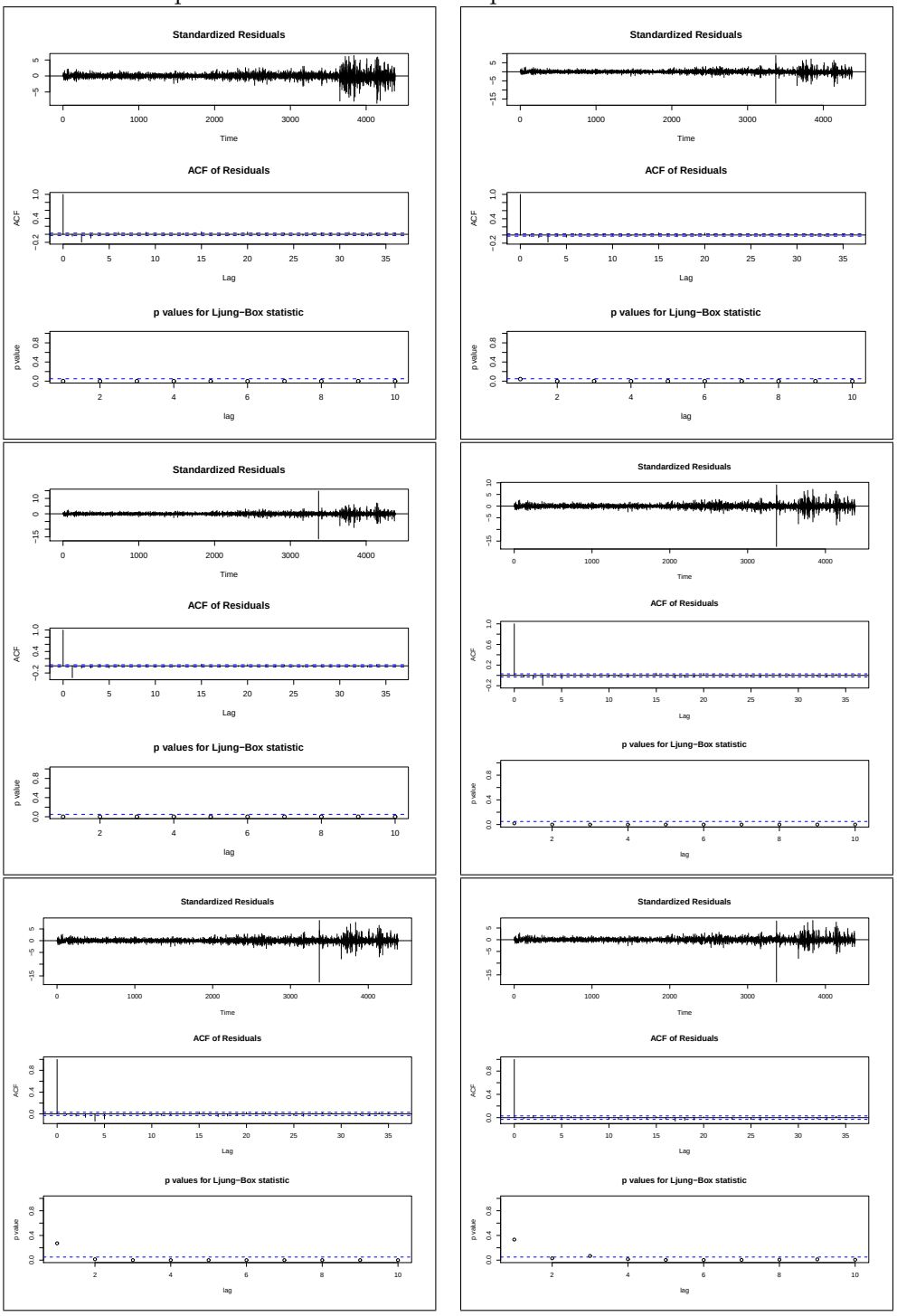
Also, the p-values for Ljung-Box statistic are all very small, except for ARIMA(3,1,0) and ARIMA(1,1,1) that have a value of about 0.36 at lag 1. In summary the residuals appear to be non-random in nature and there is a lot of residual variation between 2006 and 2007 which could be as a result of fewer obervations in these 2

years.

Table 3.9: Comparison of Ljung-Box test for the residuals of the six ARIMA models, type="Ljung-Box". The results give the chi-squared, degrees of freedom and p-values

```
 ARIMA(2,0,0)$residuals
X-squared = 310.6135, df = 25, p-value < 2.2e-16
------------------------------------------------
ARIMA(3,0,0)$residuals
X-squared = 237.0911, df = 25, p-value < 2.2e-16
------------------------------------------------
ARIMA(1,0,1)$residuals
X-squared = 585.7212, df = 25, p-value < 2.2e-16
------------------------------------------------
ARIMA(2,1,0)$residuals
X-squared = 284.5121, df = 25, p-value < 2.2e-16
------------------------------------------------
ARIMA(3,1,0)$residuals
X-squared = 194.9454, df = 25, p-value < 2.2e-16
------------------------------------------------
ARIMA(1,1,1)$residuals
X-squared =80.4731, df = 25, p-value = 9.618e-08
```

Figure 3.14: Comparison of diagnostic plots for different ARIMA models for daily $SO_2$ concentrations. The top-left box represents ARIMA(2,0,0), the top-right box is ARIMA(3,0,0), the middle-left box is ARIMA(1,0,1), the middle-right box is ARIMA(2,1,0), while the bottom-left and bottom-right boxes are ARIMA(3,1,0) and ARIMA(1,1,1) respectively. In each box, the top panel is standardized residuals, the middle is the ACF plot and the bottom is the p-value



We next decompose the time series data. The results are presented in Figures 3.15 and 3.16. These consists of the non-stationary time series of the logarithm of the mean daily $SO_2$ concentration and their three time-scale decompositions, after imputting the missing observations.

Having decomposed the data, we observe that both the long-term trend and seasonal/cyclical effects contribute significantly to variation in $SO_2$ levels across the years. The long-term component in Figure 3.15 shows variation in $SO_2$ levels over the years. The long-term series decreases between 1996 and 1999, before a gentle rise in 2000. The mean level increases gradually again to 2005 before finally falling to a very low-level in 2007. We have two prominent spikes between 2003 and 2005, and the approximate mean level here is about 3.7 $\mu g/m^3$.

For the main trend in Figure 3.16, the actual trend becomes more apparent, but shows a relatively constant pattern with little fluctuation between 1997-2005. It shows a declining pattern between 2006-2007 and a little rise thereafter. There are three prominent peaks in the trend between 2003-2006 which as high as 1.4 $\mu g/m^3$. The seasonal pattern also shows a high regular periodic fluctuation in levels between 1996 and 2001 and a low regular periodic variation in levels between 2002 and 2005. There is a higher irregular periodic fluctuations in mean level between 2005-2007. These are the usual winter and lower summer peak variation in levels.

The short term residual component is left after the removal of the trend and seasonal components, and this also shows moderate fluctuation in levels and is not always random in nature. The residuals are small between 1996-2005, and larger between 2006-2007 data. There is an high residual variation in levels within this later interval where the data are more sparse which is similar to what we earlier observed in ACF plots in Figure 3.14.

Figure 3.15: Long-term trend of EM imputed $log(S0_2)$ concentration

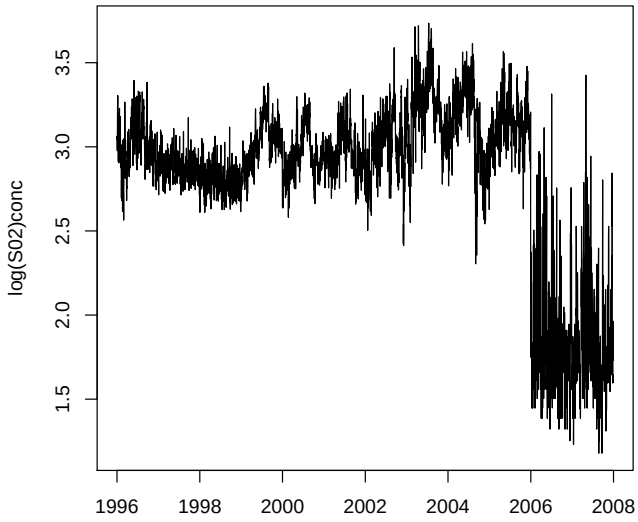
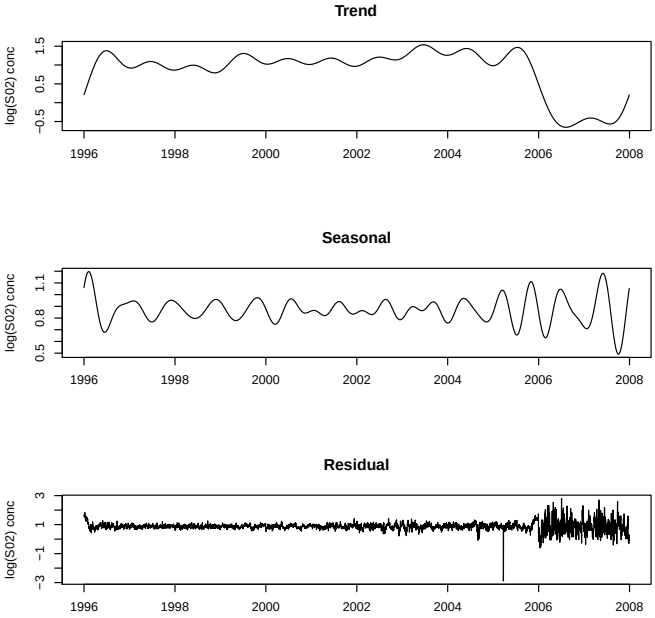



Figure 3.16: Timescale decomposition of $log(S0_2)$ concentration




## 3.6 Conclusion

In this chapter we explore various imputation techniques and apply them on several time series models. We first explore the ACF to check for any evidence of

serial correlation in our dataset, and we are able to see that some of the recording stations in 1996, 2000 and 2007 have correlated data. We later proceeded in the analysis by examining the PACF plots, to choose a suitable order for the autoregressive model. Three imputation techniques are also considered, but all have similar effects. We used the EM imputed dataset for most of the analysis in this chapter.

In the last part of the chapter, ARIMA models and time series decomposition are also considered. The ARIMA (1,1,1) model is able to model the $SO_2$ data very well. We observe that both the long-term trend and seasonal/cyclical effects contribute significantly to variation in $SO_2$ levels across the years for most of the stations we considered. We continue our further analysis based on the EM method, as this is suitable for data that are not MCAR, and results were similar for all three imputation methods. We next continue the modelling by considering spatial analysis in Chapter 4, as this will enable us to account for the spatial variation and autocorrelation in the $SO_2$ data.

# Chapter 4

# Spatial analysis of $SO_2$ levels in Scotland

## 4.1 Objectives of spatial analysis

Chapter 3 considered the time series modelling of the data. We now turn to spatial analysis in this chapter, before using aspects of both space and time in Chapter 5.

Section 4.1 gives the main objectives of this chapter and a general introduction to spatial analysis as well as outlining a number of topics from the theory of spatial stochastic analysis, focussing on concepts of autocorrelation.

Section 4.2 deals with the theory of spatial autocorrelation and highlights the methods involved in spatial analysis. Section 4.3 deals with variograms and kriging theory, the statistical theory surrounding estimation of the variogram, and fitting of parametric covariance models, as well as detailed description of the Bayesian approach.

Section 4.4 discusses the methods adopted and presents results of our analysis of the $SO_2$ data as well as comparing the impact of different parameter estimation techniques and the use of prior knowledge. Section 4.5 presents conclusions and suggestions for further work.

### 4.1.1 Aims of work

The main objective of this analysis is to identify the sources of spatial variation in $SO_2$ concentration levels and to estimate the pollution level at unmonitored spatial locations as well as to predict outside the range of the present data. In our analysis, we assume a separable model for spatial correlation (Hobert et al.,

1997; Bowman et al., 2009).
The dataset contains log (mean $SO_2$) for forty-one monitoring stations in Scotland. We will consider only years 1996, 2000 and 2005 but the kriging model will focus only on 1996 dataset. The data include the Northing and Easting of each station. We calculated average annual concentration for each station per year after imputation. Most of the stations are concentrated in Central Scotland and the North-Eastern part of Scotland. This limits the performance of our prediction results because the stations are not randomly distributed widely across Scotland. Having computed the MCAR test on our dataset in Chapter 3, for better imputation of missing observations we adopted the SPSS EM method to impute the missing observations before log-transforming the completed data, in line with Smith & Kolenikov (2003) who also chose to use the EM algorithm (in Fortran and Stata software) to deal with missing $PM_{2.5}$ observations and direct use of log-transformation in line with Bowman et al. (2009).

## 4.2 Spatial analysis and spatial autocorrelation

### 4.2.1 Spatial autocorrelation

Spatial analysis consists of techniques and models that use spatial referencing related with each dataset, including observations using topological, geometric and geographic properties. Spatial analysis techniques have been developed in various areas. Spatial analysis starts with mapping, surveying and geography, but modern spatial analysis focuses on computer based techniques because of the large amounts of data which may be involved, and the use of modern statistical and geographic information systems (GIS) software in the computational modelling.
Spatial dependence can be described as the co-variation of properties at proximal locations within geographic space, either positive or negative co-variation. It is possible that there is a simple spatial correlation relationship involving an observation at any particular location that also causes similar (or different) observations in nearby locations. Also, like temporal autocorrelation, spatial autocorrelation estimates the degree of dependence among observations but in a particular geographical area.
Spatial autocorrelation means that spatially related observations of the same phenomenon are associated, which violates statistical assumption of independence among observations. It also complicates statistical analysis by altering the variance of variables. Spatial dependence could be seen as additional information on a phenomenon rather than being a source of problems in spatial analysis. Het-

erogeneity of a spatial process means that overall parameters estimated for the entire system are different at different locations (Cressie, 1993; Chiles & Delfiner, 1999; Wackernagel, 1995; Griffith, 1999).

Positive spatial autocorrelation is an indication that nearby geographical locations shows clusters of similar observations, while significant negative spatial autocorrelation indicates that neighbouring data are more widely dispersed than expected by ordinary variation in levels.

The measures of spatial autocorrelation include Moran's I and Geary's C (Griffith, 2006). These statistics require constructing a spatial weights matrix that reflects the strength of the geographic relationship between locations, for instance the distances between two locations, the length of a shared border, or whether they fall into a specified directional class. Geary's C is defined by

$$C = \frac{(N-1)\sum_i \sum_j w_{ij}(X_i - X_j)^2}{2W\sum_i (X_i - \bar{X})^2}, \tag{4.1}$$

where $N$ is the number of spatial locations, X is the random variable, $\bar{X}$ is the mean of $X$, and $w_{ij}$ is a spatial weight relating locations $i$ and $j$, and $W$ is the sum of all $w_{ij}$. Geary's C is inversely related to Moran's I, but is more sensitive to local spatial autocorrelation, while Moran's I is a measure of general spatial autocorrelation.

Moran's I is defined as

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2}, \tag{4.2}$$

where $N$ is the number of spatial locations, $X$ is the random variable, $\bar{X}$ is the mean of $X$, and $w_{ij}$ is a spatial weight as above. The two statistics can be computed using function $Moran.I$ and *geary.test* in the *ape* and *spdep* package of R.

We calculate Moran's I by first generating a matrix of inverse distance weights which is supplied to the $Moran.I$ function. The null hypothesis of the test is that there is no correlation, based on the assumption of normality. If the observed value of I is significantly greater than the expected value, this indicates positive autocorrelation in the data, otherwise if $I_{observed} < I_{expected}$ this indicates negative autocorrelation (Gittleman & Kot et al., 1990).

### 4.2.2 Methods for spatial analysis

Spatial interpolation estimates or predicts the variables at unobserved locations in a geographic area based on the values at observed locations. Basic methods include inverse distance weighting, which reduces the value of the variable with decreasing proximity from the observed location.

*Kriging* is a modern method that interpolates across space according to a spatial lag relationship that has both systematic and random components. This can accommodate a wide range of spatial relationships to describe the hidden values between observed locations. Kriging provides optimal estimates of the relationship between spatially related variables, and error estimates can also be examined to determine whether spatial patterns exist. Details are given below and kriging is applied to the $SO_2$ data.

Spatial regression methods capture spatial association in regression analysis, avoiding statistical problems such as unstable parameters and unreliable significance tests, as well as providing information on spatial relationships among the variables involved. However, weighted regression is like a spatial regression that generates parameters separated by spatial units of analysis. This assesses the heterogeneity of spatial patterns in the estimated relationships between the independent and dependent variables (Longley & Batty, 1997). Spatial models are used as part of kriging and are also fitted in Chapter 5.

## 4.3 Geostatistics, variograms and kriging

### 4.3.1 Geostatistics

Geostatistics is the study of phenomena that vary spatially or temporally. It is a collection of numerical methods that deal with the characterization of spatial attributes, using stochastic models in a manner similar to the way in which time series analysis characterizes temporal data (Olea, 1999).

The correlogram, the covariance and the semivariogram or variogram are the three main functions usually used in geostatistics for spatial analysis description (Cressie, 1993; Ripley, 1991). The variogram is the main function among these three, as it is used for the correlation model of observed data in spatial analysis. Variogram covariance models are usually estimated by the linear, spherical, Matern, Gaussian or Exponential functions.

Kriging is an optimal interpolation technique which generates a best linear unbiased estimate at each location and employs a semivariogram model (Chils &

Delfiner, 1999; Isaaks & Srivastava, 1989), while co-kriging is an interpolation technique that gives better estimation of kriging if the distribution of a secondary variate sampled more intensely than the primary variate is known, that is it involves multiple variables. This is not available here but potentially we could use distance to main roads or population density as covariates.

### 4.3.2 Variogram

A variogram is a function used for describing the degree of spatial dependence of a spatial random field or stochastic process $Z(x)$. Empirical variograms are used to explore the spatial structure of the observed spatial data.
The main function for empirical variogram computation is *variog* in the *geoR* package of R. The classical moment estimator is a commonly adopted method in the computation of the semi-variance. It is computed as (Wackernagel, 2003; Ribeiro & Diggle, 2000),

$$2\gamma(x_1, x_2) = E\left(|Z(x_1) - Z(x_2)|^2\right), \tag{4.3}$$

for spatial positions $x_1$ and $x_2$ in which $\gamma(x_1, x_2)$ represents the *semivariogram*, and $\gamma_s(h) = \gamma(0, 0 + h)$, is a function of the distance $h = |x_2 - x_1|$ between locations only. In general

$$\gamma(x_1, x_2) = \gamma_s(|x_2 - x_1|). \tag{4.4}$$

As $\gamma(x_1, x_2) = E(|Z(x_1) - Z(x_2)|^2) = \gamma(x_2, x_1)$, the semivariogram is a symmetric function. Also $\gamma_s(h) = \gamma_s(-h)$, is an even function. The semivariogram is non-negative, $\gamma(x_1, x_2) \geq 0$. At distance 0, $\gamma(x, x) = \gamma_s(0) = E\left((Z(x) - Z(x))^2\right) = 0$. Also, the variogram satisfies

$$\sum_{i=1}^{N}\sum_{j=1}^{N} w_i \gamma(x_i, x_j) w_j \leq 0, \tag{4.5}$$

for all weights $w_1, \ldots, w_N$, such that $\sum_{i=1}^{N} w_i = 0$, and for locations $x_1, \ldots, x_N$. The variogram and covariance functions are related by

$$2\gamma(x_1, x_2) = C(x_1, x_1) + C(x_2, x_2) - 2C(x_1, x_2), \tag{4.6}$$

where $C(x_1, x_2) = Cov(Z(x_1), Z(x_2))$ (Cressie, 1993; Chiles & Delfiner 1999; Wackernagel, 2003), so

$$\gamma(x_1, x_2) = \frac{1}{2}[V(Z(x_1)) + V(Z(x_2)) - 2Cov(Z(x_1), Z(x_2))], \tag{4.7}$$

and for a stationary process we have that $\gamma(h) = C(0) - C(h)$, writing $C(h) = C(x_1, x_2)$ for locations $x_1$, $x_2$ such that $|x_1 - x_2| = h$ (Schabenberger & Gotway, 2005).

For any observations $z_i,\ i = 1, \ldots, k$, at locations $x_1, \ldots, x_k$ the empirical semivariogram $\hat{\gamma}(h)$ is defined as

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{(i,j) \in N(h)} |z_i - z_j|^2, \tag{4.8}$$

in which observations $i, j$ are such that distance $|z_i - z_j| = h$, and $|N(h)|$ is the number of pairs in the set. The natural estimator of $C(h)$ is

$$\frac{1}{|N(h)|} \sum_{(i,j) \in N(h)} (Z(x_1) - \bar{Z})(Z(x_2) - \bar{Z}). \tag{4.9}$$

The empirical variogram is used in spatial statistics as an initial estimate of the variogram needed for spatial interpolation by kriging. Kriging is usually computed for shorter distances, thus in fitting a model to the sample variogram, the values of the sample variogram for longer lag distances are not so important.

### 4.3.3 Parametric covariance function

Several models are used to fit an empirical variogram or corresponding covariance function. The Matern model for the covariance function $C(h)$ between two points at distance $h$ apart is defined as

$$C(h) = \frac{\sigma^2}{2^{v-1}\Gamma(v)} (\varphi h)^v \beta_v(\varphi h), \tag{4.10}$$

for $v, \varphi, h > 0$, $\Gamma$ is the gamma function, and $\beta_v$ is the modified Bessel function of the second kind of order $v$. The parameter $v > 0$ represents a smoothness parameter, and the Matern model reduces to the Exponential model for $v = \frac{1}{2}$ and to the Gaussian model as $v \to \infty$.

The Gaussian covariance model is given by

$$C(h) = \sigma^2 exp(-\varphi h^2) = \sigma^2 exp(-3\frac{h^2}{\alpha^2}), \tag{4.11}$$

the Exponential covariance model is given by

$$C(h) = \sigma^2 exp(-\varphi h) = \sigma^2 exp(-3\frac{h}{\alpha}), \tag{4.12}$$

and the spherical covariance model is

$$C(h) = \sigma^2(1 - \frac{3h}{2\alpha} + \frac{1}{2}(\frac{h}{\alpha})^3), h \leq \alpha, \text{and 0 otherwise} \tag{4.13}$$

(Schabenberger & Gotway, 2005; Chiles & Delfiner, 1999; Cressie, 1993; Handcock & Stein, 1993).

Distance computation is very important in spatial analysis. Inter-site distance computations are used in variogram analysis to compute the degree of spatial relationship, which helps in specifying priors for the range parameter in Bayesian modelling (Ecker & Gelfand, 1997), and for initial values of the non-linear least squares algorithms in classical analysis (Cressie, 1993).

In the spherical coordinate system, let $P_1 = (\lambda_1, \theta_1)$ and $P_2 = (\lambda_2, \theta_2)$ be two locations on the Earth's surface given by longitudes $\lambda_1$ and $\lambda_2$ and latitudes $\theta_1$ and $\theta_2$. The *geodesic* distance is the arc length of a circle joining any two points and is obtained as $R\varphi$, where $R$ (6371 km) is the radius of the earth and $\varphi$ (in radians) is the angle between the two points, and is given by

$$d = R\varphi = R\ arccos(sin\theta_1 sin\theta_2 + cos\theta_1 cos\theta_2 cos(\lambda_2 - \lambda_1)). \tag{4.14}$$

If the number of data locations is $n$ then there will be a total of $n(n-1)/2$ possible pairs of locations to consider. The field package in R computes geodesic distance using the function $rdist.earth()$. The distance summary will enable us to create reasonable lag intervals for the range parameter in the variogram computation.

### 4.3.4 Kriging

Kriging is a geostatistical technique for interpolating the value of a unknown random observation from data $Z(x)$ observed at known locations. Kriging is usually applied for interpolating environmental measurements (Switzer, 1977 & 1989). For kriging, the data $Z(\mathbf{x}) = (Z(x_1), \ldots, Z(x_N))$ are assumed to arise

from a random field defined on the spatial area of interest, such that the mean

$$E(Z(\mathbf{x})) = \mu(\mathbf{x}) \tag{4.15}$$

and $cov(Z(\mathbf{x})) = \Sigma$. Kriging uses a linear model for interpolation of $Z(x_0)$ at a location $x_0$, taking

$$\hat{Z}(x_0) = \sum_{i=1}^{N} w_i(x_0) Z(x_i), \tag{4.16}$$

for observed locations $x_1, \ldots, x_N$, and where the coefficients or weights $w_i$ are estimated to minimize the variance of prediction error,

$$V(\hat{Z}(x_0) - Z(x_0)) = \sum_{i=1}^{N}\sum_{j=1}^{N} w_i(x_0) w_j(x_0) C(x_i, x_j) + V(Z(x_0)) - 2\sum_{i=1}^{N} w_i(x_0) C(x_i, x_0), \tag{4.17}$$

subject to

$$E(\hat{Z}(x_0) - Z(x_0)) = \sum_{i=1}^{N} w_i(x_0)\mu(x_i) - \mu(x_0) = 0, \tag{4.18}$$

where $\mu(x)$ is a trend, $\mu(x) = E(Z(x))$ and $C(x_1, x_2) = Cov(Z(x_1), Z(x_2))$ is the covariance function of the random field $Z(x)$ relating observations at locations $x_1$ and $x_2$. "$\hat{Z}(x)$ is an unbiased estimator". There are various forms of kriging, namely *simple, universal, ordinary* and *Bayesian* kriging, using different forms for $\mu(\mathbf{x})$, the mean or trend. Simple kriging is based on the assumption that the trend function is a known constant that can take different values for different locations, and ordinary kriging uses a trend that is unknown but constant across locations, while the more general form for universal kriging takes

$$\mu(x) = \sum_{i=1}^{p} \beta_j f_j(x) = \beta^T f(\mathbf{x}), \tag{4.19}$$

where $\beta^T = (\beta_1, \ldots, \beta_p)$ are unknown regression parameters, and $f(\mathbf{x})^T = (f_1(x), \ldots, f_p(x))$ are known covariates depending on spatial location (e.g. Easting and Northing). For any form of kriging the nature of $\Sigma$ is specified, through a covariance function relating $Z(x)$ and $Z(y)$, so that

$$cov(Z(x), Z(y)) = \alpha K_\theta(\mid x - y \mid), \tag{4.20}$$

where $\alpha > 0$ is a scale parameter, and $\theta^T = (\theta_1, \ldots, \theta_q)$ is a vector of real-valued structural parameters for the covariance function (Handcock and Stein, 1993; De

Oliveira et al., 1997). The kriging estimate of $Z(x_0)$ is then given by

$$\hat{Z}_\theta(x_0) = k_\theta^T K_\theta^{-1} Z + b_\theta^T \hat{\beta}(\theta), \tag{4.21}$$

in which $k_\theta^T = (K_\theta(x_0, x_1), \ldots, K_\theta(x_0, x_N))$, $K_\theta$ is the $N \times N$ matrix with $(i, j)$th element $K_\theta(x_i, x_j)$ which is a function relating the covariance function with the distance between pair of locations $i$ and $j$, $b_\theta = f(x_0) - F^T K_\theta^{-1} k_\theta$, $F$ is the $N \times p$ matrix $\{f_j(x_i)\}$ and

$$\hat{\beta}_\theta = (F^T K_\theta^{-1} F)^{-1} F^T K_\theta^{-1} Z \tag{4.22}$$

(Handcock and Stein, 1993). This assumes the covariance function and its parameters to be known. Commonly a form of variogram model is fitted to the empirical variogram, or a covariance model is fitted to the empirical covariance function, and parameter estimates for $\alpha$ and $\theta$ are found by maximum likelihood estimation, weighted least squares estimation or an ad hoc method (e.g. to obtain a best visual match) (Handcock and Stein, 1993). The parameter values are then treated as if they were known for the purposes of kriging, and the estimate $\hat{Z}(x_0)$ then uses the estimated parameters. This ignores the uncertainty in the estimates.

Bayesian kriging takes account of uncertainty about the nature of the covariance function. Inference is based on the Bayesian predictive distribution $p(Z(x_0)) \mid Z(x))$, using the mean (or mode or median) of this as an estimate $\hat{Z}(x_0)$. Handcock and Stein (1993) outline a Bayesian framework for prediction of random fields which are assumed to be Gaussian. The predictive distribution is derived from $p(Z(x_0) \mid \theta, Z)$ and $p(\theta \mid Z)$ by integrating out their product over the unknown parameters $\theta$, after specifying suitable probability models. Numerical integration may be necessary.

It is not always appropriate to assume a Gaussian model for the random field data. Trans-Gaussian kriging involves finding a transformation of the spatial data for which the transformed data are approximately normally distributed, using kriging to find the best linear unbiased predictor for the transformed data and then transforming back, with a correction for bias in the original scale of measurement (De Oliveira et al., 1997; Cressie, 1993).

De Oliveira et al. (1997) extend Handcock and Stein (1993) to random fields which can be transformed to follow a Gaussian distribution, using a particular class of transformations defined by a single parameter, but avoiding the need to select any one particular transformation. This parameter becomes part of the uncertainty in the Bayesian modelling approach. They compared trans-Gaussian

kriging with their Bayesian approach and found that their method did better in terms of average prediction error and coverage of 95% prediction intervals.
An approximation to the predictive density can be obtained as

$$p(Z_0|Z) = \frac{1}{m}\sum_{i=1}^{m} p(Z_0|Z, \theta^{(i)}), \tag{4.23}$$

where $\theta^{(i)}$ is the $i^{th}$ sample $i = 1, 2, \ldots, m$, from the posterior distribution $p(\theta|Z)$. The Markov Chain Monte Carlo (MCMC) procedure using the Gibbs sampling method can be used to efficiently simulate from the posterior distribution for the parameters $\theta$. Gibbs sampling involves sampling from conditional distributions. It requires an initial value for the vector $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_q^{(0)})$, and obtains a new value $\theta_1^{(1)}$ from the conditional distribution

$$p\left(\Phi_1|\Phi_2 = \theta_2^{(0)}, \ldots, \Phi_q = \theta_q^{(0)}\right). \tag{4.24}$$

We then obtain a new value $\theta_2^{(1)}$, from the conditional distribution

$$p\left(\Phi_2|\Phi_1 = \theta_1^{(1)}, \Phi_3 = \theta_3^{(0)}, \ldots, \Phi_q = \theta_q^{(0)}\right), \tag{4.25}$$

and generation of $\theta_q^{(1)}$ continues from the conditional distribution

$$p\left(\Phi_q|\Phi_1 = \theta_1^{(1)}, \ldots, \Phi_{q-1} = \theta_{q-1}^{(1)}\right). \tag{4.26}$$

This same procedure is repeated for the new vector $\theta^{(1)}$, and we then return to $\theta_1$ and $\theta^{(2)}$ again until convergence is attained. Futher details about Bayesian geostatistics are described by Handcock and Stein (1993); Ribeiro and Diggle (1999) & (2002); Chiles and Delfiner (1999) and Wackernagel (1995).
The function *krige.bayes* in the *geoR* package of R performs Bayesian analysis of geostatistical data, and it incorporates uncertainty in the estimation procedure. It gives the estimate of the posterior distribution for the model parameters. Bayesian inferences for point and interval estimation as well as hypothesis testing are obtained using the posterior samples. It is possible for different prior distributions to have different effects on the posterior, thus, in our analysis we also assessed the implications of using different priors (Handock & Stein, 1993; Cressie, 1993; De Oliveira et al., 1997; Fuentes, 2007).

## 4.4 Analysis of $SO_2$ data

We computed Moran's I test on the complete dataset to check first for the presence of spatial correlation. After imputation, we first computed the geodata by taking the logarithm of the annual mean $SO_2$. A distance matrix was also obtained from distance weights using the coordinates (longitude and latitude) of the stations. We then take the inverse of the distance matrix after substituting the diagonal elements with zero. Each off-diagonal element $(i, j)$ from the resulting inverse distance matrix is equal to $1/$(distance between points $i$ and $j$).
We consider years 1996, 2000 and 2005 as in the previous analysis in Chapters 1 and 2. The results are presented in Table 4.1, and in this case the output is only significant for the dataset in 1996. The null hypothesis that there is no spatial autocorrelation present in the data is rejected at the 0.05 significance level, which implies that our data are positively correlated though the correlation is very low (corr= 0.1407265) from the observed value of the Moran's I test statistic. The year 2000 and 2005 datasets are not statistically significant in terms of Moran's I. We therefore based most of the analysis in this chapter on 1996 as there are more sites and evidence of spatial correlation.

Table 4.1: Summary of Moran's I test for the datasets in years 1996, 2000 and 2005. The table gives both the observed and expected values as well as the standard deviation and p-values for each year

| | 1996 | 2000 | 2005 |
|---|---|---|---|
| observed | 0.1407 | 0.0801 | -0.0567 |
| expected | -0.025 | -0.0370 | -0.0714 |
| sd | 0.0541 | 0.0826 | 0.0795 |
| p.value | 0.0022 | 0.1562 | 0.8538 |

### 4.4.1 Variogram estimation results

The variogram enables us to look at the variance of the differences of logarithms of annual mean $SO_2$ concentration among pairs of stations at different geodesic distances. Table 4.2 gives the distance summary for the 1996 dataset. The maximum separation distance between any pair of stations is about 270 km and the median distance is approximately 56 km.

Table 4.2: The geodesic distance summary for our geodata in 1996

| distance (km) | | | | | |
|---|---|---|---|---|---|
| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
| 0.001414 | 29.8 | 55.9 | 82.4 | 139.8 | 270.3 |

We calculate a variogram using the Easting and Northing of the stations to determine the geodesic distances that will give us reasonable estimates. Based on the summary of distances in Table 4.2, we consider 13 lag distance intervals for the empirical variogram and created a break vector covering the range of distances for the variogram computation. The lag distance of more than 13 results has an unusual variogram shape in Figures 4.1 and 4.2.

Firstly, we use a visual method to fit the variogram by continously changing the parameters until we have a reasonable estimate in terms of the shape of the variogram. We consider both constant mean and linear trend on the coordinates in the computation of parameters for the estimation of the variogram model. The results are shown in Figures 4.1 and 4.2 for a constant mean and linear trend respectively.

The results seem similar except that the linear trend in Figure 4.2 has lower semi-variance (with maximum value of 0.5) as compared to the constant mean trend (maximum value of 1). The semi-variance increases with increase in lag distance and decreases again beyond 100 km. The nugget parameters are approximately 0.23 and 0.25 for constant mean and linear trends respectively. Both variograms have no clearly defined shape. There is no distinct sill parameter.

We later plotted both the theoretical and empirical variograms together to visually compare them based on the results of initial parameter estimation in Figures 4.1 and 4.2. Having tried various lags distance, we now consider 8 lag distance in order to get a better shape for the variogram. Figures 4.3 and 4.4 show the empirical and estimated theoretical variograms using the Exponential (blue line), Spherical (pink), Matern (green) and Gaussian (red) functions as covariance models.

The sill is not clearly identified at the distances considered and the semi variances do not start at zero, so there is a nugget effect (0.17) in both cases. The empirical variogram possesses a nugget effect (a non-zero limit in the empirical variogram as the distances between stations becomes very small). None of the covariance models fits all of the empirical variogram well. The kriging is useful for prediction at shorter distances within Central Scotland.

Figure 4.1: Empirical variogram for the logarithm of mean annual $SO_2$ levels using a constant mean trend in 1996

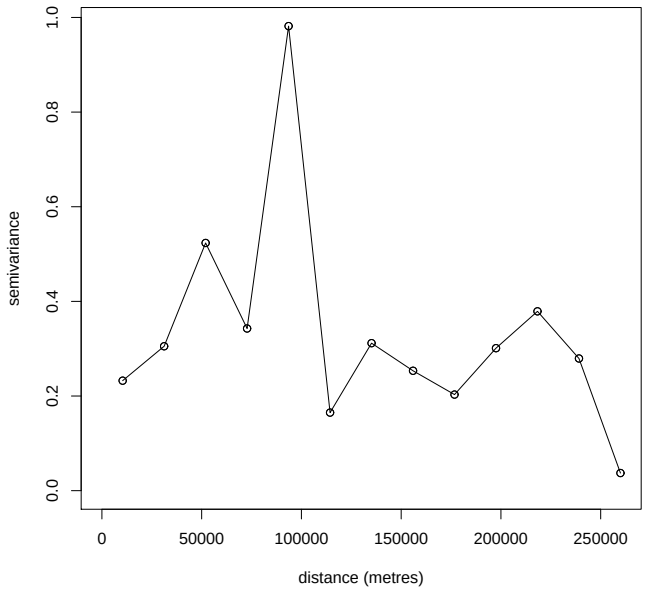



Figure 4.2: Empirical variogram for the logarithm of mean annual $SO_2$ levels using a linear trend in 1996

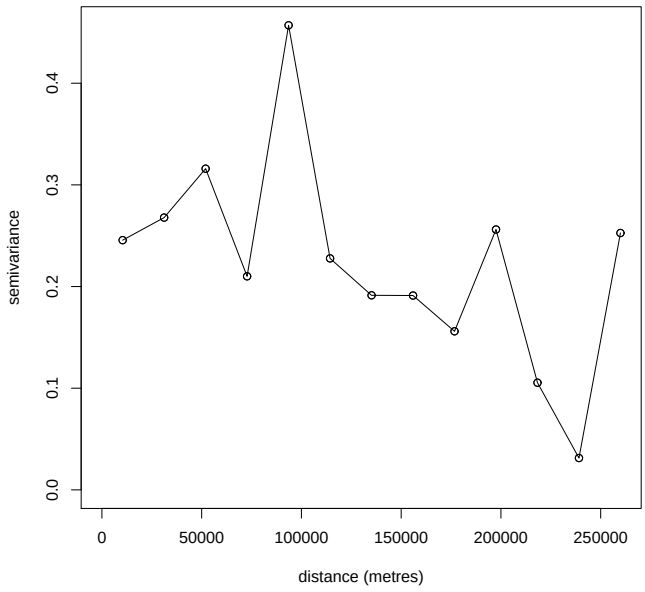


Figure 4.3: Empirical and theoretical variogram for the logarithm of mean annual $SO_2$ levels using a constant mean trend in 1996

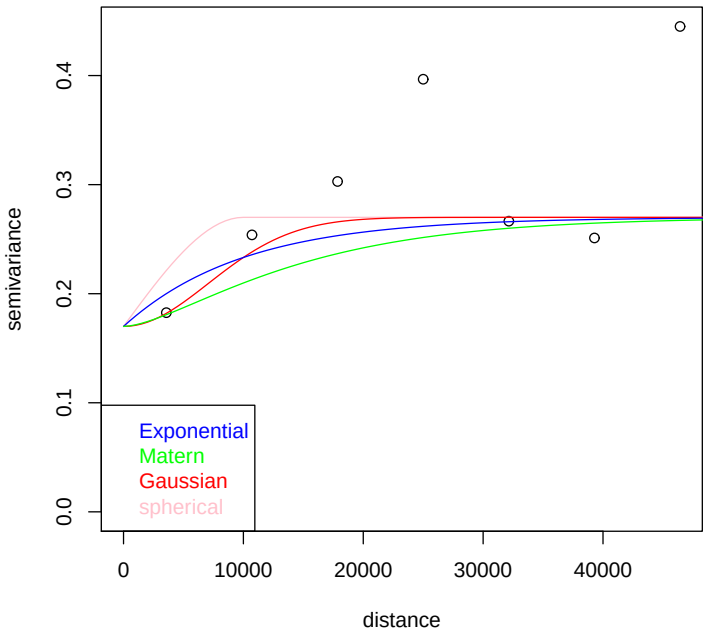
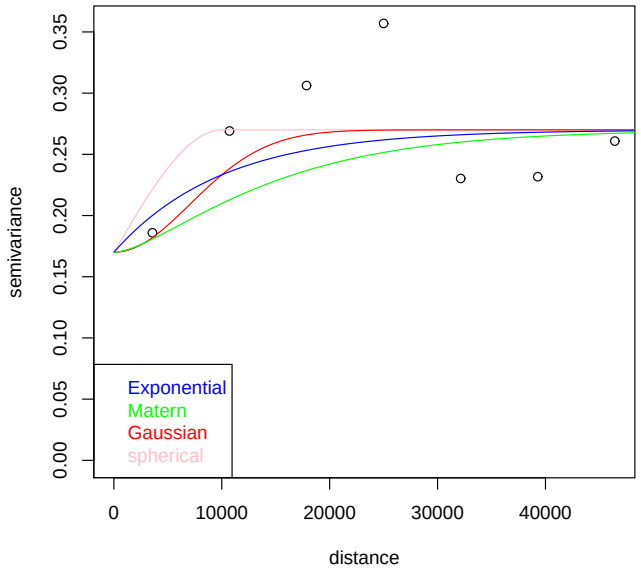



Figure 4.4: Empirical and theoretical variogram for the logarithm of mean annual $SO_2$ levels using a linear trend in 1996



### 4.4.2 Fitting a model variogram

We try to fit an optimal model to our data by computing different models for the sample variogram, and produce a model variogram that fits the sample variogram adequately using the AIC criterion. The variogram model can be used to predict the covariance between points separated by a certain distance and kriging weight for kriging analysis.

We estimated the model parameters using the maximum likelihood method. We used maximum likelihood because it is more computationally feasible than the least squares method and the procedure depends on the assumption of normality, as well as providing effective model comparison and selection criteria such as the likelihood ratio and AIC. Diggle et al. (2003) prefer the likelihood based method, because it is optimal under the model assumptions.

The kriging algorithm requires initial parameter values, and we obtained approximate values for these parameters from the fitted empirical variograms in Figures 4.3 and 4.4. We considered the Exponential, Matern, Spherical and Gaussian covariance models as above, but present results for both Exponential and Matern covariance only because both methods give the minimum AIC and BIC, and the maximum value for the log-likelihood. Table 4.3 shows the estimated variogram parameters using different covariance functions.

Table 4.3 enables us to choose the best (optimal) combination of parameters for the variogram estimation. The model with constant mean trend has the same values for AIC and BIC (75.11, 81.96) for both Matern and Exponential functions respectively, but the Exponential function has the higher log likelihood (-33.55). The linear trend gives the same estimates for both methods and is better than the model with constant mean trend because it has lower AIC and BIC as well as higher log likelihood.

The estimated $\beta$ for the Exponential function (2.774) is slightly higher than for Matern covariance function (2.757). The $\sigma^2$ for Matern is also higher than that of the corresponding Exponential value. Their range parameters $\varphi$ are the same (50 km) (this is a measure of spatial correlation). For the linear trend both sigmasq and phi are relatively low and the two methods give the same estimated values.

Figure 4.5 shows explanatory analysis before the main modelling. The top left panel shows the spatial locations using different colour codes for data in different quartiles to draw more attention to spatial patterning of the variable logarithm of mean $SO_2$. Stations in blue have values less than or equal to the $1^{st}$ quartile, the green ones have values between the $1^{st}$ and $2^{nd}$ quartile, the yellow ones are between the $2^{nd}$ and $3^{rd}$ quantile while the red have values greater than the $3^{rd}$

quantile. The stations are not evenly spatially distributed, but cluster around Central Scotland. We also observe that stations with higher mean level are located within and around Central Scotland as indicated by colour red on the data map.

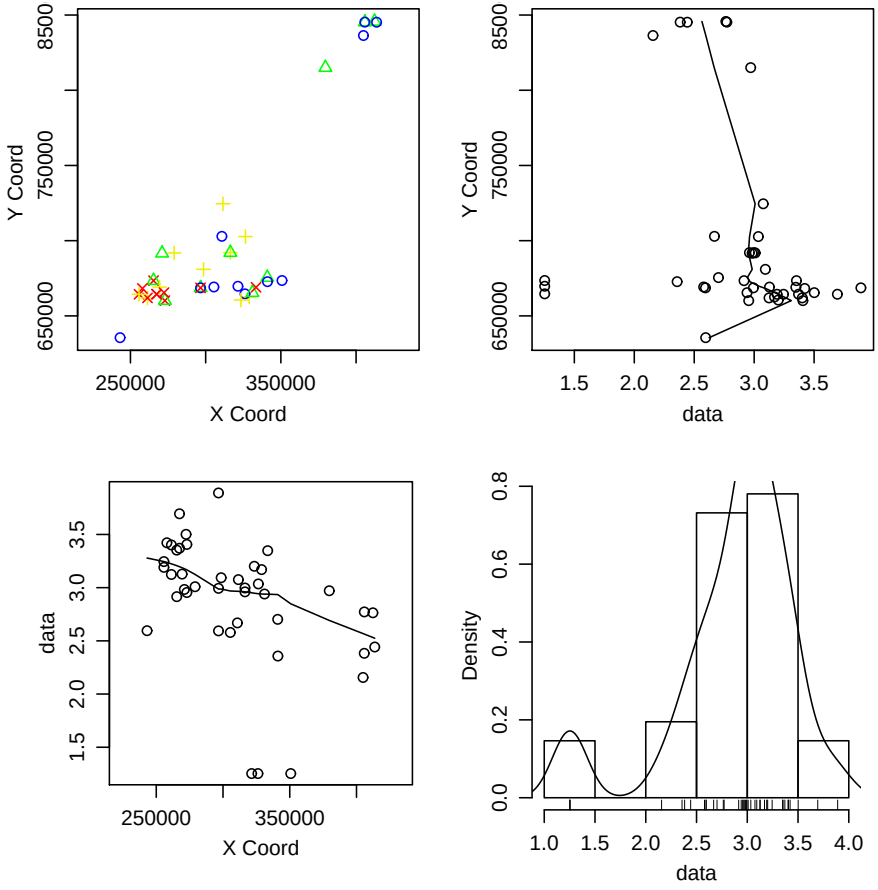
The top right and bottom left panels show the data value against Y (Northing) and X (Easting) coordinates respectively, and the bottom right panel is a histogram of the data values overlaid with an estimated density. The smoothed lines on Easting and Northing suggest that a spatial trend is present in the data (linear trend in x). The histogram is left-skewed and the majority of the observations are between 2.5-3.5.

Table 4.3: Estimated model parameters with constant mean and linear trend using likelihood method for variogram estimation

| constant mean trend | | | | | | | |
|---|---|---|---|---|---|---|---|
| function | beta | tausq | sigmasq | phi | log.L | AIC | BIC |
| matern | 2.757 | .2556 | .1128 | 50000 | -33.70 | 75.11 | 81.96 |
| exp. | 2.774 | .2437 | .09930 | 50000 | -33.55 | 75.11 | 81.96 |

| linear trend | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | beta_0 | beta_1 | beta_2 | tausq | sigmasq | phi | log.L | AIC | BIC |
| mat. | 2.6726 | -0.0011 | 0.0005 | 0.2363 | 0.0018 | .082995 | -28.6 | 69.2 | 79.48 |
| exp. | 2.6726 | -0.0011 | 0.0005 | 0.2363 | 0.0018 | .082995 | -28.6 | 69.2 | 79.48 |

Figure 4.5: Plotting data locations and values. Stations in blue colour coding have values less than or equal to the $1^{st}$ quartile, the green ones have values between the $1^{st}$ and $2^{nd}$ quartile, the yellow ones are between the $2^{nd}$ and $3^{rd}$ quantile while the red have values greater than the $3^{rd}$ quantile




### 4.4.3 Ordinary and Bayesian kriging

We now use kriging for spatial interpolation. We consider both ordinary and Bayesian kriging in our analysis. We use likelihood estimation of parameters, considering both Matern and Exponential covariance models for kriging, and compare results for both constant mean and linear trends.

For our initial ordinary kriging modelling we choose both Matern and Exponential functions with constant mean trend, using the likelihood method because it gives a high log likelihood (-33.55) and lowest AIC (75.11) in Table 4.3. The estimated variogram parameters are $\tau^2$ (0.2437) which is a measure of the relative nugget effect, $\sigma^2$ (0.09930), a measure of the sill and $\varphi$ of 50 km which is also a measure of the range parameter (the distance at which the sill is reached) from Table 4.3. These values are used in the subsequent kriging estimation.

The ordinary kriging results are shown in Tables 4.4 and 4.5, which give the detail

of estimated model parameters ($\beta$, $\tau^2$, $\sigma^2$ and $\varphi$) and the summary statistics of the estimated predicted mean and its kriging variance. The graphical results are shown in Figures 4.6-4.14.

In model 1a (Table 4.4), we utilized default options in the package, that is a Matern covariance model with constant mean trend, while model 1b in Table 4.5 used the same parameter values but with the Exponential function. In Table 4.4, the estimated parameters are $\beta$ of 2.77, which is an estimate of the mean, $\tau^2$ of 0.244 which is an estimate of the nugget effect, $\sigma^2$ of 0.0993 which estimates the variance of the model and $\varphi$ of 50 km which is a measure of the range for the correlation (which simply implies that any observation within this distance may be assumed to be spatially correlated). The asymptotic range is about 149786.6 (150 km) which is 3 times the value of estimated $\varphi$ (any observations beyond this distance are not correlated). The median and mean predictions are similar (2.774). The mean and median of kriging variance are also very low.

In model 1b (Table 4.5), the Exponential function was utilized rather than Matern, but the summary results are identical to the Matern estimation. In Figure 4.6, we observe that the estimated mean concentration is very high in Central Scotland, and there is a decreasing trend in mean concentration towards the boundary, i.e there is reduction in the mean level indicated by gradual darkening of the estimated surface. We will not pay attention to the unusual high predictions observed outside the Scotland map as these predictions are not reliable due to the scarcity of data in the sea area (Figures 4.6, 4.8, 4.10, 4.13, 4.16 and 4.18).

Central Scotland comprises the Glasgow area and Edinburgh, and the high predictions we observe here may be due to presence of heavy industries and high population density in this region. The remote stations toward the North-Eastern region (Aberdeen 3, Peterhead 1, Peterhead 2, Peterhead 3, Longside 2 and Hatton 1) have very low mean concentrations, which may be due to the low population density in the area. In Figure 4.7 variance generally increases towards the coast with Central Scotland having the least predicted variance, which could be due to clustering of stations and availability of data in this region.

In Figure 4.8, we have similar results to the Matern covariance model with Central Scotland still having the highest predicted mean, which is also largely due to presence of more data in the region. The North-Eastern region has the least predicted concentration. In Figure 4.9, the results are the same as for the Matern model with variance increases towards the coast, and Central Scotland and North-Eastern region have very low predicted variance.

In summary, the predictions in the West are higher than in Edinburgh and Ab-

erdeen (Figures 4.6 and 4.8). This is due to a lack of data thus extrapolating too far outside the spatial range of the stations.

Table 4.4: Likelihood fit result for the estimated model parameters of the ordinary kriging using the Matern covariance function with constant mean trend and summary statistics of the estimated prediction mean and variance

```
Model 1a
beta     tausq   sigmasq  phi
2.774    0.2437  0.09930  50000
Practical Range with cor=0.05 for asymptotic
range: 149786.6
likfit: maximised log-likelihood = -33.55
-------------------------------------------
> summary(kr1a$predict)
   Min. 1st Qu.  Median    Mean   3rd Qu.  Max.
  2.473   2.760   2.772   2.774   2.782   3.181
-------------------------------------------------
> summary(kr1a$krige.var)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2714  0.3647  0.3756  0.3673  0.3787  0.3802
```

Table 4.5: Likelihood fit result for the estimated model parameters of the ordinary kriging using the Exponential covariance function and summary statistics of the estimated prediction mean and variance using constant mean trend

```
Model 1b
beta    tausq    sigmasq  phi
2.774   .2437   .09930   50000
Practical Range with cor=0.05
for asymptotic range: 149786.6
likfit: maximised log-likelihood = -33.55
-------------------------------------------------
summary(kr1b$predict)
   Min.   1st Qu.  Median  Mean    3rd Qu. Max.
  2.473   2.760    2.772   2.774   2.782   3.181
-------------------------------------------------
> summary(kr1b$krige.var)
   Min. 1st Qu.  Median  Mean 3rd Qu.    Max.
 0.2714  0.3647  0.3756  0.3673  0.3787  0.3802
```

Figure 4.6: Ordinary kriging predicted mean using maximum likelihood (method=ml, cov=matern, trend=constant mean trend). The high predictions observed outside the map region are not reliable

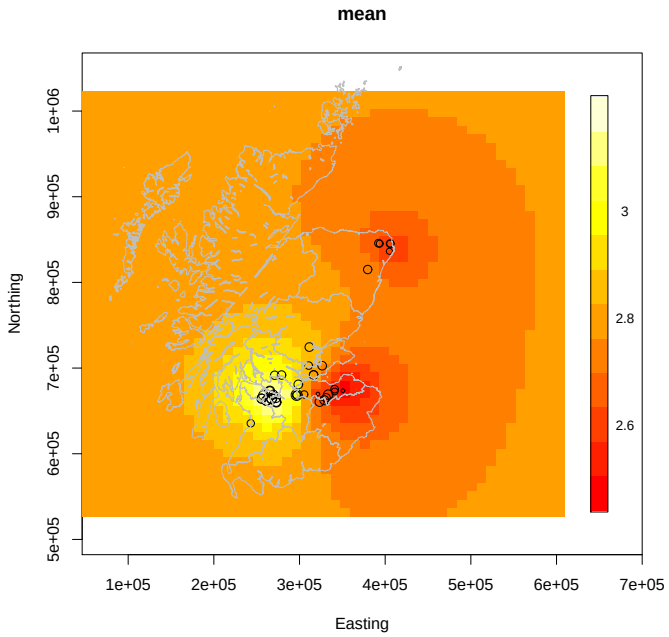



Figure 4.7: Ordinary kriging predicted variance using maximum likelihood (method=ml, cov=matern, trend=constant mean trend)

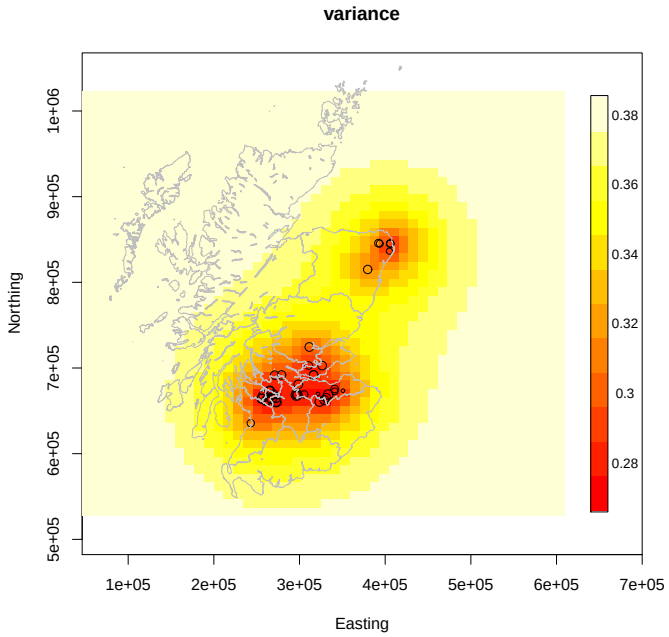


Figure 4.8: Ordinary kriging predicted mean using maximum likelihood (method=ml, cov=exponential, trend=constant mean trend). The high predictions observed outside the map region are not reliable

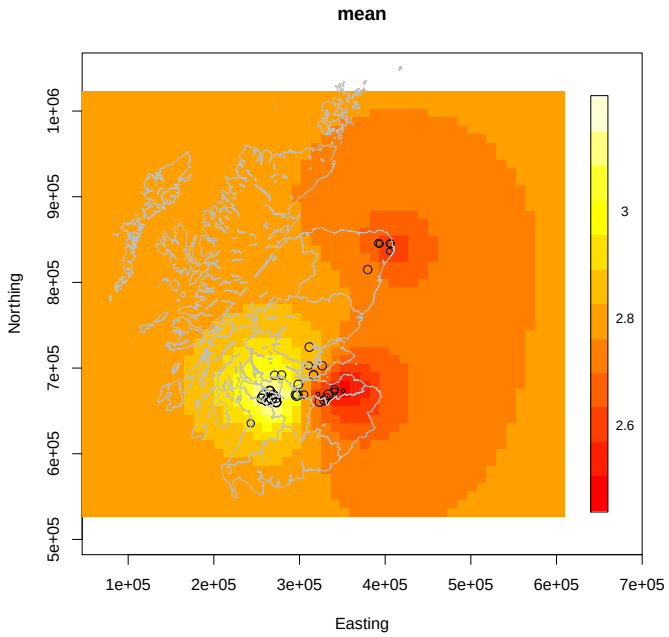
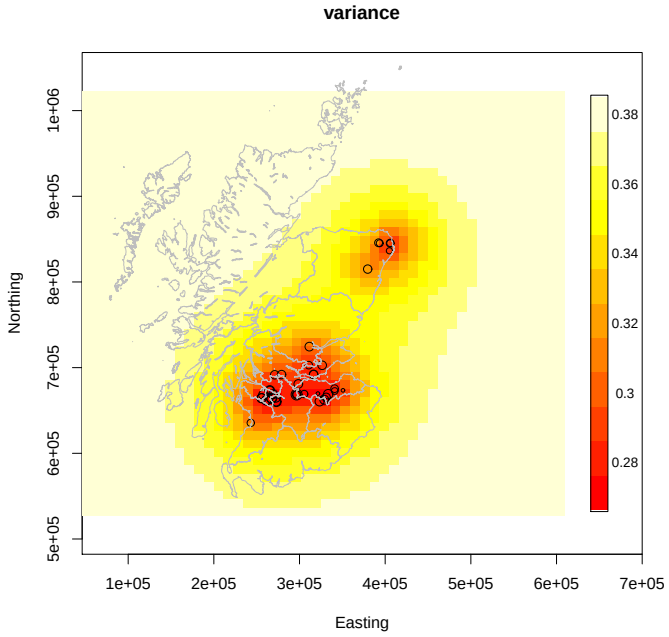



Figure 4.9: Ordinary kriging predicted variance using likelihood (method=ml, cov=exponential, trend=constant mean trend)



### 4.4.4 Bayesian results

We also utilized Bayesian kriging to estimate model parameters and obtain predictions for the logarithm of the mean annual $SO_2$ concentration at unobserved locations within Scotland for us to see if we could get similar or better results than in ordinary kriging as a little prior knowledge will be assumed here. Bayesian inference on the parameters involves setting up priors. The default setting takes the prior for the mean $\beta$ to be $flat$, $reciprocal$ for the parameter $\sigma^2$ and $uniform$ for the range parameter $\varphi$. Samples of 1000 simulations are taken from the posterior distribution as well as the predictive distribution.

Firstly, we used these default prior specifications with a constant mean trend and the Matern covariance model. The prediction results and posterior distribution of the parameters are given in Table 4.6 and Figures 4.10 and 4.11 (Model 2a). In Table 4.6, the Bayesian predictive mean and median are the same (2.76) and slightly lower than the ordinary kriging result in Table 4.4. The mean predictive variance is (0.00546) much lower than that of ordinary kriging. In Table 4.7 (model 2b) shows results for the corresponding Exponential covariance model. We investigate this model in order to assess the sensitivity of the results to the priors. The range of the predictive mean is also very low with predictive mean of 2.765 which is similar to the Matern model in Table 4.6. The estimated Bayesian variance is similar to that of corresponding Matern function.

In Figure 4.10, the result is similar to the ordinary kriging. There is a variation in predictive mean level in different locations on the map, with a higher mean level of $SO_2$ concentration observed in Central Scotland as indicated by the lightening of the surface, and the North-Eastern stations still have lower predicted mean levels. We observe that Bayesian kriging seems to have more impact on the predicted mean in the North-Eastern region, as Aberdeen now has a higher predicted mean than the other North-Eastern stations. There is a wider range in the predictive mean than in the ordinary kriging.

In Figure 4.11, there is an decrease in the estimated variance towards the boundary (higher variance at data points) and the estimated Bayesian variance is lower than that of ordinary kriging. There is lower variance in the areas away from the stations and Central Scotland still has high Bayesian variance. Figure 4.12 shows the histograms of the posterior distributions for the Bayesian parameter estimates. For $\beta$, this is symmetric, for $\sigma^2$ it is right skewed, and for $\varphi$ it is more or less a uniform distribution which implies that there is no correlation at higher distance.

Figures 4.13 and 4.14 are very similar to Figures 4.10 and 4.11. The results are

not too sensitive to the choice of covariance model because the predictive mean and variance estimates results are similar in the two models.

Table 4.6: Likelihood fit result for the estimated model parameters of Bayesian kriging using the Matern covariance function and constant mean trend with a flat distribution for the mean $\beta$, a reciprocal prior for the variance, and a uniform distribution for the range parameter

```
Model 2a
> summary(kr2a$predictive$mean)
   Min. 1st Qu.  Median   Mean    3rd Qu. Max.
  1.624   2.759   2.760   2.760   2.760   3.351
-----------------------------------------------
> summary(kr2a$predictive$variance)
 Min.    1st Qu. Median  Mean    3rd Qu. Max.
0.00512 0.00523 0.00529 0.00546 0.00537 0.0328
```

Table 4.7: Likelihood fit result for the estimated model parameters of Bayesian kriging using the Exponential covariance function and a constant mean trend with a flat distribution for the mean $\beta$, a reciprocal prior for the variance, and a uniform distribution for the range parameter

```
Model 2b
 summary(kr2b$predictive$mean)
   Min. 1st Qu.  Median   Mean 3rd Qu.    Max.
  1.625   2.765   2.765   2.765   2.765   3.351
-------------------------------------------------
> summary(kr2b$predictive$variance)
 Min. 1st Qu.     Median  Mean    3rd Qu. Max.
 0.00512  0.00523 0.00529 0.00546 0.00537 0.0328
```

Figure 4.10: Mapping of Bayesian predictive mean estimate for the Matern function using the default setting for the priors (mean $\beta = flat$, variance $\sigma^2$ = reciprocal, range parameter $\varphi$ = uniform, trend=constant mean trend). The high predictions observed outside the map region are not reliable

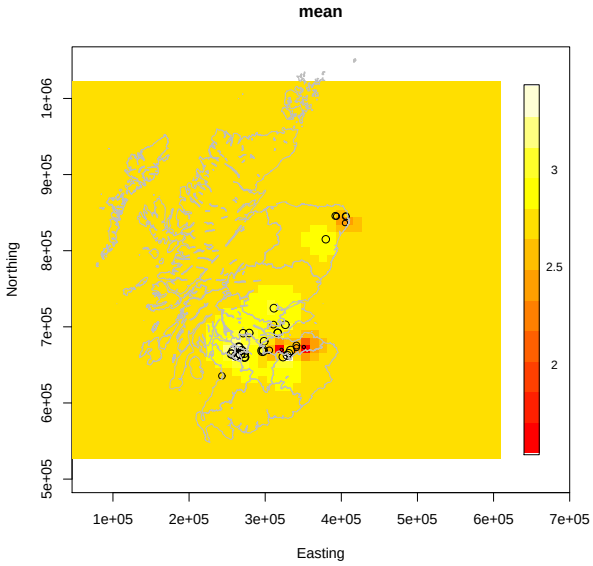



Figure 4.11: Mapping of Bayesian predictive variance estimate for the Matern function using the default setting for the priors (mean $\beta = flat$, variance $\sigma^2$ = reciprocal, range parameter $\varphi$ = uniform, trend=constant mean trend)

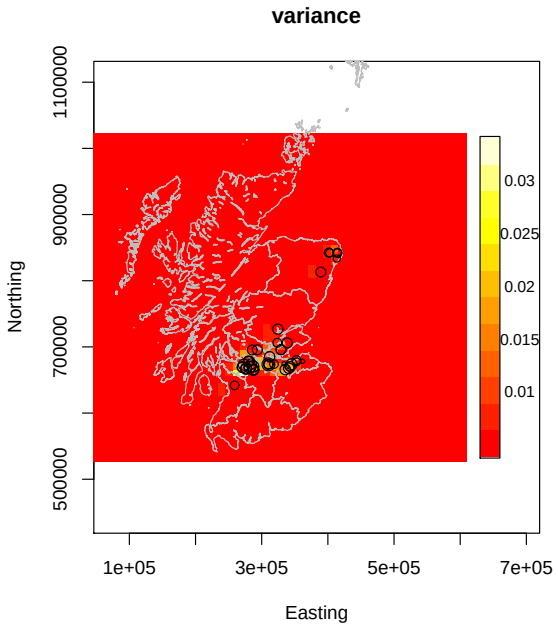


Figure 4.12: Histograms of the posterior distribution of the Bayesian predictive estimates ($\beta$, $\sigma^2$, $\varphi$) using the default setting for the priors (mean $\beta = flat$, variance $\sigma^2$ = reciprocal, range parameter $\varphi$ = uniform, trend=constant mean trend). The upper box is for Matern and the lower box is for the Exponential function

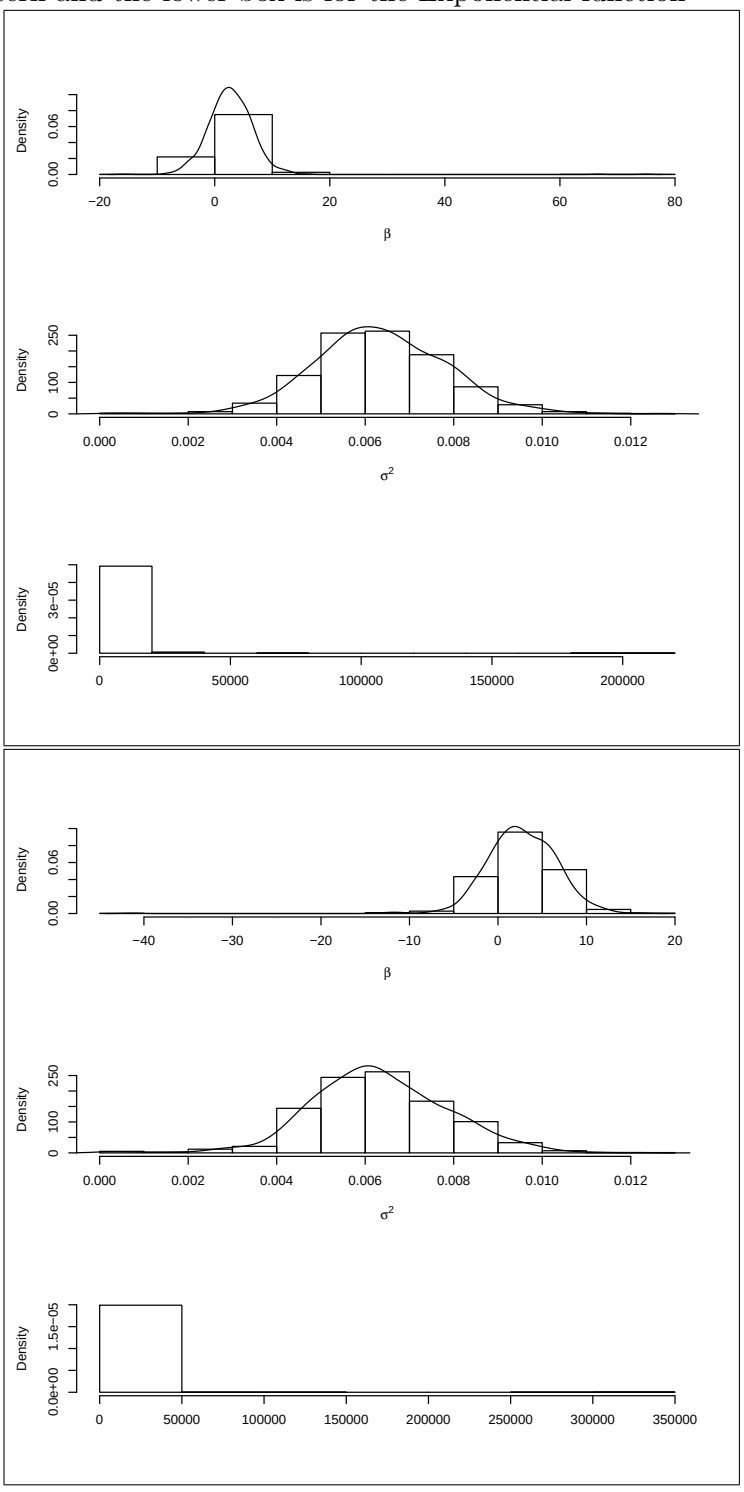


Figure 4.13: Mapping of Bayesian predictive mean estimate for the Exponential function using the default setting for the priors (mean $\beta = flat$, variance $\sigma^2$= reciprocal, range parameter $\varphi$ =uniform, trend=constant mean trend). The high predictions observed outside the map region are not reliable

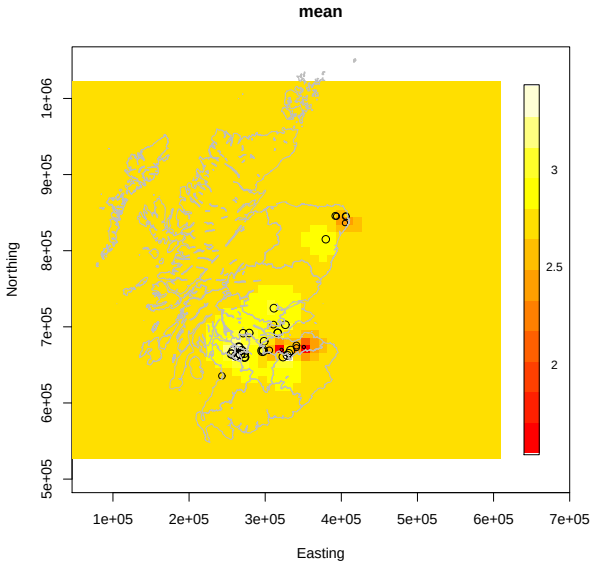
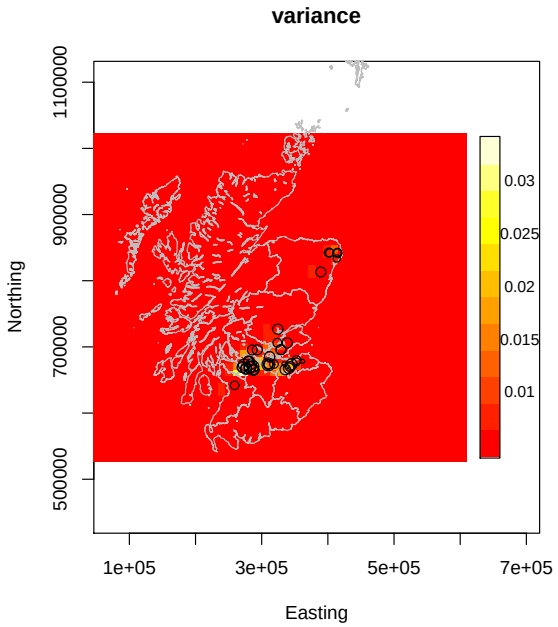



Figure 4.14: Mapping of Bayesian predictive variance estimate for the Exponential function using the default setting for the prior (mean $\beta = flat$, variance $\sigma^2$= reciprocal, range parameter $\varphi$ = uniform, trend=constant mean trend)



## Models involve linear trend

In Model 3, we continue our analysis by constructing another set of models similar to model 1a in Table 4.4 (ordinary kriging) but using a linear trend to account for the trend to see if there is any effect on estimated parameters as a result of change in trend specification. The results are shown in Table 4.8.

In Table 4.8, the likelihood based parameters are generally very low. The parameters $\beta_1$, $\beta_2$, $\sigma^2$ and $\varphi$ are all very small (as low as 0) with $\tau^2$ of 0.2363, a little lower than for model 1a. The better log likelihood in Table 4.8 may be due to removal of trend from the series which make the general mean level become very low (the series we modelled here comprises only the seasonal component and residual variation after removal of trend). The model indicates a higher mean prediction of 3.130 than the mean observed data 2.882 (from Table 4.8), it has a wider predicted range (-1.252, 7.512), indicating some negative predictions, the mean level is also higher than that of constant mean trend, and the kriging variances are very small (close to 0 at 4 d.p), which suggest that the algorithm has not converged, thus this result may not be too reliable.

We repeated the same analysis with the same model parameters but using an Exponential function for the covariance model. The results are not shown here but show a similar pattern. For the Bayesian model with linear trend using default prior specification and Matern covariance model, Table 4.9 indicates that the mean Bayesian prediction is also higher than the average observed data ($3.041 > 2.882$). There are also some negative predictions in these results. The estimated Bayesian variance is very much lower compared to the corresponding ordinary kriging variance results. The results are not similar to the constant mean trend model.

It is generally observed in Tables 4.8 and 4.9 that the model result is not yet converged. We do not present map of predicted mean and variance for the results in this section as some predictions are negative.

Table 4.8: Likelihood fit result for the estimated model parameters of the ordinary kriging results using Matern covariance function, and summary statistics for the estimated predicted mean and variance using linear trend

```
Model 3
likfit: estimated model parameters:
beta_0  beta1     beta2     tausq  sigmasq    phi
2.6726  0.0000001 0.0000001 0.2363 0.000000  0.0000
Practical Range with cor=0.05 for asymptotic
range: 0.0001159668
likfit: maximised log-likelihood = -28.6
Observed data
> summary(g8$data)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.253   2.669   2.994   2.882   3.201   3.890
-----------------------------------------------------
summary(kr3a$predict)
   Min. 1st Qu.  Median   Mean 3rd Qu.    Max.
 -1.252   1.548   3.130   3.130   4.712   7.512
-----------------------------------------------------
> summary(kr3a$krige.var)
   Min.    1st Qu.  Median  Mean    3rd Qu.  Max.
   0.0000  0.0000   0.0000  0.0000  0.0000   0.0000
```

Table 4.9: Likelihood fit result for the estimated model parameters of Bayesian kriging using Exponential covariance function, and summary statistics for the predicted mean and variance using linear trend

```
Model 4
likfit: estimated model parameters:
beta0   beta1    beta2    tausq  sigmasq  phi
2.9700  0.0000   0.0000   0.2361 0.0000   0.0000
Practical Range with cor=0.05 for
asymptotic range: 0.0001159668
likfit: maximised log-likelihood = -28.58
-------------------------------------------------
summary(kr3b$predictive$mean)
  Min.    1st Qu. Median  Mean    3rd Qu. Max.
 -1.106   1.599   3.045   3.041   4.477   7.186
-------------------------------------------------
> summary(kr3b$predictive$variance)
Min.    1st Qu. Median  Mean    3rd Qu. Max.
0.00013 0.00051 0.00096 0.00151 0.00183 0.03100
```

We now consider Central Scotland only by computing a separate model for the stations in Central Scotland alone without the six remote stations (Aberdeen 3, Peterhead 1, Peterhead 2, Peterhead 3, Longside 2 and Hatton 1). We do this to check how sensitive our previous models would be if the stations are clustered together in a region, and to overcome the problem of high predictions in the Western Coast and over the sea we earlier observed in Figures 4.6 and 4.8. We now have more correlated data as a result of reduction in distance among pairs of stations.

Figure 4.15 shows the plots of the data values and locations for the Central Scotland stations. We observe more clustered data points with relatively low variation, than for the whole of Scotland. There is not much effect of Northing while the Easting shows a decreasing straight line trend. The histogram indicates that most of the data values fall between 2.5-3.5 and is left-skewed (compare Figure 4.5), as before.

We re-analysed Model 1a (ordinary kriging) with the remaining 35 stations. The results are presented in Table 4.10. We observe that the $\beta$ parameter of 2.852 is now higher than that of the whole of Scotland (2.774) in model 1a (Table 4.4), and that both $\tau^2$ and $\sigma^2$ values are also higher than the corresponding results in model 1a, but the $\varphi$ parameter has now reduced from 50km to 40 km. The asymptotic range has also reduced to approximately 120 km, with a higher log likelihood (-30.34). The predictive mean levels are generally higher than for model 1a, and the corresponding kriging variance value is also higher, which could easily be due to reduction in data points.

The corresponding Bayesian results are presented in Table 4.11 and we also compare these results with model 2a in Table 4.6. The predictive mean level is still higher, while the predicted variance increases as compared to a full model where all the stations are used. The mean predictions are lower than for the observed data ($2.849 < 2.933$, from Table 4.10 and $2.817 < 2.933$, from Table 4.11).

In Figure 4.16, similarly to Model 1a results, the highest prediction is found in Glasgow Centre because of concentration of heavy industry, city centre business and high population density as well as large volume of vehicles, and there is a low predicted level along the Eastern Coast. In Figure 4.17, the estimated kriging variance decreases towards the coast, with the Central region (the area with highest concentration of stations) having a very high variance level. Spatial variation is also relatively small, which may be due to clustering of the stations in the region.

The Bayesian kriging results for Central Scotland alone are shown in Figures 4.18

and 4.19, and the results are also consistent with the ordinary kriging, with highest predictions in Glasgow Centre and low predictions along the Eastern Coast. The variances are also higher in areas close to recording stations.

We also considered a separate model for these 6 remote stations by computing another model for the 6 stations using Bayesian kriging. The results are shown in Table 4.12 and Figures 4.20 and 4.21. These results under-estimated the observed data, as the mean prediction is lower than the average observed data ($2.612 < 2.88$). The estimated Bayesian variance is also very high compared to the Central Scotland model. Spatial variation is also relatively small, probably due to proximity of the stations. Also, Aberdeen 3 has higher predictions than rest of the stations in that region as indicated by white colour in Figure 4.20.

Figure 4.15: Plotting data locations and values for the Central Scotland. Stations in blue colour coding have values less than or equal to the $1^{st}$ quartile, the green ones have values between the $1^{st}$ and $2^{nd}$ quartile, the yellow ones are between the $2^{nd}$ and $3^{rd}$ quantile while the red have values greater than the $3^{rd}$ quantile

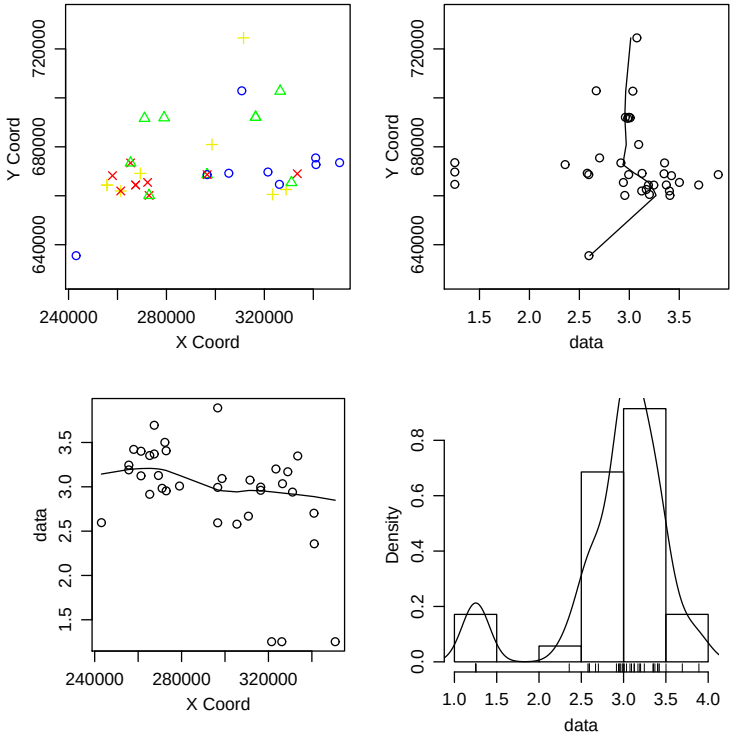


Table 4.10: Summary statistics for the estimated model parameters of the ordinary kriging using the Matern covariance function for Central Scotland only

```
Model 5a
likfit: estimated model parameters:
beta     tausq    sigmasq   phi
2.852    0 .27    0.1012    40000
Practical Range with cor=0.05
for asymptotic range: 119829.3
likfit: maximised log-likelihood = -30.34
Observed data
> summary(g8b$data)
  Min.   1st Qu. Median  Mean   3rd Qu. Max.
 1.253   2.809   3.035   2.933  3.296   3.890
-----------------------------------------------
> summary(kr7aa$predict)
   Min.  1st Qu.  Median Mean   3rd Qu. Max.
  2.502  2.845   2.849   2.849  2.854   3.187
-----------------------------------------------
> summary(kr7aa$krige.var)
   Min.  1st Qu. Median   Mean   3rd Qu.  Max.
 0.3032  0.4255  0.4351  0.4253  0.4372  0.4379
```

Table 4.11: Summary statistics for the estimated model parameters of Bayesian kriging using the Matern covariance function for Central Scotland

```
 Model 5b
likfit: estimated model parameters:
 beta     tausq    sigmasq   phi
2.852     0.27      0.1012   40000
Practical Range with cor=0.05
for asymptotic range: 114162.7
likfit: maximised log-likelihood = -30.39
Observed data
> summary(g8b$data)
  Min.   1st Qu. Median  Mean   3rd Qu. Max.
 1.253   2.809   3.035   2.933  3.296   3.890
-----------------------------------------------
> summary(kr8b$predictive$mean)
  Min.  1st Qu. Median  Mean   3rd Qu. Max.
  1.904   2.817 2.817   2.817  2.817   3.273
-----------------------------------------------
> summary(kr8b$predictive$variance)
 Min.   1st Qu. Median  Mean    3rd Qu. Max.
0.01143 0.01143 0.01143 0.01149 0.01143 0.03041
```

Figure 4.16: Ordinary kriging result for the predicted mean for Central Scotland. The high predictions observed outside the map region are not reliable

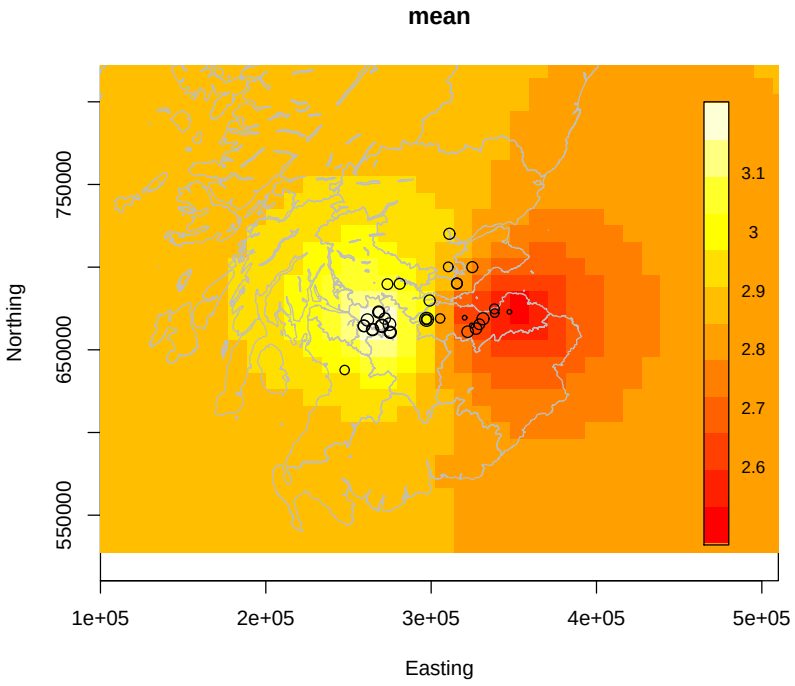



Figure 4.17: Ordinary kriging result for the predicted variance for Central Scotland

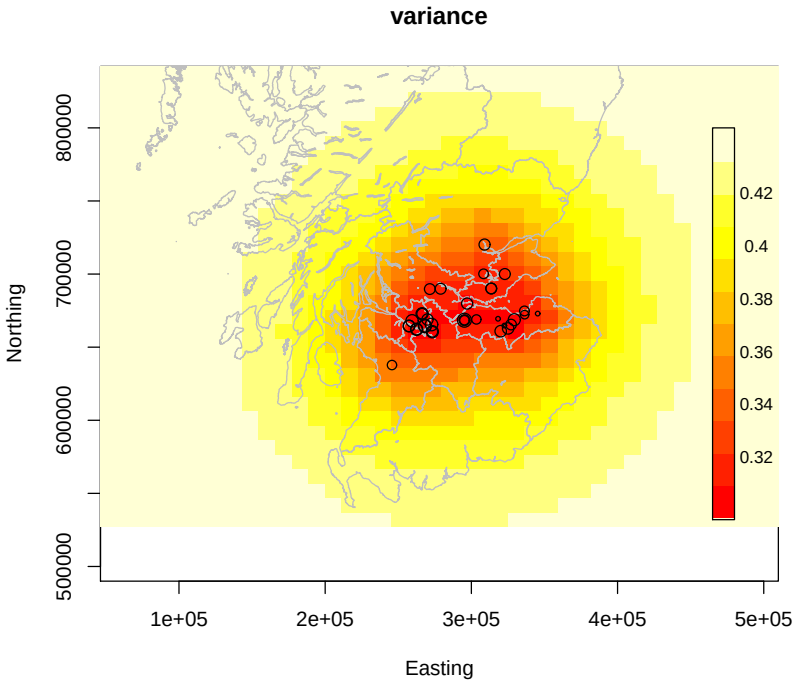


Figure 4.18: Bayesian kriging results for the predicted mean for Central Scotland. The high predictions observed outside the map region are not reliable

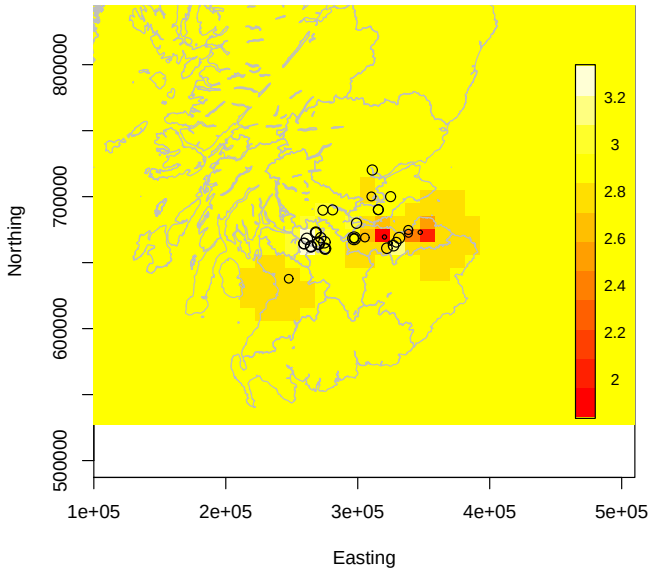



Figure 4.19: Bayesian kriging results for the predicted variance for Central Scotland

Table 4.12: Summary statistics for the estimated model parameters of the Bayesian kriging using the Matern covariance function for the remote stations

```
Model 6
Observed data
> summary(g8c$data)
 Min.   1st Qu. Median  Mean  3rd Qu. Max.
 2.156  2.398   2.603   2.881 2.770   2.971
-----------------------------------------------

summary(kr9$predictive$mean)
  Min.    1st Qu. Median  Mean   3rd Qu. Max.
  2.398   2.609   2.612   2.612  2.614   2.795
-----------------------------------------------

> summary(kr9$predictive$variance)
 Min.    1st Qu. Median  Mean    3rd Qu.  Max.
 0.04343 0.04406 0.04529 0.04984 0.04834  0.40950
```

Figure 4.20: Bayesian kriging results for the predicted mean for remote stations

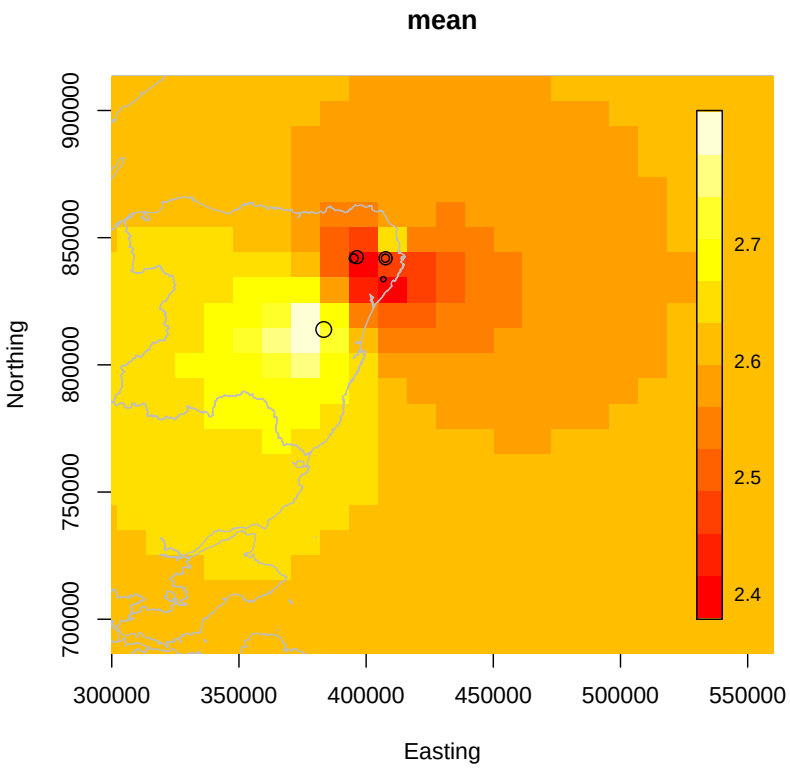
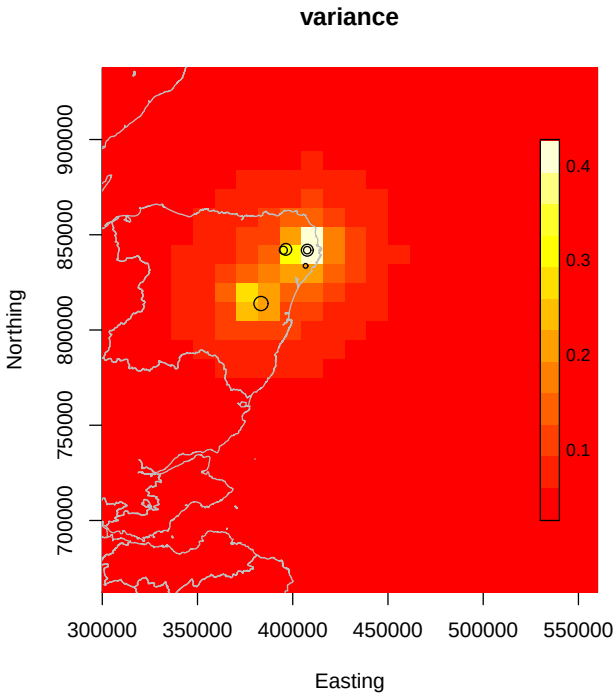


Figure 4.21: Bayesian kriging results for the predicted variance for remote stations




### 4.4.5 Model validation

In order for us to assess the spatial predictive capability of our models, we construct a validation procedure, to assess how well kriging works for our dataset. We perform model validation by comparing observed and predicted values using kriging. The various methods include leave-one-out, in which each data point is removed one by one from the given observations. The observations at unknown locations are predicted based on the remaining data at known locations. This can be done with a subset of the data points or all the data. We can also perform external validation using validation points other than data points. The function *xvalid* in geoR performs model validation within any of the three options.

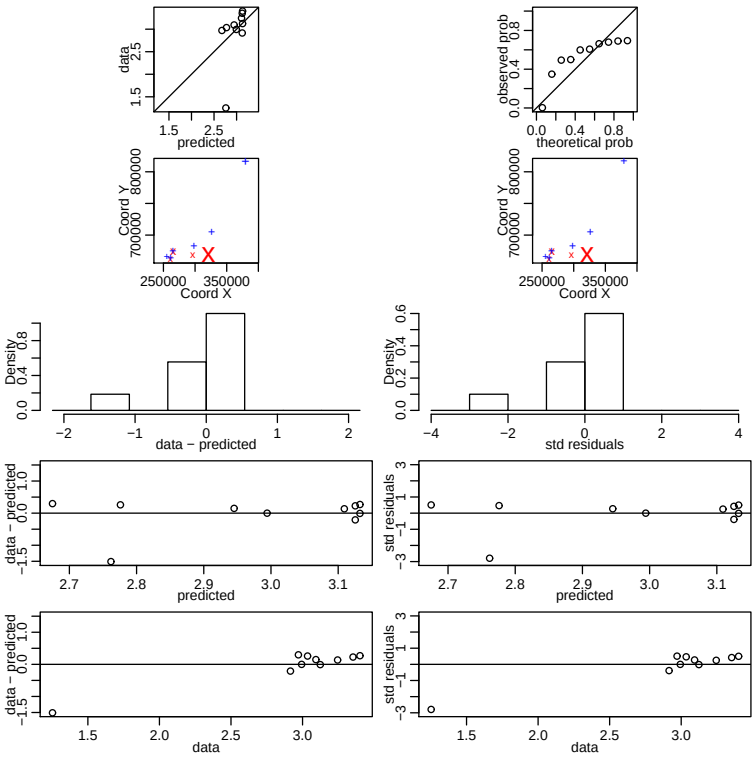
In our analysis our dataset is divided into 2 different groups, namely test and training data. The test data is obtained by randomly removing 10 stations from our dataset, while the other observations are the training dataset. We fitted an ordinary kriging using the Matern covariance function to the training data, and then used the results to validate prediction of the 10 test observations. Having previously fitted a likelihood model for parameter estimation to the logarithm of annual mean $SO_2$ concentration for all the stations, we used the results obtained to validate our models. Table 4.13 shows the summary of observed test data and

summary statistics of error as well as standard errors of prediction.
We observe that the predictions for the test data fall within the range of the corresponding full dataset, which is an indication that the validation results are compatible with the observed data. We also compare our results based on standard error (prediction error divided by the square root of the kriging variance). The standard error of the mean prediction error (-0.07841) is also very small, which is another indication of reliability of the models. The model validation plot in Figure 4.22 suggests that the model is relatively adequate, except for one point which is not well predicted (prediction=2.8, when $data < 1.5$, this point corresponds to Aberdeen 3). There is constancy of variance though the PP plot is not normal.

Table 4.13: Model validation results which show the summary statistics of the observed data, its prediction and error of prediction for the test data

```
 summary(xval1$data)
Observed data
  Min.   1st Qu. Median  Mean   3rd Qu. Max.
  1.253  2.977   3.064   2.938  3.215   3.401
-----------------------------------------------------------------
> summary(xval1$predicted)
predicted data
  Min.   1st Qu. Median  Mean   3rd Qu. Max.
  2.675  2.818   3.052   2.978  3.125   3.132
-----------------------------------------------------------------
summary(xval1)
prediction error
             Min.   1st Qu. Median    Mean    3rd Qu. Max.   sd
errors     -1.509 -0.0066  0.1420   -0.0393  0.2511  0.2961 0.5401
std.errors -2.793 -0.0122  0.2609   -0.0784  0.4546  0.5083 0.9959
```

Figure 4.22: Model validation results which show the histograms of observed data, PP plot and standardized residuals; blue indicates positive values of the error "data-predicted" and red indicates negative values




## 4.5 Summary and suggestions for further work

### 4.5.1 Summary

The results of the model validation are presented first since both inference and predictions are based on the model with the best predictive ability. Generally, the models with constant mean trend seem better than those with linear trend, because the predictions are all in the range of the observed data and there are no negative predictions, and the estimated variances are low. The Exponential covariance function tends to drastically reduce the estimated variance from kriging based on our models.

We also observed that the Bayesian models produce a lower variance than ordinary kriging estimates. In our analysis, Bayesian approach under-estimates the prediction variances. Bayesian approach outperforms ordinary kriging result as a reult of smaller kriging variance.

Ordinary kriging does not display the underlying spatial pattern very well when using all 41 stations, especially for the stations around Aberdeen. There is an increase in variance with the reduced model after removal of stations that are far

away from the rest. A low spatial variation is observed in Central Scotland and this could be due to concentration of the stations in the region. Of course we expect the data to be highly spatially correlated (dependent).

Most of the formulated models indicated high concentration of $SO_2$ in and around Central Scotland (Glasgow and Edinburgh), and elevated concentrations are also seen in parts of Eastern Scotland (which may possibly be attributed to altitude and wind direction as well as the location of the primary source of $SO_2$ in the region), while low mean level is observed in North-Eastern locations (remote stations except Aberdeen) and along the Eastern coast. There is a reduction in the Bayesian variance as the data points reduce in number.

Also, we can infer from the models that Bayesian estimation shows little sensitivity to the choice of priors and initial covariance model. The models with constant mean trend tend to have higher estimated variance than that of linear model for both ordinary and Bayesian kriging.

Non-availability of data in most of the stations we considered and the large presence of missing data limits most of our analysis to 1996 and 1997 which have most recording stations with fewest missing observations. The density of the monitoring stations in Central Scotland tends to reduce the error related to spatial variation, as is observed in the Bayesian model for Central Scotland only.

The low value of the kriging variance for the Bayesian model in Tables 4.6, 4.7, 4.9, 4.11 and 4.12 can also be attributed to incorporation of uncertainty both in the trend and covariance parameters.

Lastly, lack of adequate knowledge about the nature of the distribution of $SO_2$ may also be a source of high prediction error in our Bayesian modelling as it is not clear which are the best values of the parameters and distribution for the spatial analysis.

### 4.5.2 Further considerations

Increasing the dataset to incorporate many more stations is likely to reduce the kriging variance. Also averaging $SO_2$ levels over more than a year (rather than for only year 1996) may also reduce the prediction error to give a spatial model that will allow more effective prediction which is only valid if there is no temporal change. In Chapter 5 we fit a generalized additive model that addresses this.
Lastly, further improvement may also be gained from taking into account other covariates that affect $SO_2$ concentration, such as the wind direction and distance to the major road which we would have included in our analysis but did not because of non-availability of data on these covariates.

# Chapter 5

# Spatio-temporal analysis of $SO_2$ data in Scotland

Chapters 3 and 4 dealt with time series and spatial modelling of the data separately. We now turn to spatio-temporal analysis which involves simultaneous analysis of $SO_2$ data both spatially and temporally using the generalized additive model procedure. Many authors have considered generalized additive models in modelling of air pollution data (Samet et al., 1999 & 2002; Giannitrapani et al., 2007; Holland et al., 2000; Bowman et al., 2009; Terzi and Cengizet, 2009; Wood and Augustin, 2002).

The performance of the various models will be assessed by applying a number of quantitative approaches and standard criteria such as $R^2(adj)$, GCV, AIC and deviance explained. We shall examine errors of estimation, predict and interpolate the $SO_2$ levels across Scotland, as well as visualize the smooth function of each independent variables for any visible pattern. We shall also justify GAM as a useful tool for interpolation and prediction of $SO_2$ levels. Lastly, we shall find significant effects of spatial locations, and a long term-trend effect of year from 1996-2007. The estimated effects of year and month (seasonal variable) are clearly non-linear.

The structure of this chapter is as follows. Section 5.1 reviews the background to the theory of generalized additive models, focussing more on the method of fitting, modelling with basis functions, criteria for basis selection and dimension, and basic description of bases considered in this study. Section 5.2 describes the GAM package in R and measures of model fit. Section 5.3 presents the analysis results, sensitivity analysis and model validation. Section 5.4 gives the conclusions and further considerations.

## 5.1 Introduction

This section describes the use of the GAM procedure for fitting generalized additive models (Hastie and Tibshirani, 1990). The GAM method is a robust data analysis procedure. It involves a combination of non-parametric regression, smoothing techniques and generalized linear models. Detailed discussion of the GAM procedure described in this section can be found in Xiang (2001) and Hastie and Tibshirani (1986).

The GAM approach investigates the structural relationship between a response and a set of independent variables. The technique can be applied on Gaussian data as well as data from Binomial, Poisson, and other non-Gaussian distributions. GAM can be used as a predictive model or as an exploratory method to suggest possible transformations of the data or in a parametric model such as a GLM. It is a Generalized Linear Model where the linear predictor depends linearly on unknown smooth functions.

Let Y be a dependent variable and $X_1, \ldots, X_p$ be a set of independent variables. The linear regression model is usually based on the assumption that $E(Y)$ is linear in form, i.e.

$$E(Y) = f(X_1, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p. \tag{5.1}$$

If sample values for $Y$ and $X$ are given, estimates of $\beta_0, \beta_1, \ldots, \beta_p$ can be obtained by either the least squares or maximum likelihood methods. The additive model is a generalization of the linear model which models $E(Y)$ as

$$E(Y) = f(X_1, \ldots, X_p) = S_0 + S_1(X_1) + \ldots + S_p(X_p), \tag{5.2}$$

where $S_i(X), i = 1, \ldots, p$ are smooth functions for the explanatory variables, and a non-parametric procedure can be used to estimate these functions. A smoother is a non-parametric tool for summarizing the trend of a response measurement $Y$ as a function of one or more predictor measurements $X_1, \ldots, X_p$, but which does not rely on any assumption about the form of dependence of $Y$ on $X_1, \ldots, X_p$ (Hastie and Tibshirani, 1990). The generalized linear model is an extension of the linear model with the inclusion of a link function between a smooth functions $f(X_1, \ldots, X_p)$ and the response variable $E(Y)$.

Additive models are more flexible than linear models, and expose the true structural form and shape of the response variable without making any parametric assumptions. Generalized additive models include a random part and additive

part as well as a link function relating these two components. The response $E(Y)$ is usually from the exponential family of distributions, which includes the Gaussian, Poisson, Binomial and Gamma distributions. The probability density function is of the form

$$f_Y(y, \theta, \varphi) = exp\left\{\frac{y\theta - b(\theta)}{a(\varphi)} + c(y, \varphi)\right\}, \tag{5.3}$$

where $\theta$ is a location parameter and $\varphi$ represents a scale parameter, and a, b and c are arbitrary functions from the random part of the model. The additive component is represented by model of the form

$$\eta = S_0 + \sum_{i=1}^{p} S_i(X_i), \tag{5.4}$$

where $S_0(.), \ldots, S_p(.)$ are smooth functions, and $S_0(.) = E(\eta)$. Also, the relationship between the mean $\mu$ of the response variable and $\eta$ can be represented by a link function $g(\mu) = \eta$ (Hastie and Tibshirani, 1986; Faraway, 2006; Xiang, 2001).

### 5.1.1 Generalized Additive Model procedures

The two general methods used to fit generalized additive models are backfitting and local scoring. Hastie and Tibshirani (1990) discussed estimation of the smoothing terms $S_0, S_1(.) \ldots S_p(.)$ in the additive model. The backfitting algorithm is a general algorithm that can fit an additive model using various smoothing functions such as smoother.
To fit an additive model of the form

$$y = S_0 + \sum_{j=1}^{p} S_j(X_j), \tag{5.5}$$

using backfitting, at iteration $m = 0$, the initial value of the smoothers $S_j$ are taken as

$$S_0 = \bar{y}, S_1^0 = S_2^0 = \ldots = S_p^0 = 0, \tag{5.6}$$

then for $j = 1, \ldots p$, calculate a vector of partial residuals

$$r_j = y - \hat{S}_0 - \sum_{k \neq j}^{p} \hat{S}_k, \tag{5.7}$$

i.e. the residuals obtained by subtracting from the response all the estimated terms in the model except the $j$th smoother, which is then estimated by smoothing $r_j$ with respect to $X_j$. Doing this for $j = 1, \ldots, p$ produces a new set of estimated smoothers $\hat{S}_j$. Any unknown smoothing parameters are estimated separately in an additional stage of the algorithm, in a scoring step. The two stages alternate until the model deviance converges. Any nonparametric smoothing method can be used to obtain $\hat{S}_j$ from the residuals $r_j$ (Wood, 2006).
The GAM procedure in R library *mgcv* is estimated by penalized likelihood maximization where the penalised log likelihood is of the form

$$l(\beta) - \frac{1}{2}\sum_{j=1}^{p} \lambda_j \beta^T S_j \beta, \tag{5.8}$$

where $S_j$ is a matrix with entries of zeroes except for the coefficients of $\beta$ corresponding to the $j^{th}$ smoothing spline, Wood (2006) shows that maximizing this is equivalent to minimizing the penalised weighted least squares function (5.11) below and this is usually achieved by penalised iteratively re-weighted least squares (P-IRLS). Let the current linear predictor estimate be $\eta^{[k]}$ and the corresponding estimated mean response vector be $\mu^{[k]}$. Then weights are computed as

$$w_i \propto \frac{1}{V(\mu_i^{[k]})(g'(\mu_i^{[k]})^2}, \tag{5.9}$$

$$z_i = (g'(\mu_i^{[k]})(y_i - (\mu_i^{[k]}) + X_i\beta^{[k]}, \tag{5.10}$$

in which $var(y_i) = V(\mu_i)\varphi$, $g(.)$ is the form of link function and $X_i$ is now the $i$-th row of the matrix of predictor variables. The next step minimizes

$$||\sqrt{W}(z - X\beta||^2 + \sum_{j}^{p} \lambda_j \beta^T S_j \beta, \tag{5.11}$$

with respect to $\beta$ to obtain $\beta^{[k+1]}$, where $z = (z_1, \ldots, z_n)$, and hence

$$\eta^{[k+1]} = X\beta^{[k+1]}. \tag{5.12}$$

$W$ is a diagonal matrix such that $W_{ii} = w_i$. These two steps are repeated until convergence. This is a weighted version of the backfitting algorithm. A generalized cross validation (GCV) score can then be obtained from the final linear model in the P-IRLS iterations (Wood, 2006). Each smoother $s_j$ involves a smoothing parameter to be chosen. The GCV principle is applied in several non-parametric

regressions as the smoothing parameter selection criteria. GCV is a computationally efficient way to approximate leave-one-out cross-validation to minimise the average prediction error (average squared residual).
The above procedure assumes that the smoothing parameters $\lambda_j$, $j = 1, \ldots, p$, are known or fixed. To estimate the $\lambda_j$ also, the estimation steps above form one stage in a two stage procedure. The other stage chooses the $\lambda_j$ to minimize the chosen criterion, e.g. GCV, given the current values of the coefficients $\beta$. These two stages are repeated iteratively until convergence. For multiple smoothing parameters, an approximation is made to the GCV or UBRE score (see Section 5.2) to enable direct minimization of the score to obtain the next estimate of the smoothing parameters. Further details about this procedure can be obtained from (Wood and Augustin, 2002; Wood, 2006; Green and Silverman, 1994; Hastie and Tibshirani, 1986 & 1990).

### 5.1.2 Basis dimension and basis selection

GAM modelling involves modelling with basis functions, so we give a little introduction about basis functions. Covariate terms like $f(x)$, $f(y)$ or $g(x; y)$, without knowing the true forms of the functions, can be included in a model. Let the best representation of a function containing a smooth function of one covariate be represented by

$$y_i = f(x_i) + \varepsilon_i \tag{5.13}$$

in which $y_i$ is a response variable, $x_i$ is a covariate, $f$ is a smooth function and $\varepsilon_i$ are independent and identically distributed random variables ($N(0, \sigma^2)$. Let $f(x)$ be represented by a sum

$$f(x) = \sum_{j=1}^{k} b_j(x)\beta_j, \tag{5.14}$$

in which the $\beta_j$ are $k$ unknown parameters and $k$ is a basis dimension. So $f(x)$ is a linear combination of basis functions $b_j(x)$, and we now estimate $\beta_j$ to obtain $f$. The model can be estimated by minimising

$$\sum_{i}^{n}(f(x_i) - y_i)^2 = \sum_{i}^{n}\left(\sum_{j=1}^{k} b_j(x_i)\beta_j - y_i\right)^2 \tag{5.15}$$

(Wood, (2006)). The choice of basis dimension when using penalized regression smoothers may be expected to have a considerable effect on modelling results.

Penalized regression smoothers have computational efficiency by using a basis of relatively small size, $k$. When formulating the GAM model in R using the $s$ (isotropic smoothing) or $te$ (tensor product in which a smooth of several covariates can be constructed, especially when covariate functions are measured on different scales) (see Section 5.1.5) terms in a model formula, the basis dimension $k$ has to be chosen. The size of basis dimension sets an upper limit on the flexibility of a smoother. The three bases we consider in our GAM models are the thin-plate regression splines, cubic regression splines and P-splines.

### 5.1.3 Cubic spline basis

A cubic smoothing spline fits a smooth curve to observations using a spline function. Let $(x_i, Y_i), i = 1, \ldots, n$ be a sequence of observations using the relation $E(Y_i) = f(x_i)$. The smoothing spline estimate $\hat{f}$ of the function $f$ is defined to be a value which minimizes the twice differentiable function

$$\sum_{i=1}^{n}(Y_i - f(x_i))^2 + \lambda \int f''(x)^2\, dx, \tag{5.16}$$

where $\lambda \geq 0$ is a smoothing parameter and $\lambda \int f''(x)^2$ is a roughness penalty. When $f$ is rough, the penalty is large, but when $f$ is smooth, the penalty is small.

The choice of roughness penalty here gives the solution of a particular form, i.e. $\hat{f}$ is a cubic spline, which means that $\hat{f}$ is a piecewise cubic polynomial and has the property that $\hat{f}$, $\hat{f}'$ and $\hat{f}''$ are continuous. The joins are known as knots. Other choices of roughness penalties of higher order are also possible.

The estimation is reduced to the parametric problem of estimating the coefficients of the polynomials which can be efficiently estimated numerically. Cubic splines are the smoothest interpolators and are more or less ideal smoothers, except for the substantial problem of having many more free parameters relative to the data to be smoothed.

Cubic regression splines have a cubic spline basis. The computation is cheap, but this approach only smooths with respect to one variable at a time and does not have optimal properties. The basis dimension controls the degree of smoothing (Hastie and Tibshirani, 1990; Ruppert et al., 2003, Wood, 2006; Faraway, 2006).

### 5.1.4 Thin plate regression splines

Thin plate regression splines are low rank isotropic smoothers which can be used on any number of covariates. Isotropic means that smoothing results are not affected by the rotation of the covariate's coordinates, while a low rank means there are fewer coefficients than the observed data. They are a reduced rank version of thin plate splines and use the thin plate spline roughness penalty. Penalized thin plate regression splines give the best mean square error performance but are usually slower to set up than the other basis functions.
In summary, thin plate regression splines can smooth with respect to any number of covariates, are invariant to rotation of covariate axes, and can also select the order of penalty function with some optimality properties. But they are computationally costly for large datasets and are not invariant to covariate rescaling (Wood, 2003 & 2006).

### 5.1.5 P-splines

These are another way to represent cubic splines. A P-spline is a smoother in which the number of basis functions is less than the number of observations, but the basis terms are penalized and so it is usually referred to as a penalized spline or P-spline. They combine a B-spline basis (spline function that has minimal support with respect to a given degree and smoothness), with a discrete penalty on the basis coefficients, and different combinations of penalty and basis order are possible. They are also low-rank smoothers and perform well in tensor products (Eilers and Marx, 1996; Wood, 2006; Wahba, 1990).

### 5.1.6 Tensor product

Tensor product smooths avoid smoothing equally in all directions (isotropic smoothing), which is not always appropriate. Instead a tensor product approach is used to build up smooths of several variables, based on univariate smooths of each variable separately (Wood (2006) gives details). For three variables $x$, $z$ and $v$ for example, using a tensor product basis, if we have $f(x) = \sum_{i=1}^{I} \alpha_i a_i(x)$, $f(z) = \sum_{j=1}^{L} \psi_j d_j(z)$, and $f(v) = \sum_{k=1}^{K} \beta_k b_k(v)$, where $\alpha_i$, $\psi_j$ and $\beta_k$ are coefficients and $a_i(x)$, $d_j(z)$ and $b_k(v)$ are known basis functions, e.g. B-splines, and these have been fitted to smooth in a given direction only, then

$$f(x, z, v) = \sum_{i=1}^{I} \sum_{j=1}^{L} \sum_{k=1}^{K} \beta_{ijk} b_k(v) d_j(z) a_i(x), \tag{5.17}$$

where $\beta_{ijk}$ are also coefficients. A measure of smoothness in the multidimensional case can be obtained from the measures of smoothness achieved in each dimension separately (Wood, 2006).

## 5.2 GAM and R software

There are at least three different ways of fitting generalized additive models in R, namely *gam*, *mgcv* and *gss* packages. The *gam* package allows more choice in the smoothers and a backfitting algorithm is used, the *mgcv* package utilizes automatic choices in the degree of smoothing as well as wider functionality, and employs a penalized regression spline approach, while the *gss* package also makes use of spline-based methods. For this study we will employ the *mgcv* package of Wood (2000), similar to what was also adopted in Bowman et al. (2009).
The smoothing parameters can be selected by GCV (Generalized Cross Validation) described above, AIC (Aikake's Information Criterion) or UBRE (Un-Biased Risk Estimator) or by using regression splines with fixed degrees of freedom (Wood, 2000; Wahba and Gu, 1991). Model fit is assessed below by $R^2$-adjusted, deviance explained, GCV score, AIC, or -log likelihood. In the model fitting output, e.g. in Table 5.1, edf is the effective degrees of freedom, which depends on the number of parameters and smoothness constraints. Scale est. is an estimate of the scale parameter for the response variable ($\sigma^2$ in the normal case), and deviance explained is the proportion of the null deviance explained by the model.

### AIC, Deviance explained and UBRE

AIC is found as $2p - 2log(L)$, in term of maximized likelihood $L$ for the fitted model and where $p$ is the number of parameters in the model. For a normal response variable, this can be simplified to $AIC = 2p + n(log(RSS))$, where $n$ is the number of observations and $RSS$ is the residual sum of squares, and there are also various other equivalent expressions of AIC for model comparison. Low values indicate a better model.
Deviance explained takes a value similar to $RSS$. Up to a constant it is equal to $-2log(L)$, or equal to $2[logL(\hat{\beta}) - logL(\hat{\beta}_{max})]$ where $logL(\hat{\beta}_{max})$ is the maximum value possible of the log likelihood, found by using one parameter for each data point, and $logL(\hat{\beta})$ is the log likelihood for the current model. UBRE is equivalent to Mallow's $C_p$, from multiple regression analysis, an estimate of expected mean square error (Wood, 2006).

## 5.3 Analysis of $SO_2$ data

We model $S0_2$ concentration as a function of several explanatory variables using non-parametric smoothers. The dataframe consists of $SO_2$ concentration, with spatial location given by Northing and Easting. We also include monthly and yearly factors to measure the seasonal effects and long-term trend. Therefore, we now interpret the effects of both space and time in the analysis of the data. The log average monthly concentration for each station is computed for the 12 year range 1996-2007, after imputing the missing observations using the EM imputation technique, giving 4384 observations in total.

Firstly, we fitted an additive model using an identity function as the link function and a Gaussian model for the response, and utilized the thin-plate regression splines of the *mgcv* package because they are the default smooth for $s$ terms and are optimal smoothers for any given basis dimension (Wood, 2003). The default basis dimension of $k = 10$ is also used. The default method for smoothing parameter selection is GCV. We check the sensitivity of the analysis in the later part of the modelling using different bases for the univariate smooth while maintaining a default for the 2-dimensional spatial location smooth, using different basis dimensions as well as different methods to choose the smoothing parameters (both GCV and REML (Restricted Maximum Likelihood) are considered).

Lastly, model validation was also investigated by randomly eliminating 10 stations from our observations and fitting the models to the remaining data before predicting the left out data.

Let the basic model without the bivariate spatial location be given by

$$y = \mu + s(Year) + s(Month) + \varepsilon. \tag{5.18}$$

where $y = log(SO_2{+}0.5)$, after imputing the missing observations by EM, s(Year), s(Month) are the terms to represent the univariate smoothings for the year and month, $\mu$ is the estimated mean and $\varepsilon$ is the residual term.

The results and diagnostics plots for the preliminary analysis are shown in Table 5.1 and Figure 5.1. The QQ plot is curved, the histogram of residuals is skew to the left, and the error variance does not appear to be constant. This indicates that the $SO_2$ data residuals are not ideal but after a more complex model is fitted may be better.

In Table 5.1, only the smooth factor for year is significant with effective degrees of freedom of 8.9, although the term for month is not far from significance. The $R^2$ is 0.23 which is very low, which implies that the explanatory smooth functions

of month and year only explain about 23% of variation in the data. The GCV score is very high (0.809) whch is another indication that the model is too simple to account for the fluctuation in levels of $SO_2$. The parametric coefficient for the intercept is 2.18009 which is a measure of the mean $\mu$.

Figures 5.2 and 5.3 show the graphical display of the estimate of the smooth factor for the year and month as long term (year) and seasonal effects. The constant found beside each covariate in the y-axis label of the plots is the effective degrees of freedom estimated by the automatic method (for example, the univariate smoothing of year utilized 8.91 degrees of freedom in Figure 5.2 (see Table 5.1)). In Figure 5.2 for instance, there is a general fall in level of $log(SO_2)$ as indicated by the smooth function of year, with 2007 having the lowest concentration for the years we consider, possibly as a result of various measures taken by Goverment to reduce the concentration levels across the UK.

The monthly factor is not quite significant in this model, but from the plot in Figure 5.3 there is a slight peak between May to July (summer peak). There is a general relatively low variation in levels within year. The summer peak we observe here is in accordance with our earlier results in Chapters 1 and 3 in which the time series for most stations show evidence of higher mean level in the summer months (Figures 1.9-1.11, 1.15-1.17, and 3.1-3.3).

Table 5.1: Simple additive model without spatial interaction

```
Model 1
> summary(gam1)
Family: gaussian
Link function: identity
Formula:
log(Mean + 0.5) ~ s(Month) + s(Year)
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.18009    0.01558   140.0   <2e-16 ***
Approximate significance of smooth terms:
           edf Ref.df       F p-value
s(Month) 2.219  2.763   2.565  0.0579 .
s(Year)  8.909  8.997 110.170  <2e-16 ***
R-sq.(adj) =   0.23   Deviance explained = 23.3%
GCV score = 0.80949  Scale est. = 0.80654   n = 3324
```

Figure 5.1: Diagnostic check for residual plots of simple model

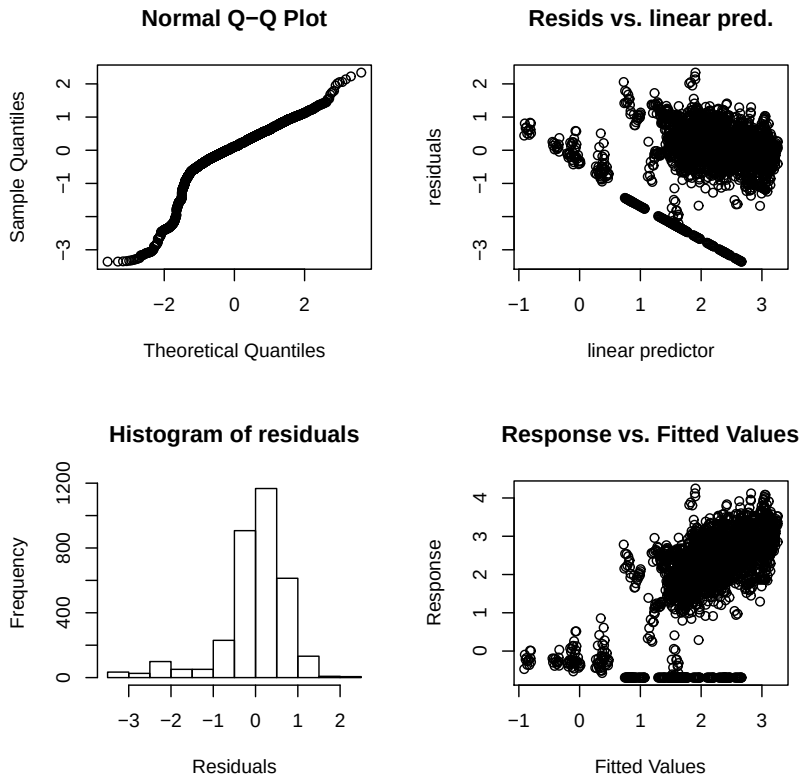


Figure 5.2: Estimate of long trend effect for the year, in the simple model; 8.91 in the y-axis label is the edf for year from the model fitting. The dotted lines show 95% confidence intervals

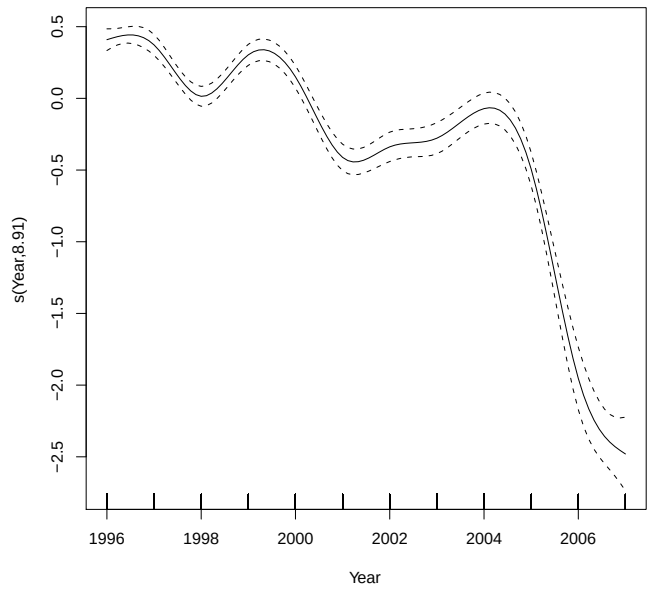
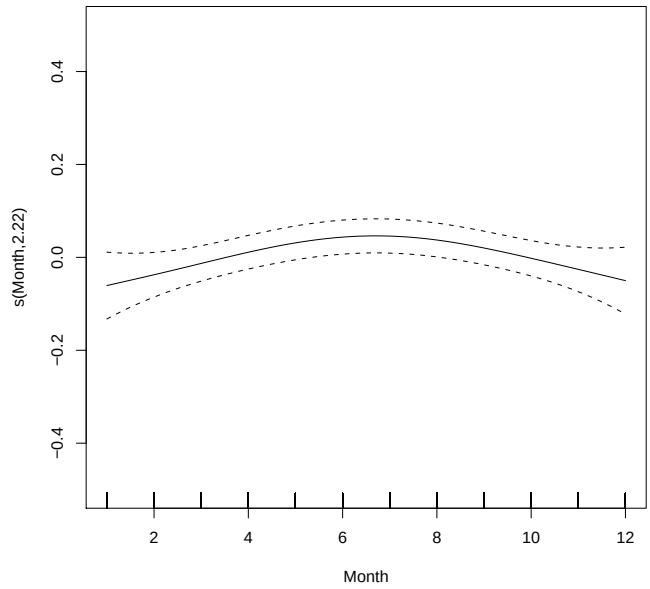



Figure 5.3: Estimate of the seasonal effect for the month, in the simple model; 2.22 in the y-axis label is the edf for month from the model fitting. The dotted lines show 95% confidence intervals



We now consider a more complex model by incorporating the bivariate smooth spatial factor s(Easting, Northing) into model 1, $log(SO_2) = \mu + s(Year) + s(Month) + s(Easting, Northing) + \varepsilon$. The essence of this additional variable is to investigate if spatial location has any effect on the variation in mean level of $SO_2$. The result is shown in Table 5.2 and Figures 5.4-5.6. We observe a substantial change from Model 1. This model performs better than model 1, as $R^2$ has increased from 0.23 to 0.382 and deviance explained from 23% to 39%. The smooth factors for the spatial location and year are both significant in this model with effective degrees of freedom of about 28 and 9 respectively. The smooth factor for month is not significant in this model.

Figure 5.4 is the bivariate spatial location for the monitoring stations and we see that most of the stations are concentrated in Central Scotland with a very high concentration of $SO_2$ in this region. We observe that the bivariate smoothing of Northing and Easting is significant (that is there is an interaction between the Northing and the Easting as displayed on the map). The yearly effect displayed in Figure 5.5 also indicates a general downward trend in $SO_2$ levels with year, very similar to the pattern for model 1.

Figure 5.6 shows the monthly effect, which is a measure of a seasonal effect. This now has a different pattern from model 1 in Figure 5.3 and it shows that there is no monthly seasonal pattern.

Figure 5.7 shows the residual analysis for the more complex model. This looks slightly better than for the simple model but still not ideal.

Table 5.2: Additive model with spatial interaction for location

```
Model 2
> summary(gam2)
Family: gaussian
Link function: identity
Formula:
log(Mean + 0.5)~s(Easting, Northing) + s(Year2) + s(Month)
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.18009    0.01395   156.3   <2e-16 ***
Approximate significance of smooth terms:
                       edf Ref.df       F p-value
s(Easting,Northing) 28.444 28.968  29.458  <2e-16 ***
s(Year)              8.968  9.000 142.551  <2e-16 ***
s(Month)             3.377  4.186   1.778   0.127
R-sq.(adj) =  0.382   Deviance explained =   39%
GCV score = 0.65531  Scale est. = 0.64707   n = 3324
```

Figure 5.4: Estimate of the bivariate spatial effect of Easting and Northing

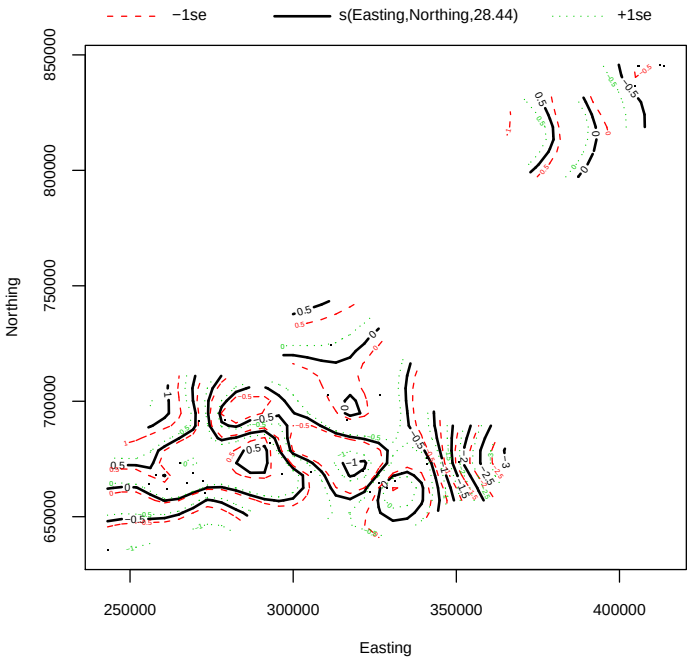


Figure 5.5: Estimate of year trend effect for the model including location; 8.91 in the y-axis label is the edf for year from the model fitting. The dotted lines show 95% confidence intervals

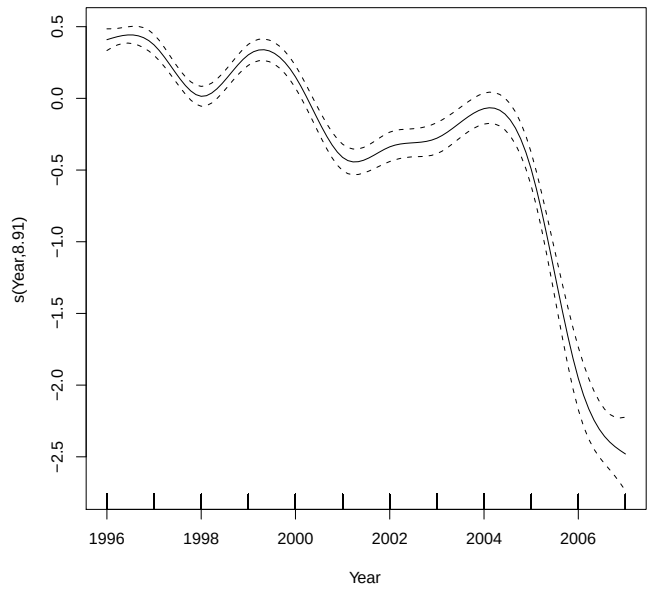



Figure 5.6: Estimate of the seasonal effect (month) for the model including location; 3.38 in the y-axis label is the edf for month from the model fitting. The dotted lines show 95% confidence intervals

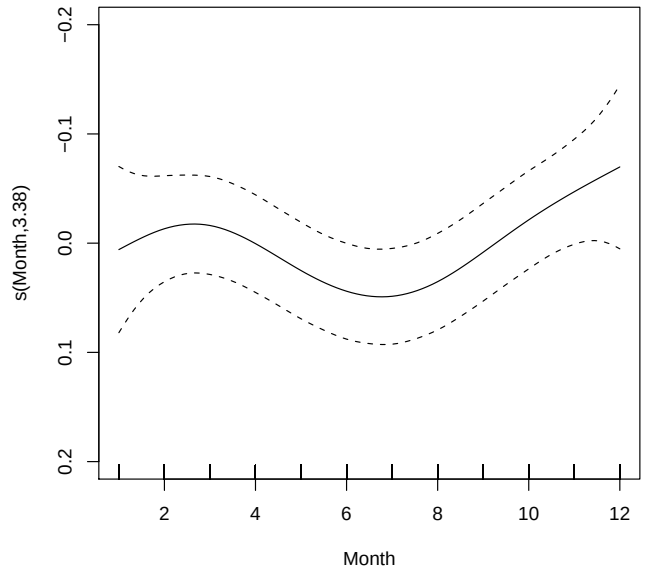
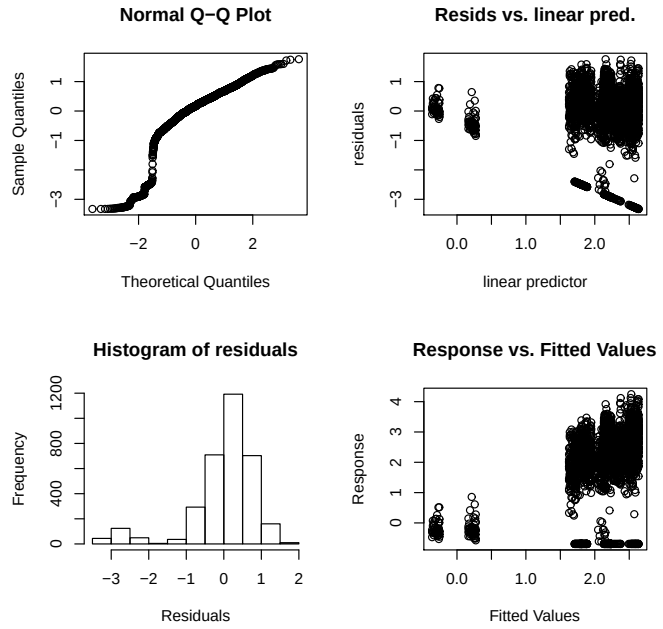


Figure 5.7: Diagnostic check for residual plots for model including location




We next split our dataset into 2 different year groups, namely 1996-2000, and 2001-2005. The aim of this is to investigate any temporal interaction, that is if the variation in long term pattern of $SO_2$ data has any effect on the models. We do this as a result of the different pattern in the long-term trend (Figures 5.2 and 5.5), in which there is a relative decrease and increase (gentle fluctuation) in trend between 1996 and 2000. Theree is also a gentle decrease in level between 2001 and 2005. We decided to divide our data into two equal groups, those between 1996 and 2000, and 2001 and 2005. We excluded years 2006 and 2007 as $SO_2$ levels fall very rapidly beween this period.

We now formulate a similar model to model 2. The results are presented in Tables 5.3 and 5.4 and Figures 5.8 and 5.9. We select a fixed basis dimension of $k = 6$ and 5 rather than 10 respectively for the two groups because of the fewer observations. Similarly to the two previous models, the smooth factor for month is not significant for 1996-2000 but it is now significant for the later years. The 2001-2005 dataset has higher $R^2$ (0.646), better than the previous models, though this may be because of the fewer observations ($n = 1056$), and so low degrees of freedom 27.2. In both cases, the effects of location and year are highly significant.

In Figure 5.8 the top and bottom panels represent the bivariate spatial location for 1996-2000 and 2001-2005 respectively, while in Figure 5.9 the two top panel

are smooth functions for year and month for the 1996-2000 datasets respectively, while the two bottom panels are the smooth functions for year and month for 2001-2005 datasets respectively. The more tightly packed contours in the bivariate spatial plot for 1996-2000 indicate levels of $SO_2$ varying more rapidly between Central Scotland stations than in 2001-2005.

The North-Eastern region also recorded a high mean $SO_2$ level within this period whereas the 2000-2005 has no recorded observation for this region, as we have fewer recording stations for $SO_2$ levels in the later years. The smooth function for year falls more rapidly between 1996-2000 than 2001-2005. We see that the level decreases between 1996 and 1998 before it rises again in 1999 and then drops to a very small level in 2000. For years 2001-2005, there is a general fall in level between 2001 and 2002 and it slightly increases in 2003 before dropping again to a very low level in 2004. The within year (monthly pattern) variation is similar for both groups, and summer months still have higher concentration levels as expected.

We consider a similar model to model 2 for Central Scotland stations only to see if there would be any effect (improvement) of excluding the remote station. The result for Central Scotland is shown in Table 5.5 and Figures 5.10 and 5.11. The two smooth functions for spatial location and year are significant. The deviance explained by this model is 39.6% and $R^2$ is 0.387 which is similar to model 2. The parametric coefficient for the intercept is a little lower (2.1029) compared to the model 2 value of 2.18, possibly due to reduction in mean $SO_2$ levels as a result of elimination of remote observations. In Figure 5.10, high mean concentration level is also visible for the $SO_2$ levels in Central Scotland. In Figure 5.11, the smooth function for year also indicates a general decrease in level, while the smooth function for month still shows high mean levels in summer months similar to models 1, 3a and 3b in Figures 5.3 and 5.9 respectively.

Table 5.3: Simple additive model with spatial interaction for the 1996-2000 dataset

```
Model 3a
 > summary(gam3a)
Family: gaussian
Link function: identity
Formula:
log(Mean[1:2172] + 0.5) ~ s(Easting[1:2172], Northing[1:2172]) +
s(Year2[1:2172], k = 6) + s(Month[1:2172], k = 6),data=mydata)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.4274     0.0166   146.2   <2e-16 ***
Approximate significance of smooth terms:
                                      edf Ref.df      F p-value
s(Easting[1:2172],Northing[1:2172]) 28.576 28.980 24.890  <2e-16 ***
s(Year2[1:2172])                     4.986  5.000 28.756  <2e-16 ***
s(Month[1:2172])                     2.377  2.927  1.714   0.163

R-sq.(adj) =  0.273   Deviance explained = 28.5%
GCV score = 0.60913  Scale est. = 0.59877   n = 2172
```

Table 5.4: Simple additive model with spatial interaction for the 2001-2005 dataset

```
Model 3b
> summary(gam3b)
Family: gaussian
Link function: identity
Formula:
log(Mean[2173:3228] + 0.5) ~ s(Easting[2173:3228], Northing[2173:3228],
k=20) + s(Year2[2173:3228], k = 5)+s(Month[2173:3228],k =5,data=mydata)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.88633    0.01715   110.0   <2e-16 ***
Approximate significance of smooth terms:
                             edf    Ref.df F p-value
s(E[2173:3228],N[2173:3228]) 18.960 18.999 97.342 < 2e-16 ***
s(Year2[2173:3228])          3.948  3.998 43.477 < 2e-16 ***
s(Month[2173:3228])          2.328  2.824  4.645 0.00381 **

R-sq.(adj) =  0.646   Deviance explained = 65.4%
GCV score = 0.31855  Scale est. = 0.31064   n = 1056
```

Figure 5.8: The bivariate spatial plot for simple additive model with spatial interaction for the separate 1996-2000 and 2001-2005 datasets. The upper panel corresponds to 1996-2000 while the lower panel is 2001-2005

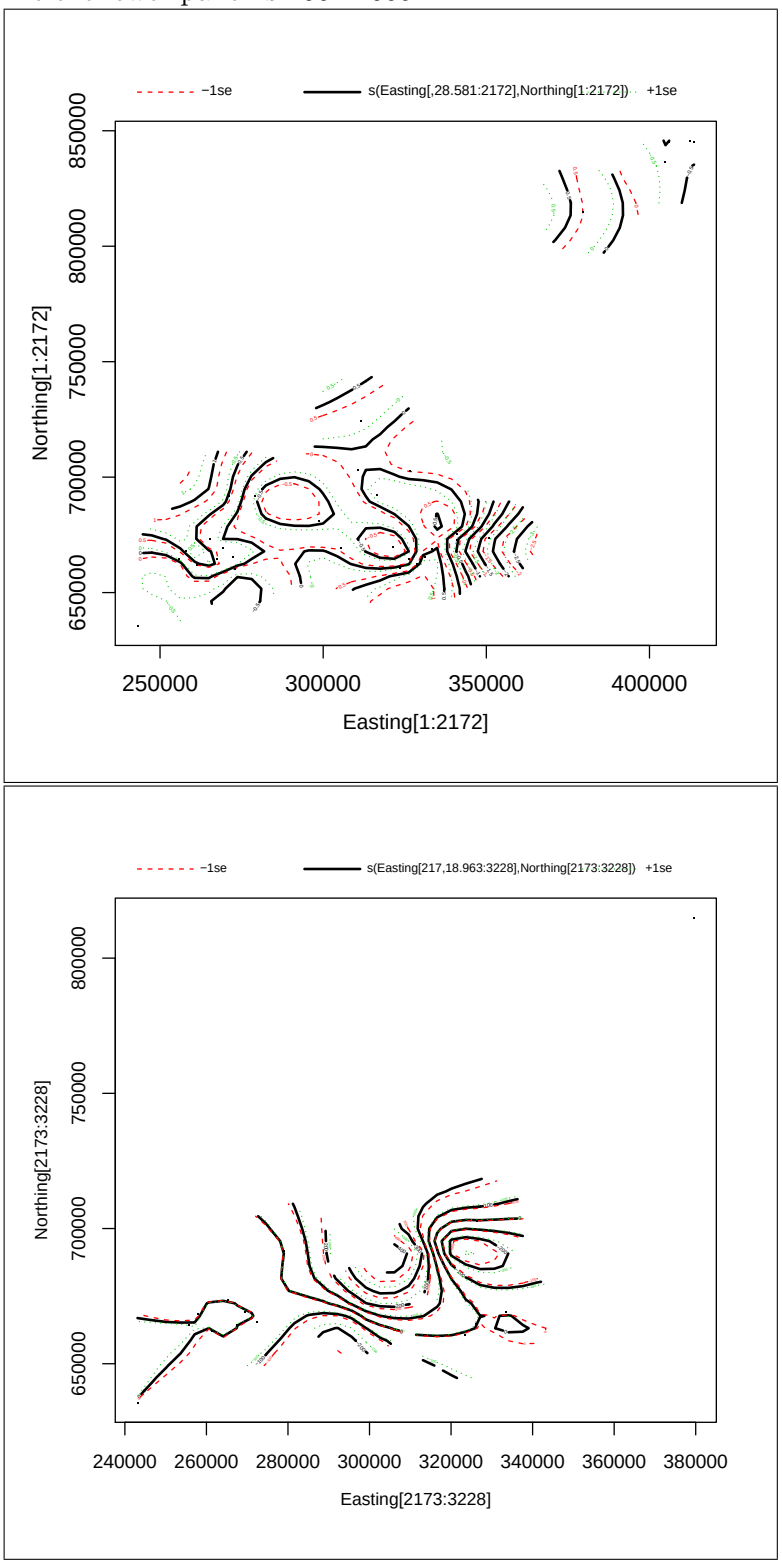


Figure 5.9: Simple additive model with spatial interaction for the separate 1996-2000 and 2001-2005 datasets. The two top panel are smooth functions for year and month for the 1996-2000 datasets respectively, while the two bottom panels are the smooth functions for year and month for 2001-2005 datasets respectively

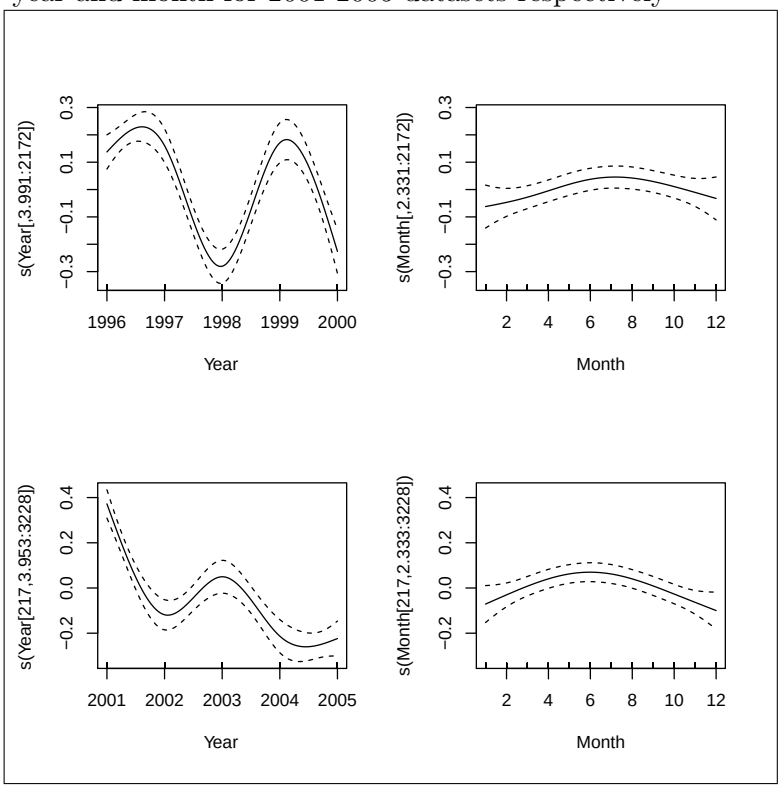



Table 5.5: Simple additive model with spatial interaction for the Central Scotland stations only

```
Model 3c
 summary(gam3c)
log(Mean + 0.5) ~ s(Easting, Northing) + s(Year) +
 s(Month), data=mydat11
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.10290    0.01689   124.5   <2e-16 ***
Approximate significance of smooth terms:
                       edf Ref.df      F p-value
s(Easting,Northing) 24.920 26.267 17.348  <2e-16 ***
s(Year)              8.925  8.998 86.421  <2e-16 ***
s(Month)             2.225  2.771  2.676  0.0502 .
R-sq.(adj) =  0.387   Deviance explained = 39.6%
GCV score = 0.65849  Scale est. = 0.64773   n = 2270
```

Figure 5.10: The bivariate spatial plot for simple additive model with spatial interaction for Central Scotland stations

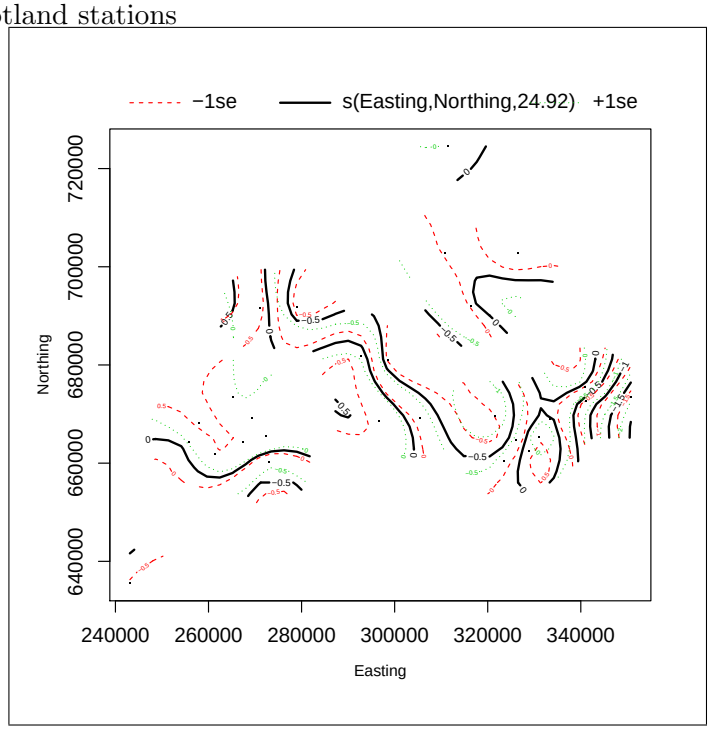
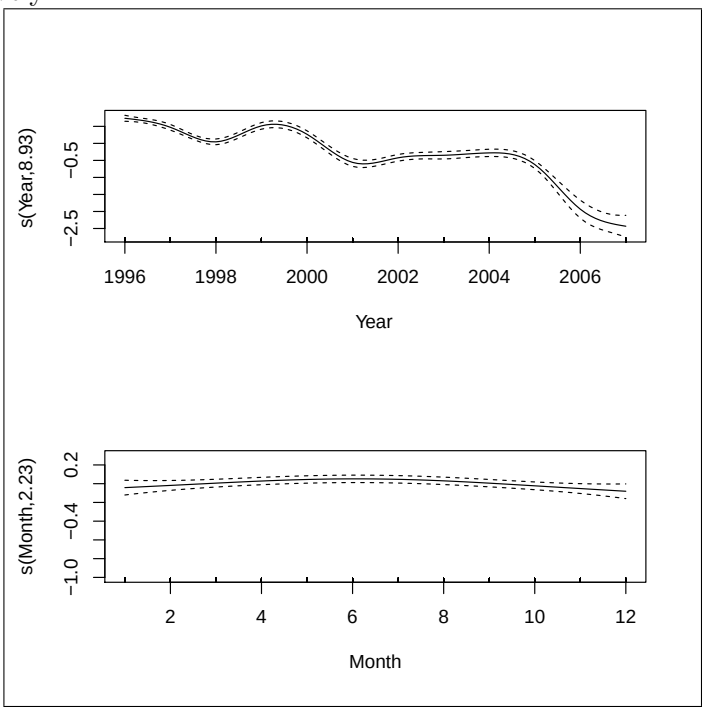



Figure 5.11: Simple additive model with spatial interaction for Central Scotland stations. The top and bottom panels represent the smooth functions for the year and the month respectively



## 5.4 Sensitivity analysis

Here, we consider the sensitivity of various models to the choice of basis and basis dimension as well as the estimation method for the degree of smoothing. We refitted model 2, for all years and stations to accommodate other bases (cubic regression and p-spline). We now have $log(SO_2) = s(Month, bs = "cr") + s(Year, bs = "cr") + s(Easting, Northing, bs = c("tp"))$ and $log(SO_2) = s(Month, bs = "ps") + s(Year, bs = "ps") + s(Easting, Northing, bs = c("tp"))$ for the univariate cases and maintaining the thin plate regression for the bivariate smoother using the whole Scotland dataset.

We also assign fixed values of basis dimension for the univariate and bivariate smoothers respectively to check sensitivity to choice of basis dimension, and lastly a REML method was used on Model 2 again to check for any significant effect of changing smoothing parameter criterion from GCV to REML. The results are shown in Tables 5.6-5.11. From Table 5.6, using cubic regression splines for the univariate smoothing in model 2, has now made all three (bivariate spatial location, year and monthly) factors to be significant, though the smooth function for month is the least significant factor. The model fit is otherwise similar. $R^2$ and deviance explained are slightly lower than for model 2 in Table 5.2, and are 0.379 and 38.6% respectively.

Table 5.7 is the result obtained by utilizing p-splines for the univariate smoothing terms, and we observe a similar pattern to Table 5.6 with all three variables still significant and both $R^2adj$ and deviance explained are also lower than for model 2 (0.372 and 37.9% respectively). Generally, the various basis combinations give similar results, thus the analysis are not too sensitive to the type of basis used in the model. Also, we refitted Model 2 by setting a fixed value of $k = 6$ and 12 and $k = 12$ and 20 for univariate and bivariate smoothers respectively in which we have $log(S0_2) = s(Month, k = 6) + s(Year, k = 6) + s(Easting, Northing, k = 12)$ and $log(S0_2) = s(Month, k = 12) + s(Year, k = 12) + s(Easting, Northing, k = 20)$.

From Table 5.8, the result indicates a large reduction in both $R^2adj$ and deviance explained from 0.382 and 39% in model 2 to 0.297 and 30% respectively. In Table 5.9, both $R^2adj$ and deviance explained are slightly lower than for model 2 with 0.374 and 38% respectively but better than in Table 5.8, which suggests that basis dimension has a considerable effect on the results. In Table 5.10, in which the estimation method is now REML rather than GCV, both $R^2adj$ and deviance explained are slightly higher than for model 2, with 0.389 and 39.6% respectively. Figure 5.12 shows a similar pattern for the smooth function of month with high

mean level in summer months irrespective of the type of basis or dimension of basis used in the model and method of estimation for models 4a-4e, which are quite different from that of model 2 in Figure 5.6. The smooth function for year has similar patterns for the models with cubic regression, p-spline, basis dimension $k = 12$ and REML estimation method terms, and these are similar to model 2 in Figure 5.5, while that of basis dimension $k = 6$ is smoother in pattern from the other sensitivity models.

In summary, we see that our model is more dependent on the basis dimension than the type of basis used and method of estimation. Also, from Table 5.11, comparing all the models considered in this chapter, model 4e seems better than the rest. This is the model with REML estimation, with a minimum AIC (7992.27) and maximum log likelihood (-3955.167), and degrees of freedom 40.967, but this was able to explain just 39.6% of the variation in the response variable. Model 3b is much better in terms of $R^2$, deviance explained, and GCV but is only for the years 2001-2005.

Models gam1, gam2, gam4a, gam4b, gam4c, gam4d and gam4e correspond to all Scotland; and years 1996-2007, gam3a is all Scotland; and years 1996-2000, gam3b is all Scotland; and years 2001-2005 and gam3c is Central Scotland; and years 1996-2007.

Table 5.6: Sensitivity to the choice of basis, for the spatial regression, using cubic splines instead of thin plate regression for the univariate terms

```
 Model 4a
> summary(gam4a)
Family: gaussian
Link function: identity
Formula:
log(Mean + 0.5) ~ s(Month, bs = "cr") + s(Year, bs = "cr") +
s(Easting, Northing, bs = c("tp"),data=mydata)
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.18009    0.01399   155.8   <2e-16 ***
Approximate significance of smooth terms:
                       edf Ref.df       F p-value
s(Month)             2.353  2.925   3.155  0.0249 *
s(Year)              8.901  8.996 140.032  <2e-16 ***
s(Easting,Northing) 28.382 28.960  29.260  <2e-16 ***

R-sq.(adj) =  0.379   Deviance explained = 38.6%
GCV score = 0.65859  Scale est. = 0.65054   n = 3324
```

Table 5.7: Sensitivity to the choice of basis, using p-splines instead of thin plate splines for the univariate terms

```
Model 4b
> summary(gam4b)
Family: gaussian
Link function: identity
Formula:
log(Mean + 0.5) ~ s(Month, bs = "ps", k = 8) + s(Year, bs
="ps",k = 8) + s(Easting, Northing, bs =c("tp"),data=mydata)
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.18009    0.01406     155   <2e-16 ***
Approximate significance of smooth terms:
                       edf Ref.df       F p-value
s(Month)             2.270  2.745   3.151  0.0278 *
s(Year)              6.988  7.000 172.779  <2e-16 ***
s(Easting,Northing) 28.425 28.965  28.506  <2e-16 ***

R-sq.(adj) =  0.372   Deviance explained = 37.9%
GCV score = 0.66525  Scale est. = 0.65751   n = 3324
```

Table 5.8: Sensitivity to the choice of basis dimension, using $k = 6$ and 12 for univariate and bivariate smoothers respectively

```
Model 4c
> summary(gam4c)
Family: gaussian
Link function: identity
Formula:
log(Mean + 0.5) ~ s(Month, k = 6) + s(Year, k = 6)
 + s(Easting, Northing, k = 12),data=mydata)
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.18009    0.01489   146.4   <2e-16 ***
Approximate significance of smooth terms:
                       edf Ref.df       F p-value
s(Month)             2.271  2.798   2.916  0.0365 *
s(Year)              4.978  5.000 236.564  <2e-16 ***
s(Easting,Northing) 10.584 10.964  37.721  <2e-16 ***

R-sq.(adj) =  0.297   Deviance explained =   30%
GCV score = 0.74089  Scale est. = 0.7367    n = 3324
```

Table 5.9: Sensitivity to the choice of basis dimension, using $k = 12$ and 20 for univariate and bivariate smoothers respectively

```
Model 4d
 > summary(gam4d)
Family: gaussian
Link function: identity
Formula:
log(Mean + 0.5) ~ s(Month, k = 12) + s(Year, k = 12)
 + s(Easting, Northing, k = 20),data=mydata)
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.18009    0.01404   155.2   <2e-16 ***
Approximate significance of smooth terms:
                       edf Ref.df       F p-value
s(Month)             2.353  2.932   2.954  0.0324 *
s(Year)             10.942 10.999 135.803  <2e-16 ***
s(Easting,Northing) 18.814 18.996  36.551  <2e-16 ***

R-sq.(adj) =  0.374   Deviance explained =   38%
GCV score = 0.66229  Scale est. = 0.6557    n = 3324
```

Table 5.10: Sensitivity to the choice of smoothing parameter estimation method, using REML instead of GCV

```
> summary(gam4e)
Model 4e
Family: gaussian
Link function: identity
Formula:
log(Mean+0.5)=s(Month) + s(Year)+s(Easting, Northing),data=mydata)
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.18009    0.01388   157.1   <2e-16 ***
Approximate significance of smooth terms:
                       edf Ref.df       F p-value
s(Month)             2.653  3.298   3.053  0.0233 *
s(Year)              8.817  8.987 148.802  <2e-16 ***
s(Easting,Northing) 27.499 28.781  29.868  <2e-16 ***
R-sq.(adj) =  0.389   Deviance explained = 39.6%
REML score = 4066.5  Scale est. = 0.64017   n = 3324
```

Figure 5.12: Comparison of sensitivity to the choice of basis, basis dimension, and estimation method for univariate smoothing of month and year, $k = 6$ and 12 for univariate and bivariate smoothers respectively; cr=cubic regression and ps=p-spline

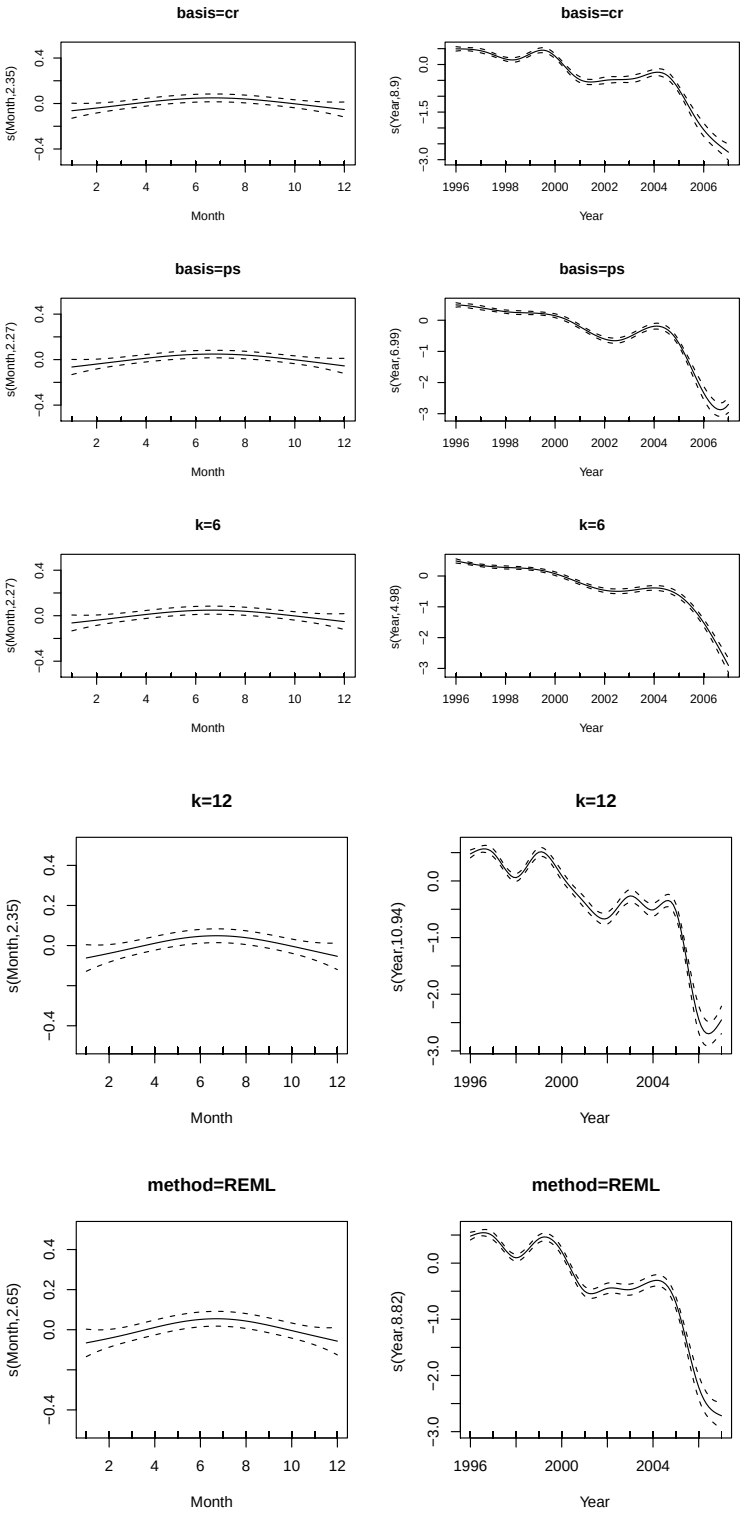


Table 5.11: Comparison of all the models considered based on their AIC, $R^2$, GCV, and deviance explained criteria; df is the sum of estimated degrees of freedom for each model; The "*" corresponds to models based on a subset of the whole data

| Model | df | logLik | AIC | R-sq | Dev. | GCV/ REML |
|---|---|---|---|---|---|---|
| gam1 | 13.1279 | -4353.146 | 8732.54 | 0.23 | 23.3% | 0.80949 |
| gam2 | 42.7885 | -3972.062 | 8029.70 | 0.382 | 39% | 0.65531 |
| gam3a* | 36.8883 | -2506.236 | 5086.24 | 0.273 | 28.5% | 0.60913 |
| gam3b* | 27.2397 | -868.2249 | 1790.92 | 0.646 | 65.4% | 0.31855 |
| gam3c* | 38.0702 | -2709.399 | 5494.93 | 0.387 | 39.6% | 0.65849 |
| gam4a | 41.6366 | -3981.53 | 8046.34 | 0.379 | 38.6% | 0.65859 |
| gam4b | 39.6833 | -4000.22 | 8079.81 | 0.372 | 37.9% | 0.66525 |
| gam4c | 19.8326 | -4199.23 | 8438.13 | 0.297 | 30% | 0.74089 |
| gam4d | 34.10861 | -3998.45 | 8065.12 | 0.374 | 38% | 0.66229 |
| gam4e | 40.96925 | -3955.16 | 7992.27 | 0.389 | 39.6% | 4066.5 REML |

```
gam1 = log(Mean+0.5)~s(Month)+s(Year),data=mydata)
gam2 = log(Mean+0.5)~s(Month)+s(Year)+s(Easting,Northing),
data=mydata)
gam3a =log(Mean[1:2172]+0.5)~s(Month[1:2172])+s(Year[1:2172])
+s(E[1:2172],N[1:2172]),data=mydata)
gam3b =log(Mean[2173:3228]+0.5)~s(Month[2173:3228])+s(Year
[2173:3228])+s(E[2173:3228],N[2173:3228]),data=mydata)
gam3c =log(Mean+0.5)~s(Month)+s(Year)+s(Easting,Northing),
data=mydat11)
gam4a= log(Mean+0.5)~s(Month,bs=cr)+s(Year,bs=cr)+s(Easting,
Northing,bs=tp),data=mydata)
gam4b=log(Mean+0.5)~s(Month,bs=ps)+s(Year,bs=ps)+
s(Easting,Northing,bs=tp),data=mydata)
gam4c=log(Mean+0.5)~s(Month,k=6)+s(Year,k=6)+s(Easting,
Northing,k=12),data=mydata)
gam4d=log(Mean+0.5)~s(Month,k=12)+s(Year,k=12)+s(Easting,
Northing,k=20),data=mydata)
gam4e=log(Mean+0.5)~s(Month)+s(Year)+s(Easting,Northing),
method=REML),data=mydata)
```

### 5.4.1 Model validation

Model validation was used to test the reliability of our models as in Chapter 4. Firstly, we completely removed data from 10 stations at random from our observations, and fitted a separate model similar to model 2 to the remaining data to use for predictions for the 10 held-out stations. We test the predictive ability of our models using the AIC, GCV and $R^2$ criteria. We perform the model validation using the dataset for all the years.

We based the new analysis on model 2. The model used a thin plate regression (default) for both univariate and bivariate smoothers. We also used the GCV estimation method similar to most of the previous models.

The results are presented in Table 5.12. The $R^2$ and deviance explained are 0.395 and 40.4%, which are similar to the results of previous models. Also, predictions from the fitted models are compared with the true value of $SO_2$. The predicted values for the held-out observations are mostly within the range of the observed data. We also compare the levels for those excluded stations using prediction errors (held-out observations - prediction). The calculated average square prediction error is 0.004761, which is very small, and which shows that the model is reliable for $SO_2$ level prediction.

Table 5.12: Model validation results for the reduced model (31 stations)

```
> summary(gam4f)
log(Mean+0.5)~s(Month) + s(Year) + s(Easting,Northing),data=mydat1)
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.07610    0.01571   132.1   <2e-16 ***
Approximate significance of smooth terms:
                       edf Ref.df       F p-value
s(Month)             2.376  2.957   3.532  0.0147 *
s(Year)              8.973  9.000 124.809  <2e-16 ***
s(Easting,Northing) 28.426 28.953  19.359  <2e-16 ***
R-sq.(adj) =  0.395   Deviance explained = 40.4%
GCV score = 0.66041  Scale est. = 0.65018   n = 2634
-----------------------------------------------------
Summary statistics for the held-out observation from 10 stations
  Min.    1st Qu. Median  Mean   3rd Qu. Max.
  0.986   1.821   2.391   2.145   2.777  4.240

-----------------------------------------------------
> held-out (model) predictions
   Min.  1st Qu.  Median   Mean   3rd Qu. Max.
  0.817   1.776   2.124    2.076  2.493   3.480
```

## 5.5 Discussion and further considerations

The main objective is to obtain a statistical model to predict and estimate $SO_2$ concentrations across Scotland spatially and temporally simultaneously. The joint effect of the predictor variables from the best model explained more than 39.3% of the variance of the dependent variable, which corresponds to the model with REML estimation. High predictions for $SO_2$ are observed in Central Scotland, in accordance with the results of the Bayesian kriging predictions of Chapter 4. We can infer that spatial location has a considerable effect on $SO_2$ levels in Scotland. This study adopted local scoring with backfitting using the penalised regression splines of the *mgcv* package in R. Each smooth in the GAM is represented by an appropriate basis set, in which the set of smoothing parameters are generally chosen by Generalized Cross-Validation (GCV). We considered a bivariate spatial term, and univariate smoothing of year and month. We adopted Akaike's Information Criterion (AIC) and $R^2$ as criteria for model comparison (Hastie and Tibshirani, 1990; Wood, 2006). A best choice for the smoothing parameter is unknown but further modelling of this data may provide some insights.

The model results rely on the basis dimension and choice of basis, but changing these did not greatly affect model fit. The higher $R^2$ in the models with the bivariate spatial term compared to the one without this term indicates that specific location generally has a large effect on the $SO_2$ level, as would be expected.

Bivariate spatial location is the most influential factor in predicting sulphur dioxide concentrations in our models. The effect of year was always significant in our models, though month was not always significant. The unexplained variation in the models we considered may be due to unmodelled covariates (wind direction, temperature, altitude, distance to the nearest major road etc.) that could have important effects on the $SO_2$ levels in Scotland. Any form of interactions between these covariates could also explain more of the variability, and any yet undiscovered interactions or linear combinations of the chosen covariates (year, month and spatial locations) could also be used to build a better model.

# Chapter 6

# Conclusion and further consideration

### 6.0.1 Conclusion

We have been able to analyse the $SO_2$ data both temporally in Chapter 3 and spatially in Chapter 4, as well as doing joint spatio-temporal modelling in Chapter 5. Chapter 1 gave an insight to the structure and location of the datasets. We are able to see that there is variation in the levels of $SO_2$ both within year (seasonal variation) and across the years, and that there is non-constancy of variance and evidence of skewness in the data, and as a result we adopted a logarithmic transformation of the data. Most of the stations we considered are concentrated in Central Scotland and are heavily characterized with missing observations.
Chapter 2 examined previous work on air pollution data and the sorts of models used to describe it. In Chapter 3, we explore various imputation techniques and apply them differently in several models. We observe that that each of the imputation techniques produces different results though we use EM as the most appropriate method in this thesis. The ARIMA (1,1,1) model is able to model the $SO_2$ data very well. There is evidence of temporal correlation in the data as suggested by the ACF and PACF. We observe that both the long-term trend and seasonal/cyclical effects contribute significantly to variation in $SO_2$ levels across the years for most of the stations we considered.
In Chapter 4, the kriging models with constant mean trend seem better than those with linear model, and the Exponential function tends to reduce the estimated kriging variance. We also observed that the Bayesian models produce a lower variance than ordinary kriging estimates. Ordinary kriging does not display the underlying spatial pattern very well, especially for the stations around

Aberdeen. Bayesian kriging gives better prediction than ordinary kriging when we based our analysis on stations in Central Scotland. There is an increase in variance with the reduced model after removal of stations that are far away from the rest. A low spatial variation is observed in Central Scotland and this could be due to concentration of the stations in the region. Of course we would expect the data to be high spatially correlated (dependent).

Most of the formulated models indicated high concentration of $SO_2$ in and around Central Scotland (Glasgow and Edinburgh), and elevated concentrations are also seen in parts of Eastern Scotland, which may be attributed to altitude and wind direction as well as the location of the primary source of $SO_2$ in the region, while low mean level is observed in North-Eastern locations (remote stations except Aberdeen) and along the Eastern coast. We also observe that there Bayesian variance increases as the data points reduce in number.

In Chapter 5, where generalized additive models are used, a penalised regression spline was adopted. The model is fitted by penalized least squares in which the set of smoothing parameters are chosen by GCV. We considered a bivariate smoothing for the spatial location, and univariate smoothing for both year and month.

The model results depend on the degree of smoothing used, basis dimension and choice of basis, and varying these choices did not affect the results very much. In most cases we allowed automatic smoothing parameter selection through generalized cross-validation. Location generally has a large effect on distribution of sulphur dioxide. The GAM technique we adopted reduces some of the problems of model mis-specification that affect linear models and generalized linear models.

The joint effect of the predictor variables from the best model explained just 65.4% of the variance of the dependent variable corresponding to the period 2001-2005. High predictions of $SO_2$ are observed in Central Scotland, as for Bayesian kriging predictions. The next best model, for all the years only explained 39.6% of the variation in $SO_2$ which corresponds to the model with REML estimation. The unexplained variation in most of the models we considered may be due to unmodelled covariates (wind direction, temperature, altitude, distance to nearest major road etc.) that could have important effects on $SO_2$ levels in Scotland.

### 6.0.2 Further considerations

Increasing the dataset to incorporate more stations is likely to further reduce the kriging variance. Averaging $SO_2$ levels over more than a year may also reduce

the prediction error. We could also consider other forms of transformation to normalize our data apart from the logarithm, and explore other imputation techniques.

Further improvement may also be gained from taking into account other covariates that affect $SO_2$ concentration, such as wind direction and distance to a major road, which we would have included in our analysis but did not include because of non-availability of data on these covariates.

Lastly, interactions between these covariates could also explain more of the variability in $SO_2$ distribution, as well as any interactions or linear combinations of considered covariates (year, month and spatial locations), and these could also be used to build a better model.

# Appendices

## Appendix 1

This section shows the R code for preliminary data description in Chapter 1. We used *maptool* and *ts* library in R.

```
A.1.1 #R-code for producing the maps of spatial
location in Figures 1.1-1.4#
>f1<-read.csv("E:/combineds.csv",sep=",")
>z=f1
>plot(HBA,col="pink",border="grey",xlim=z.xlim,ylim=z.ylim)
>text(z[,c("e1")],z[,c("n1")],seq(1,6),cex=0.4)
> legend(c("bottomright"),550000+c(0,z.plot.width),z[,"s1"]
,ncol=1,cex=.6, title(main=paste("Map of all recording station"))

>f2<-read.csv("E:/80dat.csv",sep=",")
>z=f2
>plot(HBA,col="pink",border="grey",xlim=z.xlim,ylim=z.ylim)
>title(main=paste("station according > 80% of available data"))
> legend(c("bottomright"),550000+c(0,z.plot.width),z[,"s2"],ncol=1
,cex=.6, title(main=paste("Map showing the 6 stations with greater
 than 80% of data available"))

>f3<-read.csv("E:/20dat.csv",sep=",")
>z=f3
>plot(HBA,col="pink",border="grey",xlim=z.xlim,ylim=z.ylim)
>title(main=paste("station < 20% of available data"))
> legend(c("bottomright"),550000+c(0,z.plot.width),z[,"s3"]
,ncol=1,cex=.6, title(main=paste("Map showing the 3 stations
with less than 20% of data available"))
```

```
>f3<-read.csv("E:/comb.csv",sep=",")
>z=f4
>plot(HBA,col="pink",border="grey",xlim=z.xlim,ylim=z.ylim)
>title(main=paste("station < 20\% of available data"))
> legend(c("bottomright"),550000+c(0,z.plot.width),z[,"s4"]
,ncol=1,cex=.6, title(main=paste("Map showing the stations
 with less than 20-80\% of data available"))
```

A.1.2: #R-code for producing preliminary time series plots of daily S02 and long-term trend for 1996, 2000 and 2005 in Figures 1.5-1.8#

```
> par(mfrow=c(7,1))
> for (i in 1:7)plot(ts(d1), xlim=c(1,366),main=
("Time series plot of daily mean S02 concentrations
for some stations in 1996"),xlab="day",col="green")

> par(mfrow=c(7,1))
> for (i in 1:7)plot(ts(d5), xlim=c(1,366),main=("
Time series plot of daily mean S02 concentrations
for some stations in 2000"),xlab="day",col="green")

> par(mfrow=c(6,1))
> for (i in 1:6)plot(ts(d10), xlim=c(1,366),main=
("Time series plot of daily mean S02 concentrations
 for some stations in 2005"),xlab="day",col="green")

> par(mfrow=c(2,1))
> plot(ts(d11,,start=1996,end=2005),main="Long term
 trend of daily S02 concentrations for Glasgow.73 and Glasgow.
95",xlab="year",col="green")
```

A.1.3: #R-code for producing monthly time plots for 1996, 2000 and 2005 in Figures 1.9-1.11#

```
> for (i in c(1:4))plot(ts(mth1), xlim=c(1,12),
main="Time series plot of monthly mean S02 for
some stations in 1996",xlab="month",col="green")
```

```
> plot(ts(mth5), xlim=c(1,12),main="monthly concentration
",main="Time series plot of monthly mean S02 for
some stations in 2000",xlab="month",col="green")
> plot(ts(mth10), xlim=c(1,12),main="monthly
concentration",main="Time series plot of monthly
mean S02 for some stations in 2005",xlab="month",col="green")
```

A.1.4: #R-code for producing histograms for 1996, 2000 and 2005 in Figures 1.12-1.14#

```
> par(mfrow=c(2,3))
> for(i in 1:6) hist(log(t(d1[i+3])), xlab="
daily log(conc). g/m^3",main=paste("",names(d1[i])))
> for(i in 1:6) hist(t(log(d5[i+3])), xlab="
daily log(conc). g/m^3",main=paste("",names(d1[i])))
> for(i in 1:6) hist(log(t(d10[i+3])), xlab="
daily log(conc). g/m^3",main=paste("",names(d5[i])))
```

A.1.5: #R-code for producing boxplots of daily S02 concentration for 1996, 2000 and 2005 in Figures 1.15-1.17#

```
>rownames(mth1)<-c("Jan","Feb","Mar","Apr",
"May","Jun","Jul","Aug","Sep","Oct","Nov","Dec")
>rownames(mth5)<-c("Jan","Feb","Mar","Apr",
"May","Jun","Jul","Aug","Sep","Oct","Nov","Dec")
>rownames(mth10)<-c("Jan","Feb","Mar","Apr",
"May","Jun","Jul","Aug","Sep","Oct","Nov","Dec")
>boxplot(t(mth1),main="monthly concentration in 1996"),
xlab="month")
>boxplot(t(mth5),main="monthly concentration in 2000"),
xlab="month")
>boxplot(t(mth10),main="monthly concentration in 2005"),
xlab="month")
```

A.1.6: #R-code for variance versus mean of daily S02 concentration for 1996, 2000, 2002 and 2005 in Figure 1.18#

```
>plot(mean(d1,na.rm=TRUE),(sd(d1,na.rm=TRUE)^2),
xlab="mean",ylab="variance",col="green",main="1996")
```

```
>plot(mean(d5,na.rm=TRUE),(sd(d5,na.rm=TRUE)^2),
xlab="mean",ylab="variance",col="green",main="2000")
>plot(mean(d7,na.rm=TRUE),(sd(d7,na.rm=TRUE)^2),
xlab="mean",ylab="variance",col="green",main="2002")
>plot(mean(d10,na.rm=TRUE),(sd(d10,na.rm=TRUE)^2),
xlab="mean",ylab="variance",col="green",main="2005")
```

# Appendix 2

This section shows the R code for the generation of time series plots and boxplots after imputation, the autocorrelation and partial autocorrelation as well as for autoregressive AR and ARIMA models in Chapter 3. We used *maptool*, *ts*, *tsModel* and *MICE* in R as well as *SPSS*.

```
#d96b=EM imputed data in 1996#
#d96c=Regression imputed data in 1996#
#mic1=MICE imputed data in 1996#
#d20b=EM imputed data in 2000#
#d20c=Regression imputed data in 1996#
#mic5=MICE imputed data in 2000#
#d25b=EM imputed data in 2005#
#d25c=Regression imputed data in 2005#
#mic10=MICE imputed data in 2005#
#d27b=EM imputed data in 2007#
#d27c=Regression imputed data in 2007#
#mic12=MICE imputed data in 2007#
#combdata=EM imputed for aggregate data in 1996-2007#
```

A.2.1: #R-code for producing time series plots comparison of different imputation methods for the daily mean SO2 concentrations for stations. The upper panel represents EM (green), the middle panel is regression (blue), while the bottom panel is MICE (red) imputation for each station in 1996, 2000 and 2005
Figures 3.1-3.3#

```
>par(mfrow=c(1,3))
> plot(ts(96b[1:4]), xlim=c(1,365),main="EM imputation",xlab="day",
col="green",cex.lab =.7)
>plot(ts(d96c[1:4]), xlim=c(1,365),main="Regression imputation",xlab=
"day",col="blue",cex.lab =.7)
> plot(ts(mice1[1:4]), xlim=c(1,365),main="MICE",xlab="day",col=
"red",cex.lab =.7)

>par(mfrow=c(1,3))
> plot(ts(20b[1:4]), xlim=c(1,365),main="EM imputation",xlab="day",
```

```
col="green",cex.lab =.7)
>plot(ts(d20c[1:4]), xlim=c(1,365),main="Regression imputation",xlab=
"day",col="blue",cex.lab =.7)
> plot(ts(mic5[1:4]), xlim=c(1,365),main="MICE",xlab="day",col=
"red",cex.lab =.7)

>par(mfrow=c(1,3))
> plot(ts(25b[1:3), xlim=c(1,365),main="EM imputation",xlab="day",
col="green",cex.lab =.7)
>plot(ts(d25c[1:3]), xlim=c(1,365),main="Regression imputation",xlab=
"day",col="blue",cex.lab =.7)
>plot(ts(mic10[1:3]), xlim=c(1,365),main="MICE",xlab="day",col=
"red",cex.lab =.7)
```

A.2.2: #R-code for producing corresponding boxplots of comparison of different imputation methods for the logarithm of daily mean SO2 concentrations. In each row the first panel represents EM, the middle panel is regression, while the third panel is MICE imputation for each station in 1996, 2000 and 2005 Figures 3.4-3.6#

```
>par(mfrow=c(2,2))
>boxplot(log(d96b[1:4]+3),main="EM",cex.lab =.7,xlab="c(Falkirk 8,
Glasgow 20, Glasgow 51, Glasgow69)")
>boxplot(log(d96c[1:4]+3),main="Regression",cex.lab =.7,xlab="c
(Falkirk 8,Glasgow 20, Glasgow 51, Glasgow69)")
>boxplot(log(mic1[1:4]+3),main="MICE",xlab="c(Falkirk, Glasgow 20,
Glasgow 51, Glasgow69)",cex.lab =.7)

>par(mfrow=c(2,2))
>boxplot(log(d20b[1:4]+3),main="EM",cex.lab =.7,xlab="c(Glasgow 69,
Glasgow 73, Glasgow 95, Glasgow 98)")
>boxplot(log(d20c[1:4]+3),main="Regression",cex.lab =.7,xlab="
c(Glasgow 69, Glasgow 73, Glasgow 95, Glasgow 98)")
>boxplot(log(mic2[1:4]+3),main="MICE",xlab="c(Glasgow 69,
Glasgow 73, Glasgow 95, Glasgow 98)",cex.lab =.7)

>par(mfrow=c(2,2))
>boxplot(log(d25b[1:3]+3),main="EM",cex.lab =.7,xlab="c(Glasgow 20,
```

```
Glasgow 51, Glasgow 73)")
>boxplot(log(d25c[1:3]+3),main="Regression",cex.lab =.7,xlab=
"c(Glasgow 20, Glasgow 51, Glasgow 73)")
>boxplot(log(mic3[1:3]+3),main="MICE",xlab="c(Glasgow 20,
Glasgow 51, Glasgow 73)",cex.lab =.7)
```

A.2.3: #R-code for producing comparison of autocorrelation functions for the EM, regression and MICE imputed datasets for Glasgow 51, Glasgow 73, Glasgow 95, Kirkcaldy 6, Kirkintilloch 8 and Kirkintilloch 10 in 1996, 2000, 2007 in Figures 3.7-3.9#

```
>par(mfrow=c(3,3))
>acf(log(d96b[3:8]+3),main="EM",cex.lab =.7,xlab="c( Glasgow 51,
Glasgow 73, Glasgow 95, Kirkcaldy 6, Kirkintilloch 8, Kirkintilloch 10)")
>acf(log(d96c[3:8]+3),main="Regression",cex.lab =.7, xlab="c( Glasgow 51,
Glasgow 73, Glasgow 95, Kirkcaldy 6, Kirkintilloch 8, Kirkintilloch 10)")
>acf(log(mic1[3:8]+3),main="MICE",xlab="c( Glasgow 51,Glasgow 73,
Glasgow 95, Kirkcaldy 6,Kirkintilloch 8,Kirkintilloch 10)",cex.lab =.7)

>par(mfrow=c(3,3))
>acf(log(d20b[3:8]+3),main="EM",cex.lab =.7,xlab="c( Glasgow 51,Glasgow 73,
Glasgow 95, Kirkcaldy 6, Kirkintilloch 8, Kirkintilloch 10)")
>acf(log(d20c[3:8]+3),main="Regression",cex.lab =.7,xlab="c( Glasgow 51,
Glasgow 73, Glasgow 95, Kirkcaldy 6, Kirkintilloch, 8 Kirkintilloch 10)")
>acf(log(mic5[3:8]+3),main="MICE",xlab="c( Glasgow 51,Glasgow 73,
Glasgow 95, Kirkcaldy 6,Kirkintilloch 8,Kirkintilloch 10)",cex.lab =.7)

>par(mfrow=c(2,2))
> acf(log(d27b[1:4]+3),main="EM",sub="c(Glasgow Centre, Aberdeen,
Edinburgh St. Leonards, Grangemouth")
> acf(log(d27b[1:4]+3),main="EM",sub="c(Glasgow Centre, Aberdeen,
Edinburgh St. Leonards, Grangemouth")
> acf(log(d27b[1:4]+3),main="EM",sub="c(Glasgow Centre, Aberdeen,
Edinburgh St. Leonards, Grangemouth")
```

A.2.4: #R-code for producing comparison of partial autocorrelation functions for the EM, regression and MICE imputed datasets for Glasgow 51, Glasgow 73, Glasgow 95, Kirkcaldy 6, Kirkintilloch 8

```
and Kirkintilloch 10 in 1996, 2000, 2007 in Figures 3.10-3.12#

>par(mfrow=c(3,3))
>pacf(log(d96b[3:8]+3),main="EM",cex.lab =.7,xlab="c(Glasgow 51,
Glasgow 73, Glasgow 95, Kirkcaldy 6, Kirkintilloch 8, Kirkintilloch 10)")
>pacf(log(d96c[3:8]+3),main="Regression",cex.lab =.7
,xlab="c( Glasgow 51, Glasgow 73, Glasgow 95, Kirkcaldy 6, Kirkintilloch
8 Kirkintilloch 10)")
>pacf(log(mic1[3:8]+3),main="MICE",cex.lab =.7,xlab="c( Glasgow 51,
Glasgow 73, Glasgow 95, Kirkcaldy 6,Kirkintilloch, 8 Kirkintilloch 10)")

>par(mfrow=c(3,3))
>pacf(log(d20b[3:8]+3),main="EM",cex.lab =.7,xlab="c( Glasgow 51,
Glasgow 73, Glasgow 95, Kirkcaldy 6, Kirkintilloch 8, Kirkintilloch 10)")
>pacf(log(d20c[3:8]+3),main="Regression",cex.lab =.7,xlab="c( Glasgow 51,
Glasgow 73, Glasgow 95, Kirkcaldy 6, Kirkintilloch, 8 Kirkintilloch 10)")
>pacf(log(mic5[3:8]+3),main="MICE",xlab="c( Glasgow 51, Glasgow 73,
Glasgow 95, Kirkcaldy 6,Kirkintilloch 8,Kirkintilloch 10)",cex.lab =.7)

>par(mfrow=c(2,2))
> pacf(log(d27b[1:4]+3),main="EM",sub="c(Glasgow Centre,Aberdeen,
Edinburgh St. Leonards, Grangemouth")
> pacf(log(d27b[1:4]+3),main="EM",sub="c(Glasgow Centre, Aberdeen,
Edinburgh St. Leonards, Grangemouth")
>pacf(log(d27b[1:4]+3),main="EM",sub="c(Glasgow Centre, Aberdeen,
Edinburgh St. Leonards, Grangemouth")

A.2.5: #R-code for estimating autoregressive model parameters using
both the least squares and maximum likelihood methods in Table 3.5#

>ar(log(d96b+3),method="ols")
>ar(log(d20b+3),method="ols")
>ar(log(d25b+3),method="ols")
>ar(log(d96b+3),method="ML")
>ar(log(d20b+3),method="ML")
>ar(log(d25b+3),method="ML")

A.2.6: #R-code for producing comparison of autocorrelation functions for
```

the residuals using different imputation methods for the AR(2) model for Glasgow 51 and Glasgow 73 in 1996, as well as Glasgow Centre and Aberdeen in 2007; the top-left panel is Glasgow 51 in 1996, the top-right panel corresponds to Glasgow 73 in 1996, the bottom-left panel is Glasgow Centre in 2007 while the bottom-right panel is Aberdeen in 2007 in Figure 3.13#

```
> par(mfrow=c(2,2))
>acf(log(res1b[3:4]+3),main="EM",cex.lab =.7,xlab="c( Glasgow 51,
Glasgow 73)")
>acf(log(res1c[3:4]+3),main="Regression",cex.lab =.7,xlab="c(Glasgow 51,
Glasgow 73)")
>acf(log(res1d[3:4]+3),main="MICE",cex.lab =.7,xlab="c(Glasgow 51,
Glasgow 73)")
> acf(log(res2b[1:2]+3),main="EM",sub="c(Glasgow Centre, Aberdeen")
> acf(log(res2c[1:2]+3),main="EM",sub="c(Glasgow Centre, Aberdeen")
> acf(log(res2d[1:2]+3),main="EM",sub="c(Glasgow Centre, Aberdeen")
```

A.2.7: #R-code for producing estimating ARIMA (2,0,0) model result of the MICE, EM and regression imputed dataset using maximum likelihood method for all the stations in Table 3.6#

```
>arima(x = log(rc1 + 3), order = c(2, 0, 0), method = "ML")
>arima(x = log(rc2 + 3), order = c(2, 0, 0), method = "ML")
>arima(x = log(rc3 + 3), order = c(2, 0, 0), method = "ML")
```

A.2.8: #R-code for estimating ARIMA model parameters using likelihood method for comparison ARIMA model results using the EM imputed dataset with maximum likelihood method for the combined station in Table 3.7#

```
>arima(x = log(rc2 + 3), order = c(3, 0, 0), method = "ML")
>arima(x = log(rc2 + 3), order = c(1, 0, 1), method = "ML")
>arima(x = log(rc2 + 3), order = c(2, 1, 0), method = "ML")
> arima(log(rc2+3), order =c(3,1,0), method = "ML")
> arima(log(rc2+3), order =c(1,1,1), method = "ML")
```

A.2.9: #R-code for comparison of Ljung-Box test for the residuals of the six ARIMA models, lag=25, type="Ljung-Box"; the results give the chi-squared, degree of freedom and p values in Table 3.9#

```
> ARIMA(2,0,0)$residuals
> ARIMA(3,0,0)$residuals
> ARIMA(1,0,1)$residuals
> ARIMA(2,1,0)$residuals
> ARIMA(3,1,0)$residuals
> ARIMA(1,1,1)$residuals
```

A.2.10: #R-code for comparison of Ljung-Box test for the residuals of the six ARIMA models, lag=25, type="Ljung-Box" in Figure 3.14#

```
>par(mfrow=c(2,3))
>Box.test(arima(x =log(rc1 + 3),order=c(2, 1, 0),method= "ML")$residual)
>Box.test(arima(x =log(rc2 + 3),order=c(1, 0, 1),method= "ML")$residual)
>Box.test(arima(x =log(rc3 + 3),order=c(3, 0, 0),method= "ML")$residual)
>Box.test(arima(x =log(rc2 + 3),order=c(3, 1, 0),method= "ML")$residual)
>Box.test(arima(x =log(rc2 + 3),order=c(1, 1, 1),method= "ML")$residual)
```

A.2.11: #R-code for producing long-term trend decomposition of EM imputed daily log(S02) concentrationfor in Figures 3.15 and 3.16#

```
>plot(ts(combdata,frequency=365),ylab="log(S02)conc",xlim=c(1996,2007))
>decomp1=tsdecomp(combdata,c(1,10,22,4382))
>xx=seq(as.Date("1996-01-01"),as.Date("2007-12-31"),"day")
>par(mfrow=c(1,3))
>plot(xx,decomp1[,1],type="l",ylab="log(S02)conc",main="Trend")
>plot(xx,decomp1[,2],type="l",ylab="log(S02)conc",main="Seasonal")
>plot(xx,decomp1[,3],type="l",ylab="log(S02)conc",main="Residual")
```

# Appendix 3

We used the *geoR*, *ape*, *field* and *spline* packages in *R* for our spatial analysis in Chapter 4. The function *krige.bayes* performs Bayesian analysis of geostatistical data allowing specifications of different levels of uncertainty in the model parameters. It gives results on the posterior distributions for the model parameters and on the predictive distributions for prediction locations, while *krige.conv* performs spatial prediction for ordinary kriging using a fixed covariance parameters.
The data-frame for our geodata object is represented by $g8$, in which each line corresponds to one spatial location and consists of the $log(SO_2)$, the Easting and the Northing. The dataframe for the next geodata object is represented by $g8b$ and $g8c$ which correspond to Central Scotland and remote stations respectively, and consists of the $log(SO_2)$, the Easting and the Northing.
Most of the code here relates to kriging, using a function call similar to the one below.

```
#krige.bayes(geodata, coords = geodata$coords, data = geodata$data,
locations = "no", borders, model, prior, output)#

#geodata= a list containing elements coords and data#
#coords= an n by 2 matrix where each row has the 2-dim. coordinates
of the n-data locations#
#data= a vector with n data values#
#locations= an N by 2 matrix or data-frame with the 2-d
coordinates of the N prediction locations#
#output=output.control(n.posterior, n.predictive, moments, n.back.moments,
simulations.predictive, mean.var, quantile, threshold, sim.means, sim.vars,
signal, messages)#
#trend.d= specifies the trend (covariates) values at the data locations#
#trend.l= specifies the trend (covariates) at the prediction locations#
#cov.model= string indicating name of model for the correlation function#
```

A.3.1: #R-code for producing Moran's I test for the datasets in 1996, 2000 and 2005, and for geodetic distance summary in Tables 4.1 and 4.2#

```
>g8=as.geodata(mean96,coords.col=3:4,data.col=2)
>mydist1=as.matrix(dist(cbind(g8$coords[,1], g8$coords[,2])))
>mydist1a=1/mydist1
>diag(mydist1a)<-0
```

```
>Moran.I(g8$data,mydist1a)

> mydist2=as.matrix(dist(cbind(g9$coords[,1], g9$coords[,2])))
> mydist2a=1/mydist2
> diag(mydist2a)<-0
> Moran.I(g9$data,mydist2a)

> mydist3=as.matrix(dist(cbind(g10$coords[,1], g10$coords[,2])))
> mydist3a=1/mydist3
> diag(mydist3a)<-0
> Moran.I(g10$data,mydist3a)

>summary(dist(g8$coords))
```

A.3.2:#R-code for empirical and theoretical variograms in Figures 4.1-4.2#

```
> plot(variog(g8,trend = "cte"),xlab="distance (metres)")
> lines(variog(g8,trend = "cte"))
> plot(variog(g8,trend = "linear",),xlab="distance (metres)")
> lines(variog(g8,trend = "linear"))
```

A.3.3:#R-code for empirical and theoretical variograms in Figures 4.3-4.4#

```
> plot(variog(g8,trend = "cte"),xlab="distance (metres)")
> lines(variog(g8,trend = "cte"))
> lines.variomodel(variog(g8,trend = "cte"),cov.model
 = "spherical",cov.pars=c(4.1,2.1),col="pink",kappa=1)
> lines.variomodel(variog(g8,trend = "cte"),cov.model
= "gaussian",cov.pars=c(4,1.2),col="red",kappa=1)
> lines.variomodel(variog(g8,trend = "cte"),cov.model
 = "exponential",cov.pars=c(4.1,1.1),col="blue",kappa=1)
> lines.variomodel(variog(g8,trend = "cte"),cov.model
 = "matern,,cov.pars=c(4.1,1.1),col="green",kappa=1)

> plot(variog(g8,trend = "1st"),xlab="distance (metres)")
> lines(variog(g8,trend = "1st"))
> lines.variomodel(variog(g8,trend = "linear"),cov.model
= "spherical",cov.pars=c(4.1,2.1),col="pink",kappa=1)
```

```
> lines.variomodel(variog(g8,trend = "linear"),cov.model
= "gaussian",cov.pars=c(4,1.2),col="red",kappa=1)
> lines.variomodel(variog(g8,trend = "linear"),cov.model
= "exponential",cov.pars=c(4.1,1.1),col="blue",kappa=1)
> lines.variomodel(variog(g8,trend = "linear"),cov.model
= "matern,,cov.pars=c(4.1,1.1),col="green",kappa=1)
```

A.3.4: #R-code for Plotting data locations in Figures 4.5#

```
> plot(g8)
```

A.3.5: #R-code for estimated model parameters with Matern covariance function for constant mean trend using likelihood method in Tables 4.3, 4.4 and 4.5#

Model 1a

```
> mgrid3=expand.grid(seq(46343.45,603656.55,,50)
,seq(532000,1018000,,50)
>out1=output.control(n.posterior=1000, n.predictive=1000,
n.back.moments=1000,simulations.predictive=T, mean.var=T)
>mod1a=likfit(geodata = g8, ini.cov.pars = c(0.35, 50000),
kappa = 0.5,cov.model = "matern",trend="cte")
> summary(mod1a)
>kr1a=krige.conv(g8,loc=mgrid3,krige=krige.control
(obj.m=mod1a,trend.d="cte",trend.l="cte"),out=out1)
```

Model 1b

```
>mod1b=likfit(geodata = g8, ini.cov.pars = c(0.35, 50000),
kappa = 0.5,cov.model = "exponential", trend="cte")
> summary(mod1b)
>kr1b=krige.conv(g8,loc=mgrid3,krige=krige.control
(obj.m=moda 11,trend.d="cte",trend.l="cte"),out=out1)
```

A.3.6: #R-code for plotting the likelihood fit result of the ordinary kriging using Matern and Exponential functions in Figures 4.6-4.9#

```
>image(kr1a,grid=mgrid3,xlim=c(46300,700000),ylim=c(532000,1018000),
 x.leg=c(640000,660000), y.leg=c(532000,1015000),vert=TRUE)
>plot(HBA,border="grey",xlim=c(46300,700000),ylim=c(532000,1018000))
points(g8,add=T)
```

```
>image(kr1a,val="variance",xlab="Easting",ylab="Northing",main="
variance",grid=mgrid3,xlim=c(46300,700000),ylim=c(532000,1018000),
 x.leg=c(650000,670000), y.leg=c(532000,1018000),vert=TRUE)

>image(kr1b,grid=mgrid3,xlim=c(46300,700000),ylim=c(532000,1018000)
, x.leg=c(640000,660000), y.leg=c(532000,1015000),vert=TRUE)
>plot(HBA,border="grey",xlim=c(46300,700000),ylim=c(532000,1018000))
points(g8,add=T)
>image(kr1b,val="variance",xlab="Easting",ylab="Northing",main="variance"
,grid=mgrid3,xlim=c(46300,700000),ylim=c(532000,1018000), x.leg=c
(650000,670000), y.leg=c(532000,1018000),vert=TRUE)
```

A.3.7: #R-code for likelihood fit result of Bayesian kriging with constant mean trend using both Matern and Exponential functions in Tables 4.6 and 4.7#

```
Model 2a
>kr2a=krige.bayes(g8,loc=mgrid3,model=moda1)
> moda1=model.control(trend.d = "cte", trend.l = "cte",cov.model=
"matern",kappa = 0.5)
>kr2b=krige.bayes(g8,loc=mgrid3,model=moda2)
>moda2=model.control(trend.d = "cte",trend.l = "cte",cov.model=
"exponential",kappa =0.5)
```

A.3.8: #R-code for likelihood fit result of Bayesian kriging using both Matern and Exponential functions in Figures 4.10-4.14#

```
#Figures 4.10 and 4.11#
>plot(HBA,border="grey",xlim=c(46300,700000),ylim=c(532000,1018000))
>image(kr2a,xlab="Easting",ylab="Northing",main="mean of simulation",
grid=mgrid3,xlim=c(46300,700000),ylim=c(532000,1018000),x.leg=c(640000,
660000),y.leg=c(532000,1018000),vert=TRUE)
>points(g8,add=T)
>plot(HBA,border="grey",xlim=c(46300,700000),ylim=c(532000,1018000))
> image(kr2a,val="variance",xlab="Easting",ylab="Northing",main="
variance",grid=mgrid3,xlim=c(46300,700000),ylim=c(532000,1018000),
x.leg=c(640000,660000),y.leg=c(532000,1018000),vert=TRUE)
>points(g8,add=T)
```

```
#histograms of posteriors (beta, sigmasq and phi) in Figure 4.12#
>par(mfrow=c(3,1))
>kr2a$posterior$sample$sigmasq=1/kr2a$posterior$sample$sigmasq
>hist(kr2a$posterior$sample$beta)
>hist(kr2a$posterior$sample$sigmasq)
>hist(kr2a$posterior$sample$phi)

>par(mfrow=c(3,1))
>kr2b$posterior$sample$sigmasq=1/kr2b$posterior$sample$sigmasq
>hist(kr2b$posterior$sample$beta)
>hist(kr2b$posterior$sample$sigmasq)
>hist(kr2b$posterior$sample$phi)

#Figures 4.13 and 4.14#
> kr2b$predictive$variance=1/kr2b$predictive$variance
>plot(HBA,border="grey",xlim=c(46300,700000),ylim=c(532000,1018000))
>image(kr2b,xlab="Easting",ylab="Northing",main="mean of simulation",
grid=mgrid3,xlim=c(46300,700000),ylim=c(532000,1018000),x.leg=c(640000,
660000),y.leg=c(532000,1018000),vert=TRUE)
>points(g8,add=T)
>plot(HBA,border="grey",xlim=c(46300,700000),ylim=c(532000,1018000))
>image(kr2b,val="variance",xlab="Easting",ylab="Northing",main="variance",
grid=mgrid3,xlim=c(46300,700000),ylim=c(532000,1018000),x.leg=c(640000,
660000),y.leg=c(532000,1018000),vert=TRUE)
>points(g8,add=T)
```

A.3.9: #R-code for producing likelihood fit results of the ordinary kriging using both Matern and Exponential functions using linear trend in Tables 4.8 and 4.9#

Model 3

```
>moda3=likfit(geodata = g8, ini.cov.pars = c(0.35, 50000),kappa = 0.5,
trend="linear",cov.model = "matern")
>moda4=likfit(geodata = g8, ini.cov.pars = c(0.35, 50000), kappa = 0.5,
trend="linear",cov.model = "exponential")
```

Model 4

```
>kr3a=krige.conv(geodata = g8, locations = mgrid3, krige =krige.control
(obj.m =moda3,trend.d = "linear",trend.l = "linear"), output = out1)
>kr3b=krige.conv(geodata = g8, locations =mgrid3, krige =krige.control
(obj.m =moda4,trend.d = "linear",trend.l = "linear"), output = out1)
```

A.3.10: #R-code for producing likelihood fit result of the ordinary kriging using the Matern covariance function and constant mean trend for Central Scotland in Tables 4.10 and 4.11#

```
Model 5a
mgrid3b=expand.grid(seq(243102,350752,,50),seq(635520,724512,,50)
moda5=likfit(geodata = g8b, ini.cov.pars = c(0.35, 40000),kappa =
0.5,trend="cte",cov.model = "matern")
>moda6=likfit(geodata = g8b, ini.cov.pars = c(0.35, 50000),
kappa = 0.5,cov.model = "matern")

>kr7aa=krige.conv(g8b,loc=mgrid3b,krige=krige.control(obj.m=moda6,
trend.d="cte",trend.l="cte"),out=out1)
>kr8b=krige.bayes(geodata = g8b, locations = mgrid3b,output=out1)
```

A.3.11: #R-code for producing likelihood fit results of Bayesian kriging using the Matern function and constant mean trend for Central Scotland in Figures 4.15-4.19#

```
> plot(g8b)

>mgrid3b=expand.grid(seq(243102,350752,,50),seq(635520,724512,,50)
>plot(HBA,border="grey",xlim=c(243102,375752 ),ylim=c(635520,724512))
>image(kr7aa,xlab="Easting",ylab="Northing",main="mean",grid=mgrid3,
xlim=c(46300,510000),ylim=c(532000,800000),x.leg=c(465000,480000),
y.leg=c(532000,800000),vert=TRUE)
>points(g8b,add=T)

>plot(HBA,border="grey",xlim=c(243102,375752 ),ylim=c(635520,724512))
>image(kr7aa,val=krige.var,xlab="Easting",ylab="Northing",main="
variance",grid=mgrid3,xlim=c(46300,510000),ylim=c(532000,800000),x.leg
=c(465000,480000),y.leg=c(532000,800000),vert=TRUE)
>points(g8b,add=T)
```

```
>plot(HBA,border="grey",xlim=c(243102,375752 ),ylim=c(635520,724512))
>image(kr8aa,xlab="Easting",ylab="Northing",main="mean",grid=mgrid3,
xlim=c(46300,510000),ylim=c(532000,800000),x.leg=c(465000,480000),y.leg
=c(532000,800000),vert=TRUE)
>points(g8b,add=T)

> kr8aa$predictive$variance=1/kr8aa$predictive$variance
>plot(HBA,border="grey",xlim=c(243102,375752 ),ylim=c(635520,724512))
>image(kr8aa,val="variance",xlab="Easting",ylab="Northing",main="variance",
grid=mgrid3,xlim=c(46300,510000),ylim=c(532000,800000),x.leg=c(465000,
480000),y.leg=c(532000,800000),vert=TRUE)
>points(g8b,add=T)
```

A.3.12: #R-code for producing likelihood fit result of Bayesian kriging using Matern covariance function and constant mean trend for the 6 remote stations in Table 4.12 and Figures 4.20 and 4.21#

```
Model 6
>kr9=krige.bayes(geodata = g8c, locations = mgrid3,output=out1)
> summary(kr9$predictive$mean)
> summary(kr9$predictive$variance)

> image(kr9,xlab="Easting",ylab="Northing",main="mean",grid=mgrid3,
xlim=c(300000,450000),ylim=c(700000,900000),x.leg=c(530000,540000)
,y.leg=c(700000,900000),vert=TRUE)
> par(new=T)
> plot(HBA,border="grey",xlim=c(300000,560000),ylim=c(700000,900000))
> points(g8c,add=T)

> kr9$predictive$variance=1/kr9$predictive$variance
> image(kr9,val="variance",xlab="Easting",ylab="Northing",main=
"variance",grid=mgrid3,xlim=c(300000,560000),ylim=c(700000,
900000),x.leg=c(530000,540000),y.leg=c(700000,900000),vert=TRUE)
> points(g8c,add=T)
> par(new=T)
> plot(HBA,border="grey",xlim=c(300000,560000),ylim=c(700000,900000))
```

A.3.13: #R-code for model validation tests, we randomly removed 10 stations from the 41 recording stations, we fit a model to the remaining 31 stations, and also used it to predict the held-out observations in Table 4.13 and Figure 4.22#

```
#g8wa=dataframe containing 31 remaining stations (training data#
#g8w=dataframe corresponds to 10 held-out stations (test data)#

>moda7= likfit(geodata = g8wa, ini.cov.pars = c(0.35, 50000),
kappa = 0.5, cov.model = "mat")
>xval1=xvalid(g8wa,model=moda7,reest=TRUE,locations.xvalid=
g8w$coords,data.xvalid=g8w$data
>summary(xval1$data)
>summary(xval1$predicted)
>summary(xval1)
> par(mfrow=c(5,2))
>plot(xval1)
```

# Appendix 4

We used both *mgcv* and *akima* for fitting a generalized additive model (GAM). The dataframe for our gam object for the whole of Scotland is stored in *mydata*, which consists of the $SO_2$, and is represented by the "Mean", the Easting and Northing, and the factors for the year and month. We use a Gaussian distribution and identity link function in all the models with default setting for the basis and basis dimension (except in sensitivity analysis of the models 4a-4f). The following code generated the various gam models in Chapter 5.

```
A.4.1:#R-code for simple additive model without bivariate spatial location#
#mydata= dataframe for whole of Scotland#
#mydat11= dataframe for Central Scotland stations#
#mydat1=dataframe for training data for model validation#
#Mean= SO2#
#R-code for plotting diagnostics check in Figure 5.1#

Model 1
>gam1 =gam(log(Mean+0.5)=s(Year) + s(Month),family="Gaussian",data=mydata)
> gam.check(gam1)

#generate Table 5.1#

>summary(gam1)
#plot yearly and monthly effects in Figure 5.2 and 5.3#

>plot(gam1,select=1,xlab="Year")
>plot(gam1,select=2,xlab="Month")

A.4.2:#R-code for simple additive model with bivariate spatial location#

Model 2
>gam2 = gam(log(Mean + 0.5) ~ s(Easting, Northing) + s(Year) +s(Month),
family="Gaussian", data=mydata)

#generate model summary in Table 5.2#
>summary(gam2)

#plot bivariate spatial location in Figure 5.4#
```

```
>plot(gam2,select=1,xlab="Easting",ylab="Northing")

#plot yearly and monthly effects in Figures 5.5 and 5.6

>plot(gam2,select=2,xlab="Year")
>plot(gam2,select=3,xlab="Month")

#diagnostics check in Figure 5.7
> gam.check(gam2)
```

A.4.3:#R-code for additive model with spatial interaction for the 1996-2000 and 2001-2005 datasets for Tables 5.3 and 5.4#

```
Model 3
> gam3a=log(Mean[1:2172] + 0.5) = s(Easting[1:2172], Northing[1:
2172])+ s(Year2[1:2172], k = 6) + s(Month[1:2172],k = 6,data=mydata)
>gam3b=log(Mean[2173:3228]+0.5)=s(Easting[2173:3228],Northing[2173:
3228],k = 20) + s(Year2[2173:3228], k = 5) + s(Month[2173:3228],
k = 5,data=mydata)
>summary(gam3a)
>summary(gam3b)

#plot additive model effects with spatial interaction for the separate
1996-2000 and 2001-2005 dataset in Figures 5.8 and 5.9#


>plot(gam3a,select=1,cex.lab=2.2,cex.axis=3.5,cex.main=2.5)
>plot(gam3b,select=1,cex.lab=2.2,cex.axis=3.5,cex.main=2.5)

>par(mfrow=c(2,2))
>plot(gam3a,select=2,xlab="Year")
>plot(gam3a,select=3,xlab="Month")
>plot(gam3b,select=2,xlab="Year")
>plot(gam3b,select=3,xlab="Month")
```

A.4.4: # R-code for simple additive model with bivariate spatial location for Central Scotland stations#

```
Model 3c
>gam3c = gam(log(Mean + 0.5) ~ s(Easting, Northing) + s(Year)+
s(Month),family="Gaussian", data=mydat11)

#generate model summary for Central Scotland in Table 5.5#
>summary(gam3c)

#plot bivariate spatial location, yearly and monthly in Figures
 5.10 and 5.11#

> plot(gam3c,select=1,cex.lab=1.2,cex.axis=1.5,cex.main=1.2)
> par(mfrow=c(1,2))
> plot(gam3c,select=2,xlab="Year")
> plot(gam3c,select=3,xlab="Month")
```

R-code for analysing sensitivity to the choice of basis for the cubic regression. Model 2 was extended to accomondate other bases (cubic regression and p-spline) for the univariate cases and basis dimension of ($k = 6$ and 12), and ($k = 12$ and 20) for univariate and bivariate smoothers respectively and inclusion of REML estimation method still considering the whole of Scotland data.

```
A.4.5: #R-code for additive model showing sensitivity to choice of basis
for the cubic-splines in Table 5.6#

Model 4a
>gam4a= gam(log(Mean + 0.5) = s(Month, bs = "cr")+s(Year, bs = "cr")+
s(Easting, Northing, bs = c("tp"),data=mydata)
>summary(gam4a)

#R-code for additive model showing sensitivity
 to the choice of basis for the p-splines in Table 5.7#

Model 4b
>gam4b=gam(log(Mean + 0.5)=s(Month, bs = "p",k=8)+s(Year, bs="p",k=8)+
s(Easting, Northing, bs = c("tp"),data=mydata)
>summary(gam4b)
```

```
#R-code for additive model showing sensitivity to choice of basis
dimension with k=6 and 12 for univariate and bivariate smoothers
respectively in Table 5.8#

Model 4c
>gam4c= gam(log(Mean + 0.5) ~ s(Month, k = 6) + s(Year, k = 6)
+ s(Easting, Northing, k = 12),data=mydata)
>summary(gam4c)

#R-code for additive model showing sensitivity to choice of basis
dimension k=12 and 20 for univariate and bivariate smoothers
respectively in Table 5.9#

Model 4d
>gam4d = gam(log(Mean + 0.5) =s(Month, k = 12) + s(Year, k = 12)+
s(Easting, Northing, k = 20),data=mydata)
>summary(gam4d)

#R-code for additive model showing sensitivity to the REML method
for the smoothing parameter estimation in Table 5.10#

Model 4e
>gam4e = gam(log(Mean + 0.5)=s(Month) +s(Year)+s(Easting,Northing),
method="REML",data=mydata)
>summary(gam4e)

A.4.6: # R-code for model summary in Table 5.11 and Figure 5.12#
#R-sq, Dev. exp., and GCV/ REML parameters are obtained from
Tables 5.1, 5.2, and 5.6-5.10#

>logLik(gam1, gam2, gam4a, gam4b, gam4c, gam4d, gam4e)
>AIC(gam1, gam2, gam4a, gam4b, gam4c, gam4d, gam4e)



>par(mfrow=c(3,2))
>plot(gam4a, select=1,main="basis=cr",xlab="Month")
>plot(gam4a, select=2,main="basis=cr",xlab="Year")
>plot(gam4b, select=1,main="basis=p",xlab="Month")
```

```
>plot(gam4b, select=2,main="basis=p",xlab="Year")
>plot(gam4c, select=1,main="k=6",xlab="Month")
>plot(gam4c, select=2,main="k=6",xlab="Year")
>par(mfrow=c(2,2))
>plot(gam4d, select=1,main="k=12",xlab="Month")
>plot(gam4d, select=2,main="k=12",xlab="Year")
>plot(gam4e, select=1,method=REML",xlab="Month")
>plot(gam4e, select=2,main="method=REML",xlab="Year")
```

A.4.7: #R-code for model validation for reduced model in Table 5.12#

Model 4f

```
>gam4f=log(Mean+0.5)=s(Month)+s(Year)+ s(Easting, Northing),data=mydat1)
> summary(gam4f)
```

# References


1 http://www.airquality.co.uk/data

2 http://www.naei.org.uk

3 http://www.npemap.org.uk/api/geocodes.shtml

4 http://www.geopostcodes.com/index.php

5 http://www.freemaptools.com/download-uk-postcode-lat-lng.htm

6 http://www.postcodefinder.org.uk/web/Postcode/Scottish

7 https://www.ordnancesurvey.co.uk/opendatadownload/products.html

8 Abramson, M. and Voigt, T. (1991). Ambient air pollution and respiratory disease. *Medical Journal of Australia*, $154, 543 - 553$.

9 Armstrong, M. and Jabin, R. (1981). Variogram models must be positive-definite. *Journal of Mathematical Geology*, $13, 455 - 459$.

10 Azadeh, A. and Salibian-Barrera, M. (2009). An outlier-robust fit for generalised additive models with applications to outbreak detection. *Manuscript, available at the second authors website.*

11 Bacarelli, A., Artinelli, I., Zanobetti, A. Grillo, P. et al. (2008). Exposure to Particulate Air Pollution and Risk of Deep Vein Thrombosis. *archives of Internal Medicine*, $168(9), 920 - 927$.

12 Baumbach, G. (1996). Air Quality Control; Formation and Sources, Dispersion, Characteristics and Impact of Air Pollutants. *Springer-Verlag*, Berlin, Germany.

13 Beeson, W.L., Abbey, D.E. and Knutsen, S.F. (1998). Long-term concentrations of ambient air pollutants and incident lung cancer in California adults, results from the AHSMOG study. *Environmental Health Perspectives*, $106(12), 813 - 823$.

14 Biggeri, A., Bellini, P.A. and Terracini, B. (2004). Meta-analysis of the italian studies on short term effects of air pollution 1996-2002. *Epidemiologia Prevenzione*, $(4-5Suppl), 4-100$.

15 Biggeri, A, Baccini, M., Bellini P.A. and Terracini, B. (2005). Meta-analysis of the Italian Studies of short term effects of air pollution (MISA) 1990-1999. *International Journal of Occupational and Environmental Health*, $11, 107-122$.

16 Biggeri, A., Baccini, M., Lagazio, C. and Saez, M. (2007). Parametric and semi-parametric approaches in the analysis of short-term effects of air pollution on health. *Journal of Computational Statistics and Data analysis*, $4324-4336$.

17 Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (1994). Time Series Analysis, Forecasting and Control, 3rd ed. Prentice Hall, Englewood Cliffs, NJ.

18 Bowman and azzalini (2002). Computational aspects of nonparametric smoothing with illustrations from the sm library. *Computational Statistics and Data Analysis*, $42, 545-560$.

19 Bowman, A., Giannitrapani, M., Scott, M. and Smith, R. (2009). Spatiotemporal smoothing and sulphur dioxide trends over Europe. *Journal of the Royal Statistical Society, c*, $58, 737-752$.

20 Brauer, M., Hoek, G., Van vliet, P., Meliefste, K., Fischer, P. (2003). Estimating long-term average particulate air pollution concentrations; application of traffic indicators and geographic information systems. *Epidemiology*, $14, 228-239$.

21 Brunekreef, B. and Holgate, S.T. (2002). Air pollution and health. *Lancet*, $360, 1233-1242$.

22 Burnett, R.T., Dale, R.E., Krewski, D., Vincent, R., Dann, T. and Brook, J.R. (1995). Associations between ambient particulate sulphate and admissions to Ontario hospitals for cardiac and respiratory disease. *American Journal of Epidemiology*, $142, 15-22$.

23 Cressie, N. (1993). Statistics for Spatial Data. *John Wiley & Sons*, New York.

24 Chiles, J. P., P. Delfiner, P. (1999). Geostatististics, Modelling Spatial Uncertainty. *Wiley & Sons*, New York.

25 Diggle, P.J. and Ribeiro Jr, P.J (2000). geoR: A package for geostatistical analysis. *R News*, $1(2), 15 - 18$.

26 Diggle, P. J. and Ribeiro Jr, P.J.(2002). Bayesian inference in Gaussian model-based geostatistics.*Geograpical Environmental Model*. $6, 129 - 146$.

27 Diggle, P.J., Knorr-Held, L., Rowlingson, B., Su, T., Hawtin, P. and Bryant, T. (2003). Towards on-line spatial surveillance. In *Monitoring the Health of Populations: Statistical Methods for Public Health Surveillance*, Brookmeyer, R. and Stroup, D., Oxford University Press. University Press.

28 De Oliveira, V., Benjamin, K., David, A. (1997). Bayesian Prediction of Transformed Gaussian Random Fields *Journal of the American Statistical Association* , $92(440), 1422 - 1433$.

29 Dockery, D.W., Pope, C.A., Xu, X., Spengler, J.D., Ware, J.H., Fay, M.E., Ferris, B.G. and Speizer, F.E. (1993). An association between air pollution and mortality in six U.S. cities. *The New England Journal of Medicine*, $329, 1753 - 1759$.

30 Dockery, D.W., Cunningham, J., Damokosh, A.I., Neas, L.M., Spengler, J.D., Koutrakis, P., Ware, J.H., Raizenne, M., and Speizer, F.E. (1996). Health effects of acid aerosols on North American children: respiratory symptoms. *Environmental Health Perspectives*, $104, 500 - 505$.

31 Dockery, D.W., Laden, F., Schwartz, J. and Speizer, F.E. (2006). Reduction in fine particulate air pollution and mortality; extended follow-up of the Harvard six cities study. *American Journal of Respiratory and Critical Care Medicine*, $173, 667 - 672$.

32 Eilers, P.H.C. and B.D. Marx (1996). Flexible Smoothing with B-splines and Penalties. *Statistical Science*, $11(2), 89 - 121$.

33 Fan, S., Burstyn, I. and Senthilselvan, A. (2008). Spatiotemporal modelling of ambient sulphur dioxide concentrations in rural Western Canada. *Environmental Modelling and Assessment*, $15(2), 137 - 146$.

34 Faraway, J.J. (2006). Extending the Linear Model with R; Generalized Linear, Mixed Effects and Non-parametric Regression Models. *Chapman & Hall / CRC*, Boca Raton.

35 Fasso, A. and Cameletti, M. (2007). A general spatio-temporal model for environmental data. *Research Group for Statistical Applications to Environmental problems (GRASPA) Working paper*, n(27).

36 Fuentes, M., Song HR, Ghosh, S., Holland, D. and Davis, J. (2006). Spatial association between speciated fine particles and mortality. *Biometrics*, $62(3), 855 - 863$.

37 Handcock, M.S. and Stein, M.L. (1993). Bayesian analysis of kriging. *Technometrics*, $35, 403 - 410$.

38 Handcock, M.S. and Wallis, J.R. (1994). An approach to Statistical spatial-temporal modelling of meteorological fields. *Journal of the American Statistical Association*, $89, 368 - 378$.

39 Harville, J. and Jeske, D.R. (1992). Mean squared error of estimation or prediction under a general linear model. *JASA*, $87, 724 - 731$.

40 Hastie, T.J. and Tibshirani, R.J. (1986). Generalized Additive Models. *Statistical science*, $1(3), 293 - 318$.

41 Hastie, T.J. and Tibshirani, R.J. (1990). Generalized Additive Models. *Chapman & Hall*, New York. Hesterberg, T. (1999). A Graphical Representation of Little's Test for MCAR. *MathSoft Research Report*, 94

42 Hobert, J.P., Altman, N.S. and Schofield, C.L. (1997). Analysis of fish species richness with spatial covariates. *Journal of the American Statistical Association*, $92, 846 - 854$.

43 Holland, D.M., De Oliveira, V., Cox, L.H. and Smith, R.L. (2000). Estimation of regional trends in sulphur dioxide over the Eastern United States. *Environmetrics*, $11, 373 - 393$.

44 Horton, N.J. and Lipsitz S.R. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *The American Statistician*, $244 - 254$.

45 Host, G. (1999). Bayesian estimation of European sulphur emissions using monitoring data and an acid deposition model. *Environmental and Ecological Statistics*, $6(4), 381 - 399$. of regional trends in sulphur dioxide over the Eastern United States. *Environmetrics*, $11, 373 - 393$.

46 Huerta, G., Sanso, B. and Stround, J.R. (2004). A spatio-temporal model for Mexico City ozone levels. *Applied Statistics*, 53(2), 231 – 248.

47 Jackson, L.S., Carslaw, D.C. and Emmerson, K.M. (2008). Modelling trends in OH radical concentrations using generalized additive models. *Atmospheric Chemistry Physics Discusson*, 8, 14607 – 14642.

48 Jesper Moller, and Murray, J.D. (2003). Spatial Statistics and Computational Methods. *Springer Verlag*, New York.

49 Kammann, E. and Wand, M.P. (2003). Geoadditive models. *Applied Statistics*, 52(1), 1 – 18.

50 Katsouyanni, K., Touloumi, G. and Schwartz, J. et al. (1997). Short term effects of ambient sulphur dioxide and particulate matter on mortality in 12 European cities: results from time series data from the APHEA project. *British Medical Journal*, 314, 1658.

51 Kauermann, Opsomer (2000). Local Likelihood Estimation in Generalized Additive Models. *Sonder for schungsbereich*, 386, 190.

52 Koren, H. S. (1995). Associations between criteria air pollutants and asthma. *Environmental Health Perspectives*, 103, 235 – 242.

53 Kyriakidis, P.C. and Journel, A.G. (1999). Geostatistical space-time models. *A review. Mathematical Geology*, 31, 651 – 684.

54 Li, K.H., Le, N.D., Sun, L. and Zidek, J.V. (1999). Spatial-temporal models for ambient hourly $PM_{10}$ in Vancouver. *Environmetrics*, 10, 321 – 338.

55 Little, R.J.A. and Rubin, D.B., (1987) Statistical Analysis with Missing Data. *Wiley*.

56 Little, R. J. A. and Rubin, D. B. (2002). Statistical Analysis with Missing Data. *Wiley*, 2nd. Edition, Hoboken, NJ.

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

58 Longley and Batty 1997. Spatial Analysis Modelling in a GIS Environment. 24 – 25.

59 Lopez-Moreno, J. I. and Nogues-Bravo, D. (2005). A generalized additive model for the spatial distribution of snowpack in the Spanish Pyrenees. *Hydrological. Processes*, $19, 3167 - 3176$.

60 MacNab, Y.C. (2004). Bayesian spatial and ecological models for small-area accident and injury analysis. *Accident Analysis and Prevention*, $36, 1019 -$ 1028.

61 Martuzzi M, M.F., Iavarone, I. and Serinelli, M., (2006). Health Impact of $PM_{10}$ and Ozone in 13 Italian cities, WHO Regional Office for Europe.

62 Mentzakis, E. and Delfino, D. (2005). Effects of air pollution and weather parameters on human health in the city of Athens, Greece. *Proceedings of the* $10^{th}$ *Int. Conf. on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes*, 566.

63 Murase, H., Nagashima, H., Yonezaki, S., Matsukura, R., and Kitakado, T. (2009). Application of a generalized additive model (GAM) to reveal relationships between environmental factors and distributions of pelagic fish and krill: a case study in Sendai Bay, Japan. *ICES Journal of Marine Science*, $66, 1417 - 1424$.

64 OBrien, L., Rago, P., (1996). An application of the generalized additive model to groundfish survey data with Atlantic cod off the northeast coast of the United States as an example. *NAFO Science Council Study*, $28, 79 - 95$.

65 Olsson, J.S. and Oard, D.W. ( 2008). Combining Speech Retrieval Results with Generalized Additive Models. *Proceedings of Association for Computational Linguistics*, $461 - 469$.

66 Pengfei Li, (2005). Box-Cox Transformations: An Overview.

67 Pilz, J., Schimek, M.G. and Spock, G. (1997). Taking account of uncertainty in spatial covariance estimation. In *Geostatistics* Wollongong, Baafi, E. and Schoefield, V(1), Kluwer, Dordrecht, $402 - 413$.

68 Pilz, J., Pluch, P. and Spock, G. (2005). Bayesian kriging with lognormal data and uncertain variogram parameters. *Geostatistics for Environmental Applications*, $51 - 62$.

69 Pilz, J., and G. Spock (2007). Why do we need and how should we implement Bayesian kriging methods. *Stochastic Environmental Research and Risk Assessment*, $22(5), 621 - 632$.

70 Prasad, N.G.N, Rao, J.N.K. (1990). On the estimation of mean square error of small area predictors. *JASA*, $85, 163 - 171$.

71 R Development Core Team(2005). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria.

72 Riccio, A., Barone, G., Chianese, E. and Giunta, G. (2006). A hierarchical Bayesian approach to the spatio-temporal modelling of air quality data. *Journal of Atmospheric Environment*, $40, 554 - 566$.

73 Roca-Pardinas, J., Gonzalez-Manteiga, W., Febrero-Bande, M. et al. (2004). Predicting binary time series of $S0_2$ using generalized additive models with unknown link function. *Environmetrics*, $15, 729 - 742$.

74 Roca-Pardinas, J., Gonzalez-Manteiga, W., Febrero-Bande, M. et al.(2005). Testing for interactions in generalized additive models: Application to $S0_2$ pollution data. *Statistics and Computing*, $15, 289 - 299$.

75 Roderick J. A., (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association*, $83(404), 1198 - 1202$.

76 Romanowicz, R., Young, P., Brown, P. and Diggle, P. (2006). A recursive estimation approach to the spatio-temporal analysis and modelling of air quality data. *Journal of Environmental Modelling and Software*, $21, 759 -$ 769.

77 Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys, *John Wiley & Sons*, New York.

78 Ruppert, D., Wand, M.P. and Carroll, R.J. (2003). Semiparametric Regression. *Cambridge University Press*, London.

79 Sabah, A. and Saleh, M. (2008). Prediction of sulphur dioxide concentration levels from the Mina Al-Fahal refinery in Oman using artificial neural networks. *American Journal of Environmental Sciences*, $4(5), 473 - 481$.

80 Samet, J.M., Dominici, F., Curriero, F.C. and Zeiger, S.L (2000). Fine particulate and air pollution in 20 US cities, 1987-1994. *New England Journal of Medicine*, $343, 1742 - 1749$.

81 Schwartz, J. (1991). Particulate air pollution and daily mortality; a synthesis. *Public Health Review*, $19(4), 39 - 60$.

82 Schwartz, J., Samoli, E., Wojtyniak, B., Touloumi, G., Spix, C., Balducci, F., Medina, S., Rossi, G., Sunyer, J., Bacharova, L., Anderson, H.R. and Katsouyanni, K. (2001). Investigating regional differences in short-term effects of air pollution on daily mortality in the APHEA project; a sensitivity analysis for controlling long-term trends and seasonality. *Environmental Health Perspective*, $109(4), 349 - 353$.

83 Skalpe, I.O. (1964). Long-term effects of sulphur dioxide exposure in pulp mills. *British Journal of Industrial Medicine*, $21(1), 69 - 73$.

84 Smith, R.L., Kolenikov, S. and Cox, L.H. (2003). Spatio temporal modelling of $PM_{2.5}$ data with missing values. *Journal of Geophysical Research*, $108(24), 9004$.

85 Software Documentation for SPSS 17.0.

86 Stein, A., Kassteele, J.V., Dekkers, A.L.M. and Velders, G.J.M. (2006). Statistical air quality mapping. *Doctoral Thesis of Wageningen University.*

87 Sunyer, J., Castellsague, J., Saez, M., Tobias, A. and Anto, J.M. (1996). Air pollution and mortality in Barcelona. *Journal of Epidemiology and Community Health*, $50(1), S76 - S80$.

88 Terzi, Y. and Cengiz, M. (2009). Using of generalized additive model for model selection in multiple poisson regression for air pollution data. *Scientific Research and Essay*, $4(9), 867 - 871$.

89 Touloumi, G., Pocock, S.J., Katsouyanni, K. and Trichopoulos, D. (1994). Short-term effects of air pollution on daily mortality in Athens; a time-series analysis. *International Journal of Epidemiology*, $23, 957 - 967$.

90 Wahba (1990). Spline Models of Observational Data. *SIAM.*

91 Waller, L.A. (2005). Bayesian Thinking in Spatial Statistics. *Handbook of Statistics.*

92 Wang Shizhen and Morishima Gary (2009). The Use of Generalized Additive Models for Forecasting the Abundance of Queets River Coho Salmon. *North American Journal of Fisheries Management*, $29, 423 - 433$.

93 Wikle, C.K., Berliner, L.M., and Cressie, N. (1998). Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics*, $5, 117 - 154$.

94 Wikle, C.K., Royle, J.A. (2002). Spatial statistical modeling in biology. *Encyclopedia of Life Support Systems*, Oxford.

95 Wood, S.N. (2003). Thin plate regression splines. *Journal of Royal Statistics Society, B*, $65(1), 95 - 114$.

96 Wood S.N. and Nicole H. Augustin, N.H (2002). GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling*, $157, 157 - 177$.

97 Wood, S.N. (2006). Generalized Additive Models; an Introduction with R. *Chapman & Hall / CRC*, Boca Raton.

98 World Health Organisation (1992). Acute effects on health of smog episodes. *European Series*, $43, 1 - 74$.

99 Xia, H. and Carlin, B.P. (1998). Spatio-temporal models with errors in covariates; mapping Ohio lung cancer mortality. *Statistics in Medicine*, $17, 2025 - 2043$.

100 Yanosky, J.D., Paciorek, C.J., Schwartz, J., Laden, F., Puett, R. and Suh, H.H. (2008). Spatio-temporal modeling of chronic $PM_{10}$ exposure for the Nurses' Health study. *Atmospheric Environment*, $42, 4047 - 4062$.

101 Yap, C., Robertson, C, Beverland, I., Hole, D.J., Angius, R., Heal, M.R., Henderson, D.E., Cohen, G. and Morris, G. (2006). Estimating long-term exposures to air pollution in Scotland. *Epidemiology*, $17(6), S240$.

102 Yee, T. and Mitchell, N. (1991). Generalized additive models in plant ecology. *Journal of Vegetation Science*, $2, 587 - 602$.

103 Zimmerman, D.L., Cressie, N. (1992). Mean squared porediction error in the spatial linear model with estimated covariance parameters. *Annal of the Institute of Statistical Mathematics*, $44, 27 - 43$.