Nonparametric methodologies for regression models with correlated data

Marco Giannitrapani

A Dissertation Submitted to the
University of Glasgow
for the degree of
Doctor of Philosophy

Department of Statistics

January 2006

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Abstract

In many spatial and temporal data sets, nonparametric techniques have recently been widely used because of their ability to model without requiring any assumptions on the distributional form of the data. However many nonparametric tools assume independent errors, that is not always the case. The present work extends some of the well established nonparametric techniques in order to make them applicable even with correlated data. Simulation studies will show the performances of the proposed methodologies. The methods are applied to air pollution data monitored over Europe in last quarter of the twentieth century by EMEP (Co-operative Programme for Monitoring and Evaluation of the long Range Transmission of Air Pollutants in Europe) and by OECD (Organization for Economic Co-operation and Development).

Chapter 1 gives a background to the air pollution problems, introduces the questions of interest and the aims of this work. It also shows some characteristics of the data that will be necessary to take into account for the analysis that will be done in the following chapters.

Chapter 2 reviews some of the existing nonparametric methodologies that, however relying on the assumption of independent errors, could be applied to the data.

Chapter 3 presents a diagnostic to detect discontinuities in a one-dimensional nonparametric regression accounting for correlated errors. A simulation study shows the performance of the proposed test, and the results of its application to air pollution data (SO_2, SO_4 in air and SO_4 in precipitation) monitored across 130 sites in Europe from 1970's to 2000, will be presented.

Chapter 4 presents the generalization of well established nonparametric techniques that can model and test correlated data. Simulation studies show the performances of the proposed modeling tools.

Chapter 5 shows applications of the methodologies presented in Chapter 4 to air pollution data.

Chapter 6 develops binned versions of the methodologies introduced in Chapter 4 allowing to fit and test models with large data sets, such as spatiotemporal ones, that show correlation.

Chapter 7 presents an analysis of the relationship between the SO_2 emissions and the monitored SO_2 concentrations.

Chapter 8 will summarize the main conclusions with a final discussion on possible future work.

Acknowledgements

My first thanks goes to my supervisors Adrian Bowman, Marian Scott and Ron Smith, without whose help this work would not have been possible. Their teaching gave me the chance to live an extremely exciting learning experience. Their professional and human support has been been an everyday vital boost for my research. I acknowledge the financial support from the Chancellors fund of the University of Glasgow, and the Centre for Ecology and Hydrology in Edinburgh.

A great thanks to all stuff and students in the department of statistics at the University of Glasgow that made me feeling at home since the first day I walked in. I thanks to all people I met in Glasgow along these unforgettable years. Thanks to all the Scottish people to make this country one of the most beautiful and friendly place on Earth! Thanks to all the international friends I had the luck to meet in this fantastic city. Thanks to everyone who gave me a sweet smile, a nice chat...a cold pint ;)! A particular thanks goes to my flatmate Chris to have been my "brother in Glasgow", a unique sincere support in every moment.

I finally dedicate this work to my Dad, my Mum, my Sister, my Grandmas and my Lord, whose constant presence in my heart is making this Life the most beautiful Life I could ever imagine to have! THANKS!

A mio Papa', mia Mamma, mia Sorella, le mie Nonne e il mio Signore, perche' la vostra presenza nel mio cuore mi sta' regalando la piu' bella vita che mai avrei immaginato di poter avere.

Contents

List of Tables

List of Figures

Chapter 1

Aim, Background & Exploratory Analysis

1.1 Aim of this work

All data present in nature have a spatial and/or a temporal structure. Several methodologies are presented in the literature concerning spatial analysis and time series data, but few of them are able to deal with spatiotemporal data where these two studies are combined. Most of the spatiotemporal analysis methods belong to the parametric approach which although powerful, is usually based on distributional assumptions that are difficult to justify in nature. The objective of this thesis is to propose a flexible methodology that is able to analyze spatiotemporal data using nonparametric techniques that do not make distributional assumptions on the data. In particular we want to build nonparametric model fitting and testing techniques that allow us to:

- include in the model an undefined number of covariates,

- include nonlinear and nonmonotonic relationships between the response and the covariates,
- analyze separately the effect of each covariate over the response,
- produce estimates and standard error bands, that account for (spatial and/or temporal) correlation of the data,
- detect abrupt changes in trend (discontinuities),
- test models or single covariates accounting for correlation,
- deal with large data sets, as spatiotemporal ones often are.

Methodological developments of such techniques will be shown in the following chapters along with simulation studies that will prove their benefits. Applications to air pollution data (mainly sulphur compounds) monitored in Europe from 1970's to 2000 will show some interesting answers to the scientific questions of interest. The remainder of this chapter will introduce the environmental background, some characteristics of the data, and some preliminary statistical analysis.

1.2 Background to the air quality problem

During the last quarter of the twentieth century, great importance has been attached to the condition of the atmosphere because of its connection with public health risks and sensitive ecosystems. Coordinated international monitoring of acidifying air pollution in Europe developed steadily from the 1950's, when the

Swedish monitoring network supervised by the Institute of Meteorology, Stockholm, was first extended to a number of other countries under the name of the European Air Chemistry Network (EACN). Around 100 sites were involved and by the 1960's observations indicated an expanding area of Europe subject to highly acidic precipitation (pH = 3-4). In 1976, in response to the observed acidification of a growing region of Europe, the Co-operative Program for the Monitoring and Evaluation of the Long Range Transmission of Air Pollutants in Europe (EMEP) was established. The programme included coordinated background measurements of acidifying air pollutants to be performed by countries themselves, with data assembled by the Chemical Coordinating Centre (CCC) of EMEP, hosted by the Norwegian Institute for Air Research (NILU). This programme has since served the technical requirements of international agreements under the 1979 Geneva Convention on Long Range Transboundary Air Pollution (CLRTAP); the first of these was the 1985 Protocol on the Reduction of Sulphur Emissions or their Transboundary Fluxes. The acidification programme itself has extended beyond sulphur to include oxidised and reduced nitrogen compounds in air and precipitation. In addition to acidification, ozone, VOC, and trace contaminant programmes have also been estabilished. In 1988 a Protocol Concerning the Control of Nitrogen Oxides or their Transboundary Fluxes was agreed which sought to stabilise emissions. In 1994, the Protocol on Further Reduction of Sulphur Emissions was negotiated. These protocols have stipulated that EMEP should oversee levels and depositions of relevant compounds across Europe. The latest agreement was signed by many countries in 1999 and was called the "multi-pollutant, multi-effect" protocol, because it related the problem of acidification to other photochemical problems. The environmental issue that is the focus of

this thesis is to assess whether the efforts that have been made during the last quarter of the twentieth century to reduce the emissions of pollutants, have led to real improvement in environmental quality and a real change in the acidifying environment. A key point is that emissions within a region may not represent the critical influences upon air quality in that region and so there has been interest in the correspondence between emission changes resulting from policy and the observed quality of the atmosphere.

1.3 The monitoring network and the data: some interesting characteristics

This work is based on the data collected as part of EMEP (Co-operative Programme for Monitoring and Evaluation of the long Range Transmission of Air Pollutants in Europe), and from the OECD program (Organization for Economic Co-operation and Development). The data that will be analyzed are the daily concentrations of SO_2, SO_4 in air, and SO_4 in precipitation from 130 European stations. A map of the locations of the EMEP sites in Europe is shown in Figure 1.1. Most of the time series have been downloaded from the EMEP web site (`www.emep.int`), with recorded data starting from the late 1970's. However for some of the other European sites, more data are available from the OECD program, which was started in the early 1970's. The data have been recorded daily, and exploratory analysis of these values showed immediately some interesting features. Some examples are reported in the following graphs (Figure 1.2, Fig 1.3, Fig 1.4).

Figure 1.1: Location of the EMEP sites in Europe.

Characteristics that immediately appear are: the presence of some outliers, the presence of shifts in level, the presence of missing values, the skewness of the distribution.

The presence of outliers was unexpected, since the data are quality controlled firstly by each country, and secondly by the central body. However some of the values (as in Figure 1.3), are clearly too extreme to be included in the analysis,

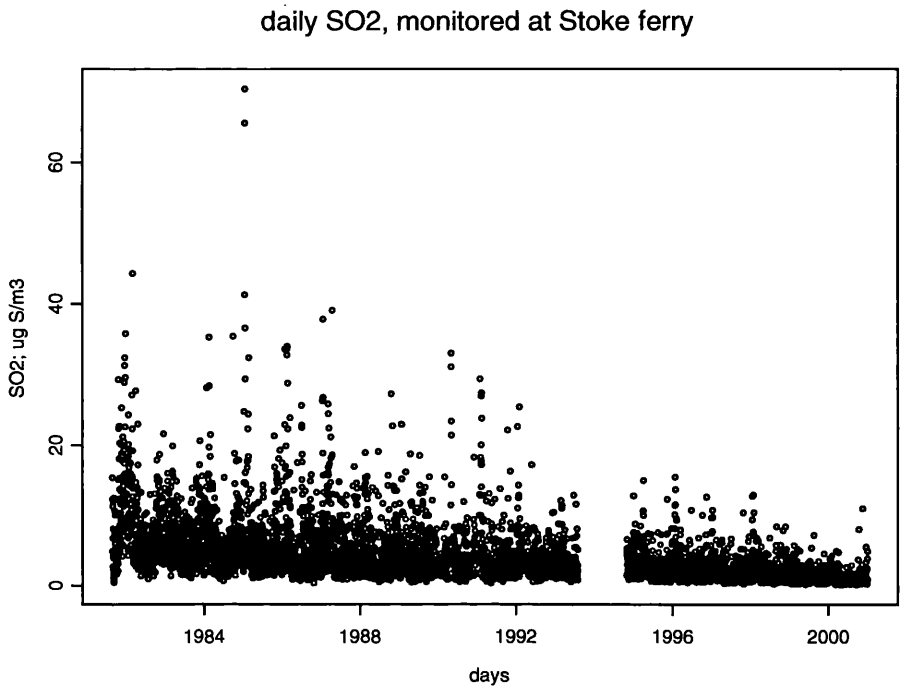

Figure 1.2: daily data for SO2 at Stoke Ferry (GB04)

making necessary their exclusion.

A further technical issue is the presence of shifts in level, clear from Figure 1.4, due mainly to the "detection limits". The "detection limits" are those levels below which the instrumentation is not able to measure concentrations accurately, and therefore the concentrations are reported as less than a constant fixed value. Over the years, the instrumentation has improved. Therefore, sometimes it is possible to observe a shift in values due to an improvement in the detection limits.

The presence of missing values, clearly seen in Figure 1.2 and Figure 1.3, has required further study.

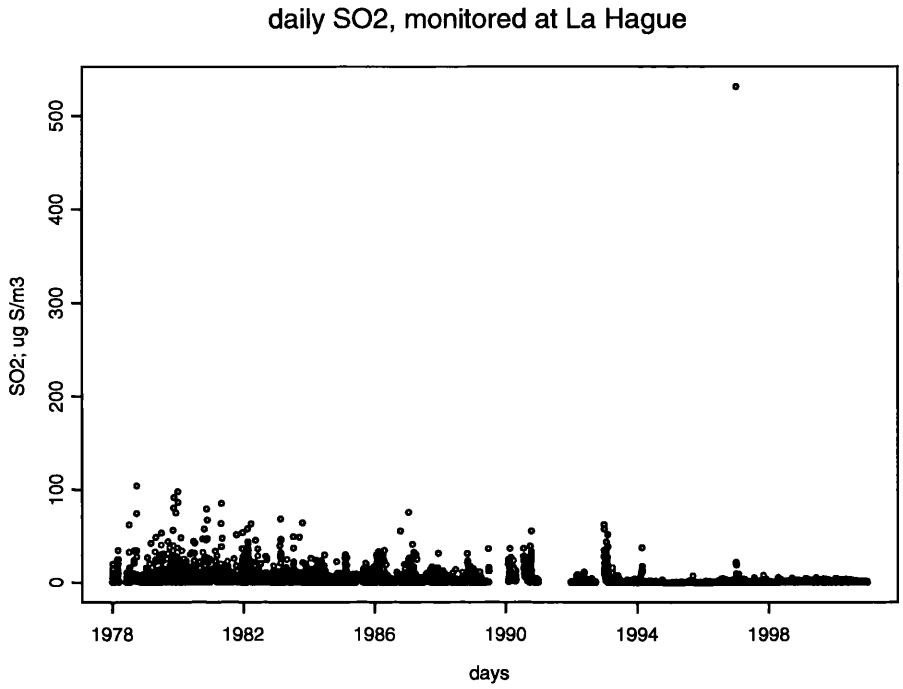

Figure 1.3: daily data for SO2 at La Hague (FR05)

- Are there compounds that have a significant number of missing values?
- Are there stations that have a significant number of missing values?
- Are the missing values uniformly spread across the observed period, or are they present as "missing blocks" (months, or even years)?

To have a clearer idea of the amount of missing data for each compound at each station, a bar graph for the counts of missing (Figure 1.5) and the percentages of missing values (Figure 1.6) for some stations are shown. Looking at these Figures, it is clear that a huge number of missing values affect SO_4 in precipitation. This ranges from 45% for the British station (GB02), to 90% for the Austrian station

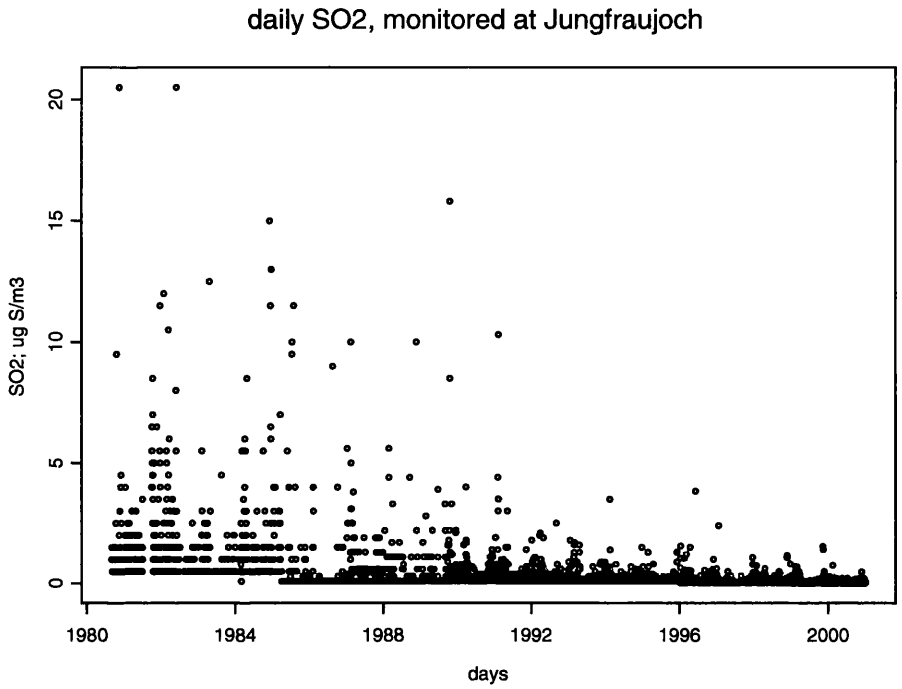

Figure 1.4: daily data for SO2 at Jungfraujoch (CH01)

(AT02). For SO_2 and SO_4 in air, except for the Austrian (AT02), the Finnish (FI04), and the two French (FR03, FR05) stations, the proportion of missing data is below 10%. The focus of this thesis will therefore be on SO_2 in air.

An analysis of the location and the structure of the missing values for each compound at each site has been performed by plotting graphs such as Figure 1.7, Figure 1.8 and Figure 1.9. The missing (1) and the observed (0) values, and a smooth curve, obtained by local linear regression, have been plotted. These three figures clearly show the different situations at different sites. For example, the Danish site shows quite a few missing observations that are spread across the whole monitoring period (Figure 1.7). The Finnish one presents instead a

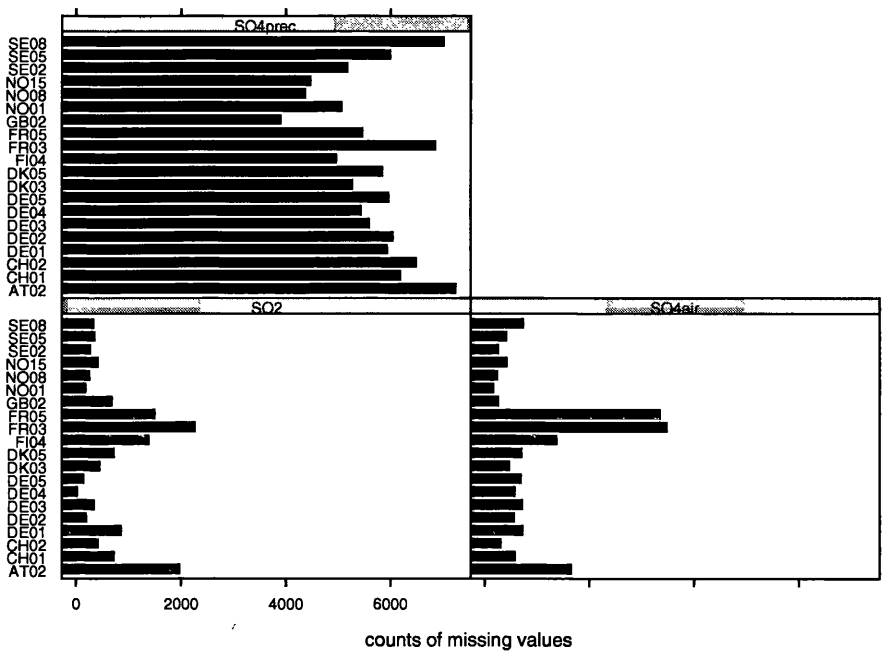

Figure 1.5: counts of missing daily data across compounds/sites

block of missing observations (Figure 1.8), while the French one presents not only blocks but also a constant amount of missing values (Figure 1.9).

This graphical analysis has been supported by a numerical one. For each compound at each station a matrix was constructed whose rows corresponds to the years that showed missing values, and whose columns corresponds to the months. The cells of those months that have more than 15 missing values, have been filled with the number of missing days. In this way it has been possible to show which months have only a very small number of observations. Examples of these data are given in Table 1.1 and Table 1.2. It is possible to note immediately that for the Danish site (Tange), very few months have blocks of missing data,

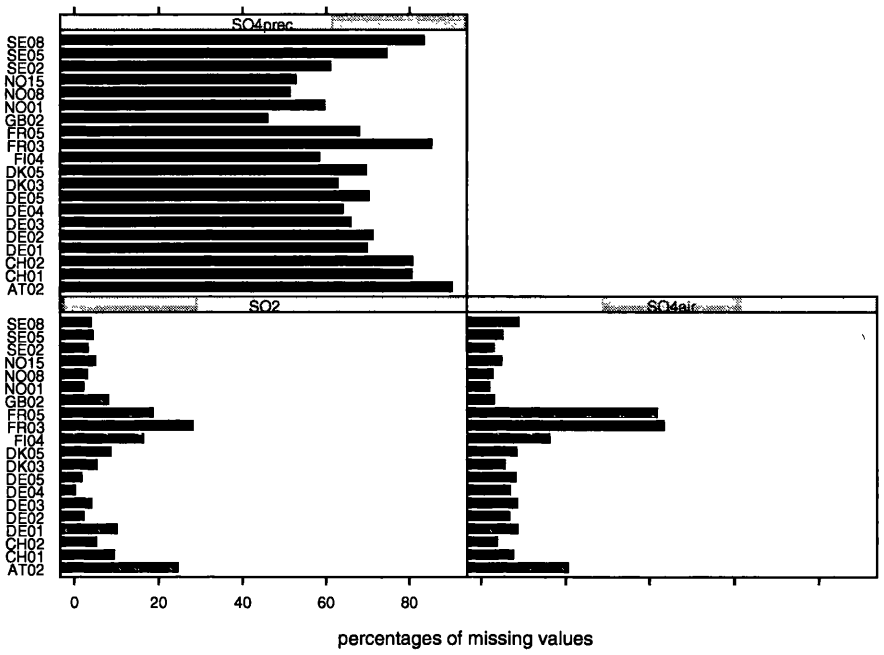

Figure 1.6: percentages of missing daily data across compounds/sites

while for the French site (La Crouzille), quite a few months have more than 15 missing values.

The results of these graphical and numerical analyses confirm the impression from the first two bar plots (Figure 1.5, Figure 1.6), that SO_4 in precipitation, exhibits a huge amount of missing values and so it will not be analyzed further. For SO_2 and SO_4 in air, interesting results for the missing values have been obtained. In fact, from the bar charts, the stations AT04, FI04, FR03 and FR05 were the only ones to have more than 10% missing values (for both SO_2 and SO_4 in air). However from the analysis of the plots of missing values over time and from the matrix of "Missing Blocks", it is clear that the Finnish station,

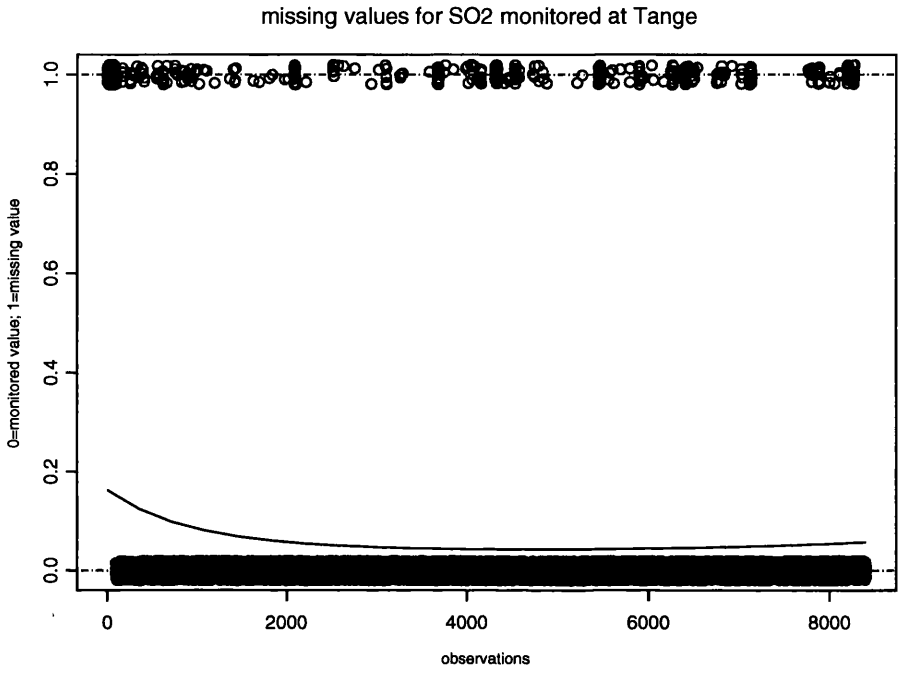

Figure 1.7: missing data for SO_2 at Tange (DK03)

Ähtäri (FI04), is a different case from the other three (AT04, FR03 and FR05). In fact its high percentage of missing values is due entirely to a missing period of three and half years (6/1997-12/2000). Apart from this, across its range of observations (10/1977-5/1997) there are very few missing observations. However, for the other three stations (AT04, FR03 and FR05), there is a high percentage of missing values across all its range of observations. All of these characteristics are extremely important for future stages of the study, and on the basis of these results it will be necessary to choose the appropriate methodologies of future analysis.

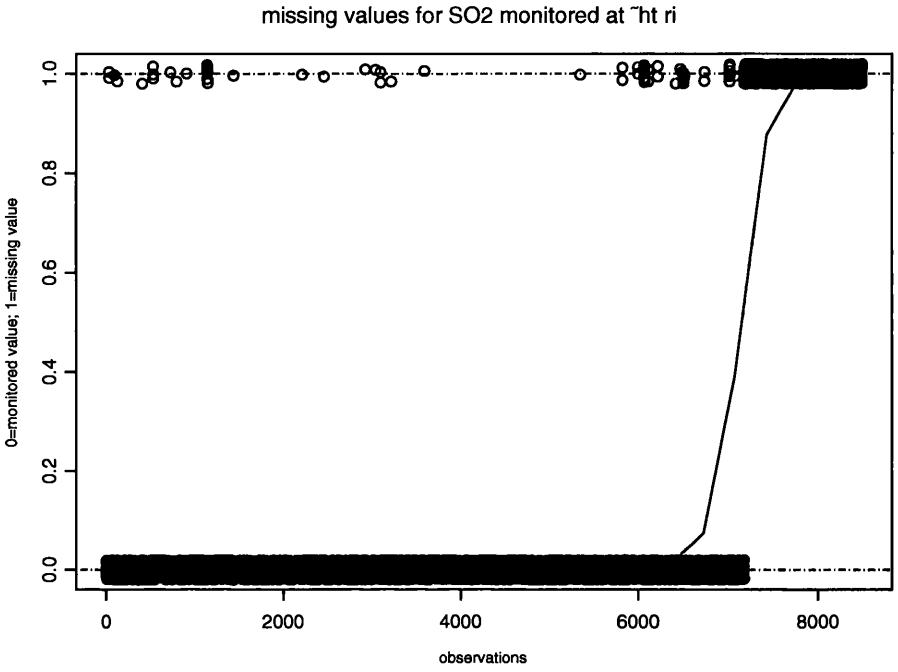

Figure 1.8: missing data for SO_2 at Ähtäri (FI04)

1.4 Exploratory Analysis of Trend and Seasonality

From the previous section it has been evident that the daily data are clearly skewed and show considerable variation, making the interpretation of the trend and seasonality difficult. The logarithm of the data over time was plotted to give a clearer idea of the main features of the data. It is necessary to point out that, because of the presence of some zero values in some data sets, a small positive offset equal to half of the minimum value of the series has been added to each data value before taking natural logs.

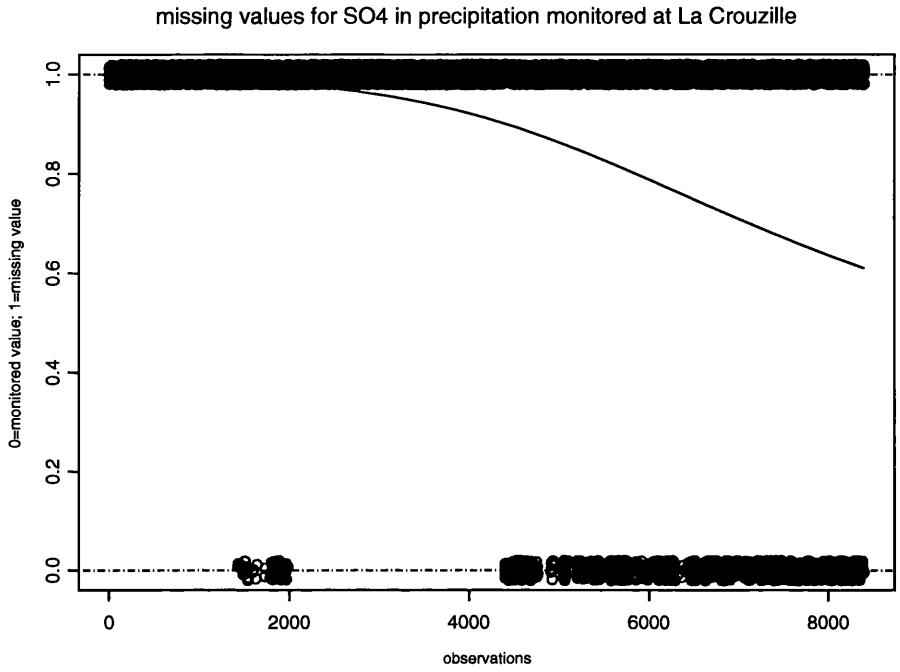

Figure 1.9: missing data for SO_4 in precipitation at La Crouzille (FR03)

For example Figure 1.10 a) presents the logarithm of the concentration over the entire period of observation. It is possible to note that the presence of straight lines are due to a detection limit problem. This graph also shows a clear downward trend, and peaks indicating the presence of seasonality. To look more carefully at the seasonal cycle, the detrended data over one year of observations have been plotted in Figure 1.10 b). The detrended data have been obtained by fitting a local linear regression, that can be thought of as a general kind of local moving average, and whose computation will be explained in more detail in the next chapter. From Figure 1.10 b) it is possible to note a seasonal yearly cycle, characterized by lower values in summer and higher values in winter.

Year \ Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1978	31	28	31									
1979												
1980												
1981												
1982												
1983									19			
1984												
1985												
1986												
1987												
1988												
1989												
1990												
1991												
1992												20
1993												
1994												
1995												
1996												
1997												
1998												
1999												
2000							16					

Table 1.1: Presence of Missing data in each month for SO_2 monitored at Tange.

Another kind of seasonality, characterized by shorter seasonal length was also studied. Indeed Figure 1.11 and Figure 1.12 show that a weekly cycle seems to characterize the data. This is possibly explained by the fact that factories are closed during the weekend, reducing the emissions of SO_2 on Saturday and Sunday. However this cycle, according to the chemistry, should not appear for SO_4 in air, in contrast with what has been noted in this analysis across most of the sites (i.e. Figure 1.12). It has been suggested by experts, that the presence

Year \ Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1979	31	28	31	30	31	30	31	31	30	31	30	31
1980	31	29	31	30	31	30	31	31	30	31	30	31
1981	31	28	31	30	31	30	31	31	30	31	30	21
1982	26	25	26	30	29	28	31	31	28	31	29	24
1983	28	24	27	26	25	29	31	31	30	31	30	31
1984	31	29	31	30	31	30	31	31	30	31	30	31
1985	31	28	31	30	31	30	31	31	30	31	30	31
1986	31	28	31	30	31	30	31	31	30	31	30	31
1987	31	28	31	30	31	30	31	31	30	31	30	31
1988	31	29	31	30	31	30	31	31	30	31	30	31
1989	31	28	31	30	31	30	31	31	30	31	30	31
1990	22		26		18	19	26	27	23	16	19	26
1991	28	28	31	30	31	23	28	31	30	29		30
1992	30	28	22	20	21		22	20	22			20
1993	24	27	26		18		17	23		17	29	24
1994	18	25	27	22	22	27	29	30	16	26	21	26
1995	20		28	29	31	28	29	28	17	21	23	29
1996	29	22	26	24	22	27	28	23	24	19		22
1997	25	18	30	30		17	23	23	25	18		18
1998	19	24	25		28	20	23	27			25	27
1999	19		22		18	25	22	20	16	19	23	17
2000	25	16	25	17	19	25	18	27	23	18	21	19

Table 1.2: Presence of Missing data in each month for SO_4 in precipitation monitored at La Crouzille.

of this cycle also for SO_4 in air may means that the filter that measures SO_4, also measures SO_2, possibly due to moisture.

In order to handle the effect of the short cycle and to reduce the variability that is affecting the data, the weekly means will be analyzed. However before computing the weekly means, it is necessary to keep in mind the problem of missing values that affect our data. In fact, because of the large number of missing values in the series, the computation of the weekly means based simply on the original daily values, would be highly biased. In some cases, the weekly

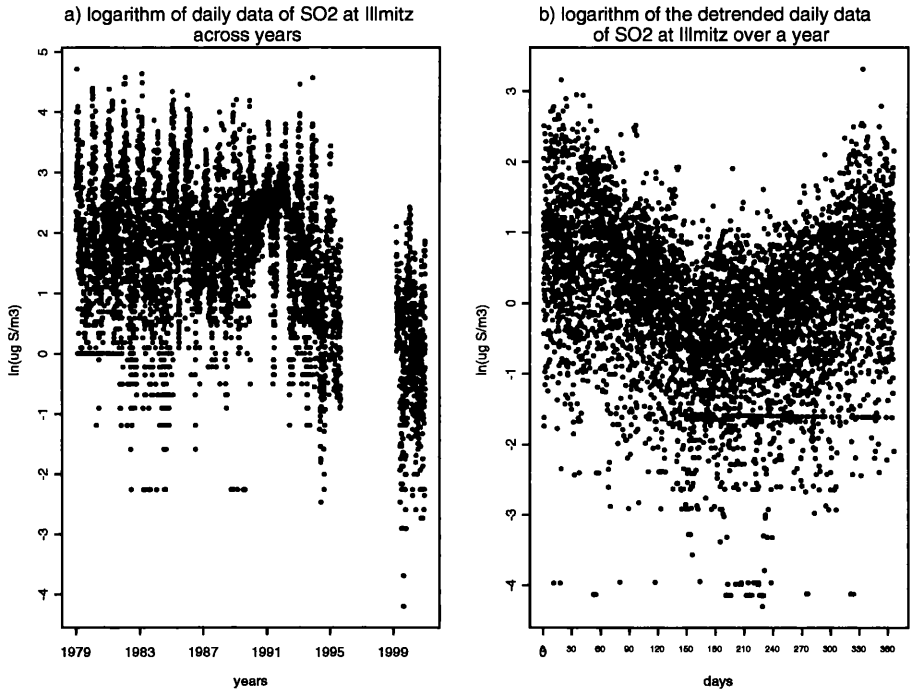

Figure 1.10: SO_2 at Illmitz (AT02): a) daily data across years, b) detrended data over one year

values would rely only on one or two days that could belong to the high or the low part of the cycle, skewing the results. So it was necessary first to remove the seasonality, and this has been done by applying a linear model, fitting the days of the week as factors, and then the de-seasonalised daily data have been used to obtain the weekly means. Figure 1.13 shows the steps that have been followed in computing the weekly means.

In particular, Figure 1.13 a) shows the natural logarithm of the daily data; Figure 1.13 b) shows the natural logarithm of the daily data after removing the "day within the week" seasonal component; Figure 1.13 c) shows the weekly means of the natural logarithm of the daily data after removing the "days within

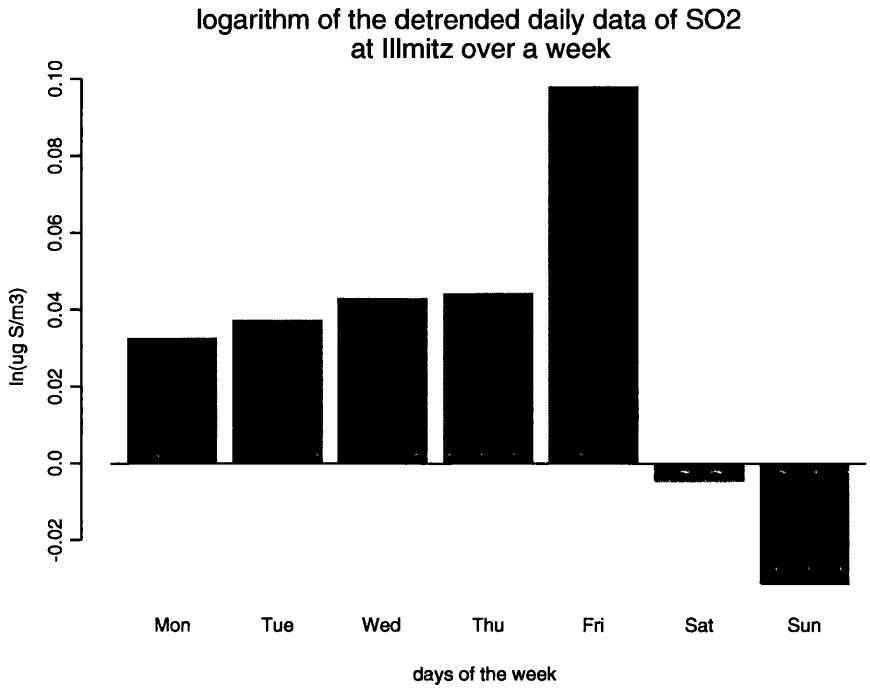

Figure 1.11: logarithm of the detrended daily data of SO_2 at Illmitz (AT02) over a week

the week" seasonal component; Figure 1.13 d) show the estimates of the "day within week" parameters of the linear model. The days of the week are considered as factors and their contrast matrix includes each level as a dummy variable, excluding the first one. So the 6 values that are present in the plot of Figure 1.13 d), represent the values for Tuesday, Wednesday, Thursday, Friday, Saturday and Sunday compared with Monday. From these plots, the presence of a daily seasonality is apparent, showing lower values during the weekend than in the rest of the week.

It is important to note that, for SO_4 in precipitation, the weighted weekly means are calculated where the weights are determined by the volume (mm) of

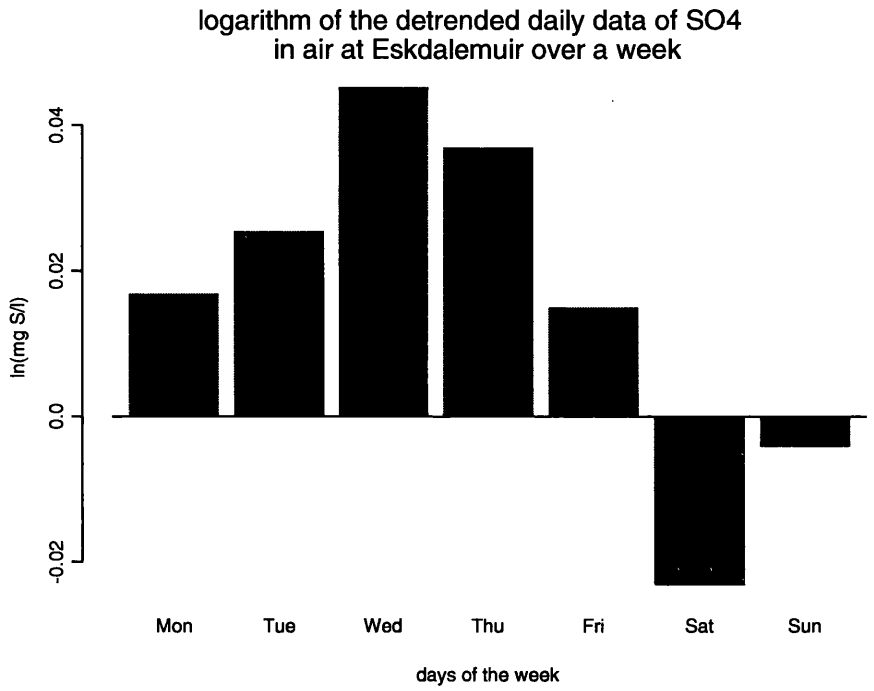

Figure 1.12: logarithm of the detrended daily data of SO_4 in air at Eskdalemuir (GB02) over a week

precipitation (for those sites where the data are present).

Figure 1.14 show an example of plotting the weekly means across years (Figure 1.14 a), and the detrended weekly data across weeks of the years (Figure 1.14 b).

Graphs like Figure 1.14 have been produced across all the sites, and most of them show that the yearly cycle is still present after removing the weekly cycle and the trend. From a quick look at the plots, it is clear that the weekly means of the logarithm of the data give a clearer pattern of the data, characterized by much less skewness and variability. In fact, taking the log first and the weekly means afterwards (Figure 1.14 a), it is possible to obtain data that are more interpretable. Therefore, most of the following analysis has been done using the

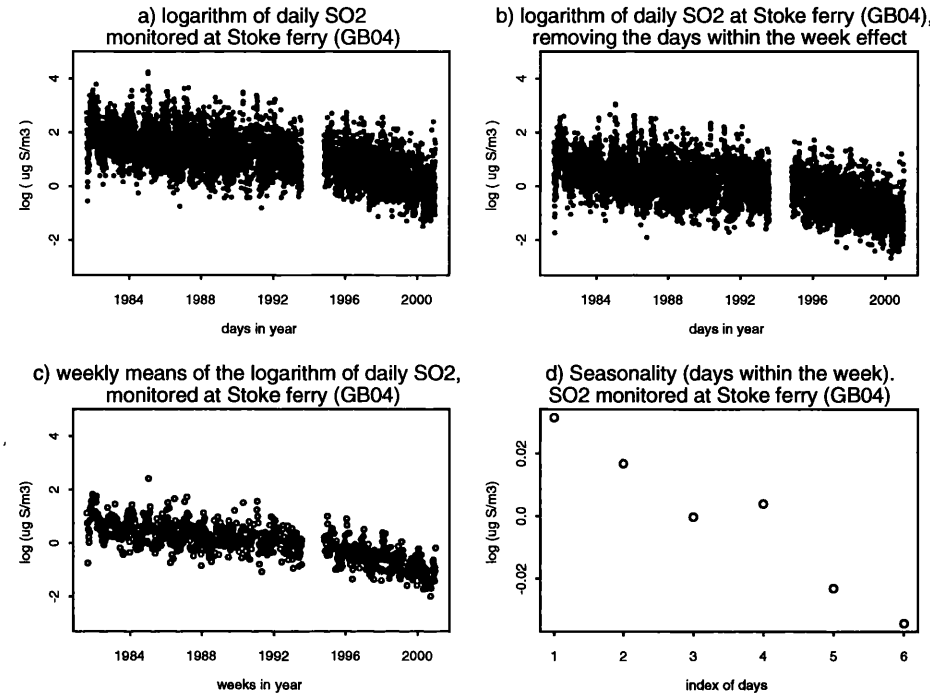

Figure 1.13: Analysis of seasonality for SO2 at Stoke Ferry (GB04). a) log of the daily data; b) deseasonalised log daily data; c) weekly means of the deseasonalised log daily data d)estimates of the "days within week" parameters (1=Tue, 2=Wed,...,6=Sun).

weekly means, rather than the daily ones.

Across the plots that have been produced, it has also been noted that the decreasing trends do not follow a linear pattern. In most cases, it seems more appropriate to model the trend by smooth curves. The following section will present some of the main approaches in analyzing trend.

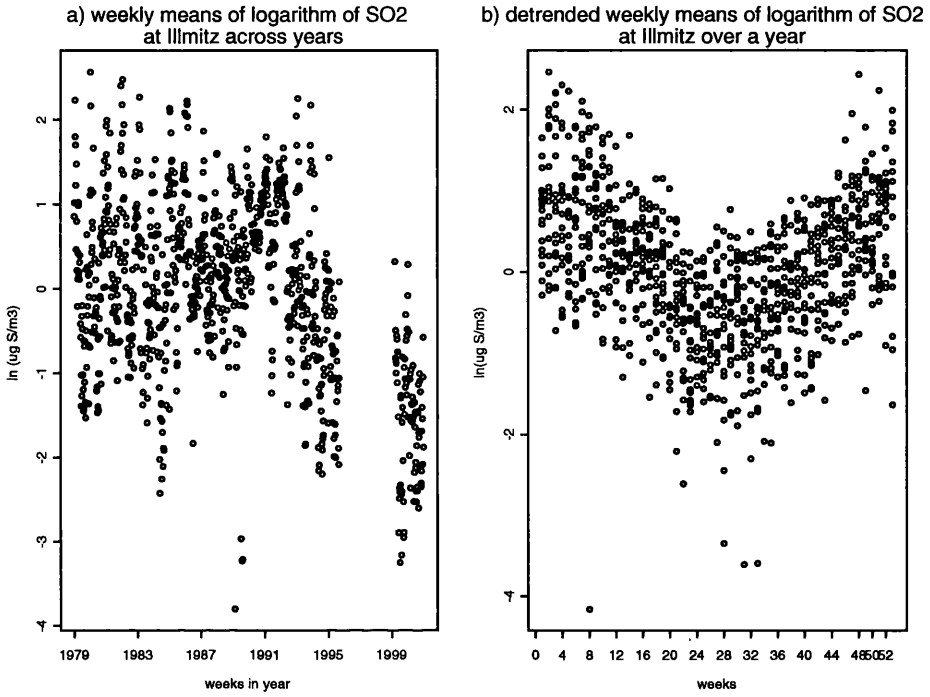

Figure 1.14: SO_2 at Illmitz (AT02): a) weekly log data across years, b) detrended weekly log data over one year

1.4.1 Literature review of trend analysis methodology

There are various ideas associated with the concept of trend and in particular different statistical approaches have been proposed for detecting and estimating trends. Trend analysis represents a huge area of study and the literature on this topic is vast. Even the definition of trend has involved many contributions. One of the most common is "long term systematic change in mean". That gives a good idea of what the trend is, but the interpretation of the words "long term" is subjective. So, a more appropriate definition is given by Granger (1966), who said that "trend in mean" comprises all the frequency components whose

wave lengths exceed the length of the observed series. Here some of the most important methodologies that are typically implemented in trend analysis are presented. These methodologies are classified into two areas: those that consider linear trends and those that relax this assumption of linearity.

1.4.1.1 Trend Analysis in Time Series: Linear Trends

An excellent review of linear trend analysis, was carried out by Hess et al. (2001), who compared seven methodologies for the analysis of linear trend. The methodologies described are:

- The Spearman Partial Rank Correlation (SPRC) to test for the presence of trend after correction for seasonality, proposed by McLeod et al. (1991).
- The Seasonal Kendall Test (SKT) proposed by Hess et al. (2001).
- The Seasonal Kendall Test for dependent data proposed by Hirsch and Slack (2002).
- Generalized Least Squares (GLS) with autoregressive errors, proposed by Reinsel and Tiao (2002), that can account for atmospheric variables, and that can combine the trend estimates from several stations, determining a general trend estimate.
- The Kolmogorov-Zurbenko (KZ) filter that Rao and Zurbenko (1994) have suggested in order to estimate trend. The KZ filter is an iterative moving average that separates trend from high frequency components.
- The t-test to assess the hypothesis that the difference between the means

of the first and the second half of observations is statistically significant, described by Hess et al. (2001).

- the deseasonalised t-test (described by Hess et al. (2001))

After presenting all these different procedures, Hess et al. (2001) presented some applications. Their results suggest that the deseasonalised t-test, and the SKT, should be preferred to the others. As well as being easily applicable, they are the ones that maintain the significance level and have high power with the variation of the trend considered.

Recently, Yue et al. (2002) investigated the interactions between a linear trend and an AR(1) process in a time series. They demonstrated through Monte Carlo simulation experiments that the serial correlation increases the size of the Mann-Kendall (MK) test statistic. Yue et al. (2002) also showed that serial correlation does not alter the central tendency or mean and the distribution of the MK statistic, but on the other hand, the presence of a trend affects the estimate of the serial correlation. Yue et al. (2002) also showed that a more accurate estimate of the true $AR(1)$, is obtained detrending the series prior to pre-whitening. They also proposed an alternative approach to detecting a significant trend in serially correlated series. This new approach comprises four steps (Yue et al., 2002):

1. The slope of the trend in the sample data is computed.

2. If the estimated slope differs significantly from zero, the identified trend is assumed to be linear and is removed from the sample data.

3. The lag-1 serial correlation coefficient of the detrended series is computed, and the AR(1) process is removed from the series.

4. The identified trend and the modified residual series are combined and the MK test is applied to this combined series to assess the significance of trend.

Weatherhead et al. (1998) analyzed the factors affecting the detection of trends with applications to environmental data. They showed how the number of years of data needed to detect a given trend is dependent on the autocorrelation and on the natural variation of the data (noise). It was also shown that the occurrence of level shifts in the data can add significantly to the uncertainty in trend estimates, thereby increasing the number of years necessary to detect a given trend.

1.4.1.2 Trend Analysis in Time Series: Non-Linear Trends

Methodologies that aim to relax the assumptions of linearity have also been developed. A wider view of methodologies has been proposed by Brillinger (1994), who presented some techniques for trend analysis in time series, classified as parametric, semi-parametric and nonparametric. Esterby (1993) compared assumptions of some of the trend analysis methods which have been used for environmental data, concluding that modeling seasonality provides a much more informative analysis than blocking on season. After presenting the Seasonal Kendall Test and least squares regression, Esterby (1993) shows that in some cases the seasonal variation may be poorly represented by sinusoidal terms and a general smoothing procedure may be more appropriate. Cleveland et al. (1990) presents a Seasonal Trend decomposition based on Loess (STL), that decomposes a time series into trend, seasonal and remainder components. STL consist of an inner loop nested inside an outer loop. In the inner loop, the seasonal and the trend components are estimated and updated at each loop. In the outer loop the robustness of the weights that will be used in the run of the following inner loop are computed.

Esterby (1993) also presented applications of locally weighted regression smoothing (loess), and of the related STL. Esterby (1993) has shown on the basis of water quality data monitored in the provinces of Alberta and Saskatchewan since the early 1950s that the STL algorithm can model gradual change from year to year of the seasonal component in a particular season, whereas, in the blocking methods, the seasonal component is assumed constant from year to year.

El-Shaarawi (1995) provides a summary of some test statistics that are often used for the detection of trends in environmental time series data. In the parametric area, he describes Rao's efficient score statistic to test simultaneously the presence of trend and correlation, requiring only computation of the Maximum Likelihood (ML) estimates under H_0. The Kendall's S-score test provides the equivalent test in the nonparametric area. The Kendall's S-score test is an extension of the Seasonal Kendall test that accounts for autocorrelation.

The effects of autocorrelation on estimates of trend have been the subject of many studies. The work of Tiao et al. (1990) discussed the characteristics of the data that affect the estimates of the time trends and of the spatial correlation. Tiao's principal findings were as follows.

1. Auto-correlations in the monthly observations affect critically the estimates of the time trends.

2. The temporal sampling rates of daily measurements under systematic sampling do not affect the estimates of the time trends and of spatial correlation between two neighboring stations.

3. The time lag between measurements taken at the two stations affects the estimate of spatial correlation between two neighboring stations.

An important tool in time series analysis is represented by Generalized Additive Models (GAMs), and several papers have been published on the use of GAMs. In particular, Hastie and Tibshirani (1987) used some applications to show that GAMs provide a flexible method for analyzing the effects of an undefined number of covariates in a variety of settings. Nonparametric estimates of the effects of the covariates can be used to suggest parametric transformations in order to perform usual linear analysis on the transformed variables. Hastie and Tibshirani (1987) pointed out that in literature there has also been a quite wide investigation and development of inferential tools for estimating and testing the covariates' effects in GAMs. In their work Hastie and Tibshirani (1987) used the local scoring algorithm to estimate the functions, using a scatterplot smoother as a building block. Hastie and Tibshirani (2000) also proposed a "general procedures for posterior sampling from additive and generalized additive models". They propose a Bayesian backfitting procedure that uses the ideas coming from Gibbs sampling and from the backfitting algorithm, in a way that at each step of the "backfitting algorithm", a noise effect is added to each component in order to obtain new realizations.

Berhane and Tibshirani (1998) proposed an extension of the GAMs that account for intrasubject correlation of longitudinal data. The fitting of these GAMs for longitudinal data is performed through the Local Quasiscoring Algorithm that is based on a multivariate form of quasilikelihood. This algorithm assumes fixed and user-supplied degrees of freedom that do not account for correlation. The testing performed by Berhane and Tibshirani (1998) is based on approximate model-based and empirical tests on the nonparametric contributions of the smooth terms that are not based on solid theoretical justification. Berhane and

Tibshirani (1998) conclude that formalization of inferential tools supported by simulation studies are needed.

Wood and Augustin (2002) presented theory and applications of GAMs with penalized regression splines, combining the idea of GAMs with Generalized Spline Smoothing (GSS) (Wahba, 1990). They noted how model selection techniques of GAMs can be improved using the GSS algorithm for estimating multiple smoothing parameters.

Some theory of penalized spline generalized additive models has also been given recently by Aerts et al. (2002) who derived simple closed form approximations to the degrees of freedom of the estimator and its components for ordinary additive models and for GAMs. Cleveland and Devlin (1988) instead investigated the use of locally weighted regression (LOESS) as a way of providing: exploratory information of the data, indication on the parametric form to model the data, and nonparametric estimates of the regression surface. Cleveland and Devlin (1988) stressed the importance of using nonlinear local procedures that can relax the assumptions of normality and constant variance of the errors that characterize most of the more traditional methodologies.

Dominici et al. (2002) discuss the use of GAMs for estimating relative rates of mortality associated with exposure to air pollution in time-series analysis of air pollution and health. Re-analyzing the data of the National Morbidity, Mortality, and Air Pollution Study (NMMAPS), Dominici et al. (2002) concluded that GAMs provide a more flexible approach for adjusting for nonlinear confounders compared with fully parametric alternatives. However the use of GAMs through statistical packages, such as Splus, requires considerable caution because the default parameters need to be more stringent.

All these methodologies that have been discussed are potential methods of modeling trend. However, with air pollution data, there is also a need to consider the effect of meteorology. The next section will present an exploratory analysis of some meteorological variables of interest. Some of the main trend analysis approaches that account for meteorology will also be discussed.

1.5 Exploratory Analysis of Meteorological Variables

The analysis of atmospheric pollutants must also consider, where possible, the effects of meteorology. Thus at many pollutant monitoring sites meteorology is measured too.

The meteorological variables that are commonly available and that could be of interest are: Temperature, in terms of Mean, Minimum and Maximum (degrees Celsius), Humidity (%), Precipitation (mm) and Wind Speed (knots), and Wind Direction (degrees). We have acquired meteorological data at 11 stations, namely at Eskdalemuir (GB02, Scotland), Westerland (DE01, Germany), Waldhof (DE02, Germany), Schauinsland (DE03, Germany), Deuselbach (DE04, Germany), Brotjacklriegel (DE05, Germany), Kosetice (CZ03, Czech Republic), Rörvik (SE02, Sweden), Bredkälen (SE05, Sweden), Hoburg (SE08, Sweden), Payerne (CH02, Switzerland). The meteorological data, collected hourly, have been aggregated to daily values first, and then weekly. It is clear that the computation of weekly values of temperature (mean, minimum and maximum), humidity, amount of rain and wind speed, is straightforward, but not for wind

direction. So two different kinds of mean for wind direction were computed. The first is the mean direction proposed by Mardia (1972). If P_i, $i = 1, \ldots, n$ are points on the circumference of the unit circle corresponding to the angles θ_i, $i = 1, \ldots, n$, then the mean direction $\bar{x}_0$ of $\theta_1, \ldots, \theta_n$ is defined to be the direction of the resultant of the unit vectors $\overrightarrow{OP}_1, \ldots, \overrightarrow{OP}_n$. The cartesian co-ordinates of P_i are $(\cos\theta_i, \sin\theta_i)$, $i = 1, \ldots, n$, so that the center of gravity of these points is $(\bar{C}, \bar{S})$ where

$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n}\cos\theta_i, \quad \bar{S} = \frac{1}{n}\sum_{i=1}^{n}\sin\theta_i \tag{1.1}$$

therefore, if

$$\bar{R} = (\bar{C}^2 + \bar{S}^2)^{\frac{1}{2}} \tag{1.2}$$

then $R = n\bar{R}$ is the length of the resultant and $\bar{x}_0$ is the solution of the equations

$$\bar{C} = \bar{R}\cos\bar{x}_0, \quad \bar{S} = \bar{R}\sin\bar{x}_0 \tag{1.3}$$

To this definition of mean direction we provide a slight modification, in which we compute the mean direction weighted by wind speed. In other words we have amended the previous definition, replacing 1.1 with

$$\bar{C} = \frac{1}{\lambda}\sum_{i=1}^{n}\lambda_i\cos\theta_i, \quad \bar{S} = \frac{1}{\lambda}\sum_{i=1}^{n}\lambda_i\sin\theta_i \tag{1.4}$$

where λ_i, $i = 1, \ldots, n$ represents wind speed, and $\lambda = \sum_{i=1}^{n}\lambda_i$.

Other than meteorological variables, two other "time" variables have been added, to account for the trend and the seasonality of the pollutants: days within the year, from 1 to 366, and years, in terms of fractions of days within the year,

i.e.: 1973.003,... 2001.997. Clearly for the weekly values, the variables created are weeks within the year, from 1 to 53, and years, in terms of fractions of weeks within the year, i.e.: 1973.019,... 2001.981.

Starting with the analysis of the daily values, a simple graphical inspection of the scatterplots of meteorology against time, is shown in Figure 1.15, Figure 1.16, Figure 1.17 (for wind direction 0 corresponds to North).

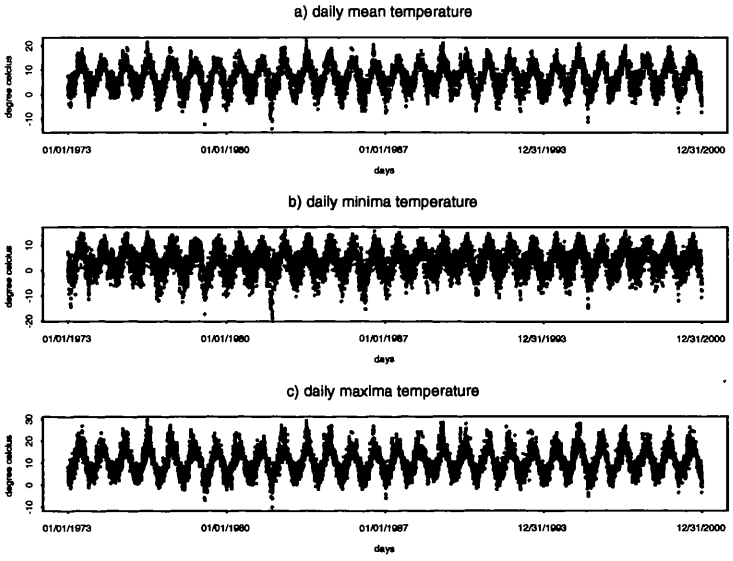

Figure 1.15: a) Mean, b) minima and c) maxima daily Temperature at Eskdalemuir (GB02).

It is possible to note from Figure 1.17 b), that the precipitation is strongly skewed. Therefore its logarithm (Figure 1.17 c), gives a clearer idea of its pattern.

However these plots still show large variation, that could be reduced using the weekly means, as can be seen from the Figure 1.18, Figure 1.19, Figure 1.20.

A first idea of the relation (if any) between pollutants and meteorology, is

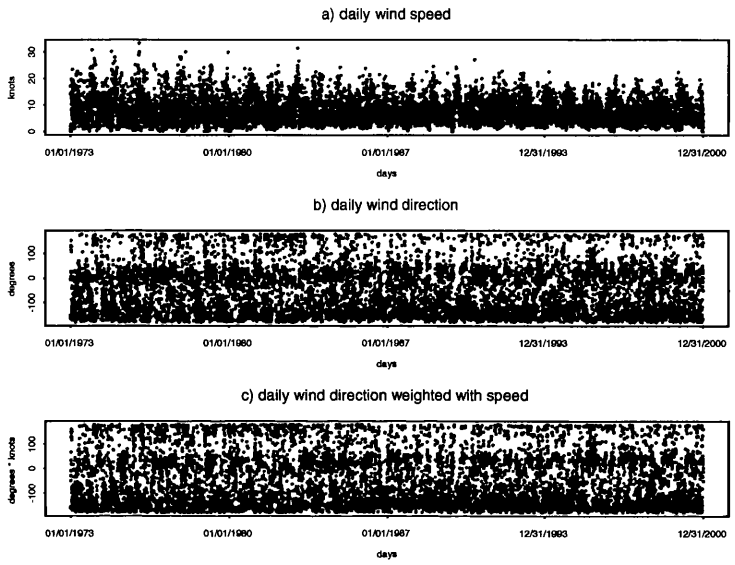

Figure 1.16: a) daily wind speed, b) daily wind direction and c) daily wind direction weighted by speed at Eskdalemuir (GB02).

given by a simple graphical inspection of the plots of pollutants against meteorology, as presented from Figure 1.21 to Figure 1.26. From these plots it is clear that some relationships between meteorology and pollutants exist. However more analysis will be necessary to assess the significance of any meteorological effect on air pollution.

1.5.1 Literature review of Meteorological adjustment in Trend Analysis

One potential assessment of policies to reduce air pollution is to evaluate whether there has been a decrease in pollutant concentrations over time. Unfortunately,

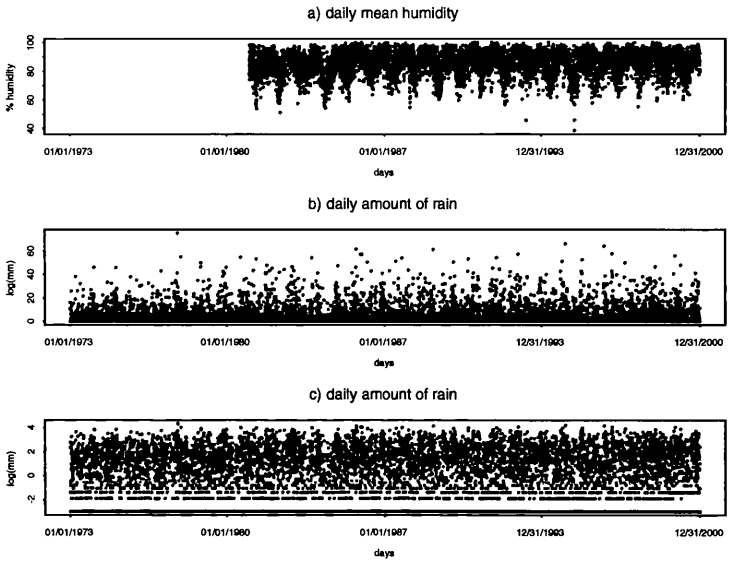

Figure 1.17: a) daily humidity, b) daily amount of precipitation and c) log of daily precipitation at Eskdalemuir (GB02).

attempts to find trends in ambient pollutant concentrations have often been confounded by the effects of meteorology on pollutant formation, accumulation, and destruction. This section presents an overview of the most relevant methodologies concerning meteorological adjustment in trend analysis.

A good review of this topic is presented by Thompson et al. (2001), who underlined how much development is going on in this area while simple linear regression, non-linear regression (Bloomfield et al., 1996), regression trees (Huang and Smith, 1999) and partial least squares (Libiseller and Grimvall, 2002) represent the most commonly used techniques. In recent years there has been an increase in the use of techniques such as cluster analysis (Davis et al., 1998), and artificial neural networks (Gardner and Dorling, 2000b), (Libiseller and Nordgaard, 2003), all concerned with meteorological adjustment for trend analysis

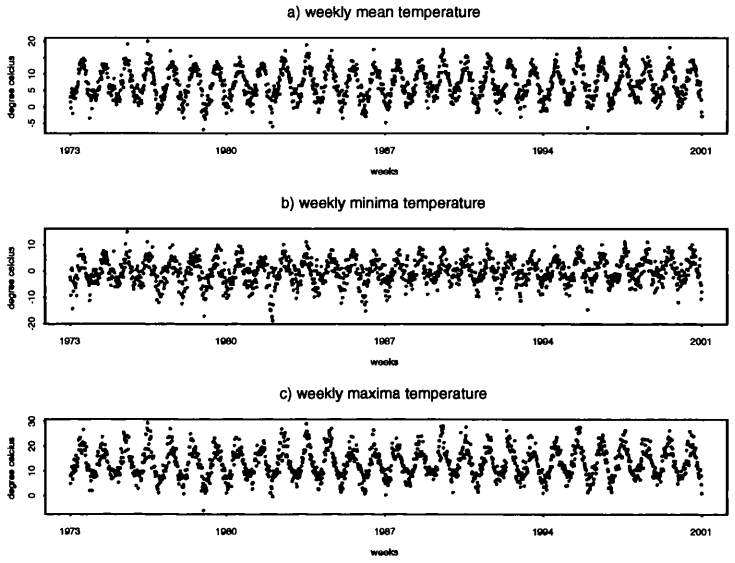

Figure 1.18: a) Mean, b) minima and c) maxima weekly Temperature at Eskdalemuir (GB02).

An important paper in this area was written by Bloomfield et al. (1996). In the past, the methodologies that have been developed to quantify the impact of meteorology on the pollutant compounds were mainly based on linear models, which have difficulty in capturing the complex relationships. Therefore Bloomfield et al. (1996) stress that complex, stratified, non-linear regression models are needed to approximate the true underlying mechanisms. The model strategy proposed by Bloomfield et al. (1996) consists first of all in using scatterplot smoothing and nonparametric regression in order to explore the effects of meteorological variable on the pollutants. On the basis of the nonparametric estimates, a parametric model is then fitted by non-linear least squares. If it is observed that the model residuals have seasonal dependence, a seasonal term is included in the model. A trend component is finally added to the model to describe the

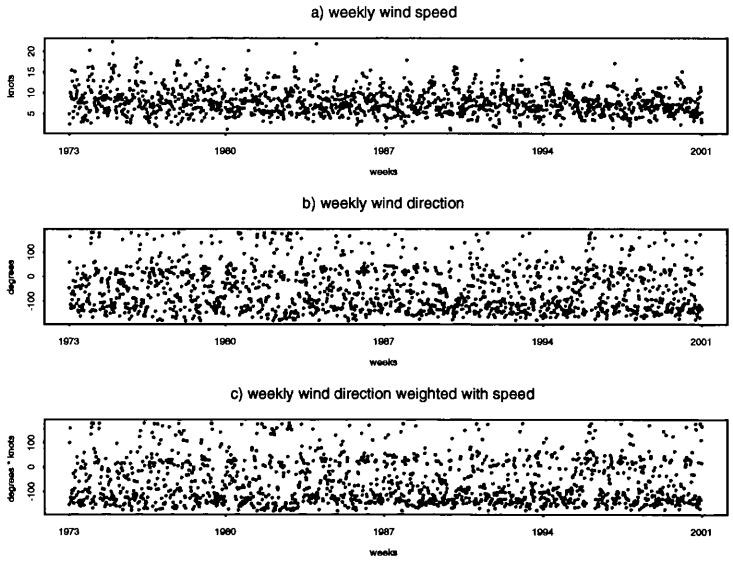

Figure 1.19: a) weekly wind speed, b) weekly wind direction and c) weekly wind direction weighted by speed at Eskdalemuir (GB02).

change in pollutant concentrations, having accounted for meteorological variation and seasonality.

Gardner and Dorling (2000a) used UK ozone data as a case study to demonstrate that statistical models of hourly surface ozone concentrations require interactions and non-linear relationships between predictor variables in order to capture the ozone behaviour accurately. The technique proposed by Gardner and Dorling (2000a) was the Multi-Layer Perceptron (MLP) neural network model that supposes a time series of ozone $O(t)$ as the sum of a long term $e(t)$, a seasonal $S(t)$ and short term $W(t)$ components: $O(t) = e(t) + S(t) + W(t)$. The long term component is due to climate and / or emission and / or background changes. The seasonal component is due to the annual cycle in solar radiation, while the short-term component is due to the day-to-day variations in weather.

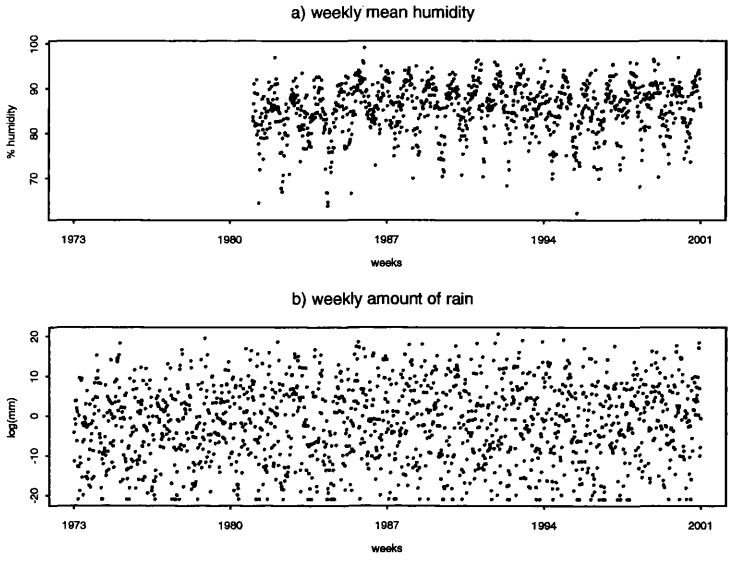

Figure 1.20: a) weekly humidity, b) weekly amount of log precipitation at Eskdalemuir (GB02).

Using UK daily maximum surface ozone concentrations, they show that MLP removed more of the meteorological variability than techniques based on time series filters and regression models. In other work, Gardner and Dorling (2000b) use UK hourly surface ozone concentrations as a case study to compare linear regression, regression tree and multilayer perceptron neural network models. Although MLP are shown to capture any smooth functional relationship between the predictors (meteorological and temporal variables) and the response (hourly ozone concentrations), the regression tree models provides easier physical interpretations. Regression trees were the object of analysis of Huang and Smith (1999), who aimed to separate what they called the genuine trends from meteorological fluctuations. An advantage of this procedure is that it allows different trends at each cluster to be considered, and the variability of the estimated trend among

the clusters is reduced using an empirical Bayesian adjustment.

Similarly to Huang and Smith (1999), Cocchi et al. (2002) proposed a tree based approach. They assumed the daily maxima of ozone concentrations follows a Weibull distribution and proposed a random effects model for the natural logarithm of the quasi-scale parameter of this distribution. The modeling of the natural logarithm of the quasi-scale parameter is very close to the proposal of Cox and Chu (1993) but differently from Cox and Chu (1993), Cocchi et al. (2002) analyzed the quasi-scale parameter in each "group by year" cluster. Besides the Cox and Chu (1993) trend estimation is based on the assumption of a linear functional form for the trend component.

Shively and Sager (1999) proposed a semi-parametric regression approach to model the long-term trend in ozone levels accounting for the effects of meteorological conditions. The authors indeed maintain that the parametric approach risks "incorporating too much prior information" into regression models, but at the same time the nonparametric models may "incorporate too little".

Niu (1996) introduces a class of additive models in which each component of the additive model is fitted using cubic smoothing splines, and, in order to account for serial correlation, an ARMA model is fitted to the error structure at each step of the backfitting algorithm. This technique combines the usual backfitting algorithm with the Box-Jenkins modeling strategy.

Davis et al. (1998) modeled the effects of meteorology on ozone in Houston in two stages. Firstly they used cluster analysis (average linkage and then k-means) and then generalized additive models are used to analyze the relationship between ozone and meteorological variables within each "meteorological regime" (i.e. cluster).

Classification methods were also studied by Cape et al. (2000), who analyzed trace gas measurements arriving at Mace Head (Ireland) at midday between 1995-1997. They combined cluster analysis of the ground-based ozone concentrations (average linkage) with cluster analysis based on the origin of the sampled air (air-mass back trajectories). This technique allows analysis of extreme events by air mass origin.

On the basis of the same classification philosophy, Tørset et al. (2001), studied trends in ambient air concentrations in relation to air mass origin by sector analysis. 5 sectors of 72 degrees were obtained, and for each sector, the Mann-Kendall Test was then performed on the annual average concentrations.

Libiseller (2003) provided a study of the comparison between "one-step" and "two-steps" approaches for trend testing purpose. The first approach is based on multivariate trend tests that try to include covariates as numerical variables in the test formula. The latter approach instead consists firstly in the application of normalization techniques that "clean" the covariate effects, followed by the application of trend detection techniques. Among the "One-Step approaches", Libiseller and Grimvall (2002) presented the Partial Mann-Kendall (PMK) test, and they showed how covariates can be introduced in a general multivariate Mann-Kendall test and in the seasonal Kendall test for trends in serially correlated data. Among the "Two-Steps approaches", Libiseller and Grimvall (2003) presented Partial Least Squares Regression (PLS) to fit linear models to the data, and Artificial Neural Networks (ANNs) to allow for non-linearity. Libiseller et al. (2003) also examined the performances of different normalization models using trajectory data as explanatory variables, along with local meteorological data.

Libiseller (2003) summarized the results, suggesting that multivariate tests

are employed solely to test the significance of trends, while normalization procedures are used primarily to elucidate underlying developments and changes, although they can also be followed by a trend test. In other words, the objective of normalization procedures is to carefully model, and thereby remove, the dependencies between the response and the covariates; whereas the multivariate tests are based on monthly or annual (mean) values, and determine whether both variables develop jointly, correcting the test statistic of the response variable accordingly.

1.6 Conclusions

Having obtained a good idea of the main features of the data, the next chapters will propose and carry out some analysis that deal with the characteristics of the data just identified.

The focus of the statistical methodologies will be on nonparametric procedures. In particular, smoothing techniques for independent data will be described in Chapter 2, and then extended to deal with the correlated case in Chapter 4. In Chapter 3 a diagnostic for detecting abrupt changes in trends will be presented. Chapter 4 will present the fitting and the testing of additive models for correlated data to model pollutant data by an undefined number of covariates. Their properties will be shown through simulation studies. An application to sulphur dioxide with meteorology will be also shown. Chapter 6 will present a spatiotemporal study accounting for both the spatial correlation across years, and temporal correlation across sites. Chapter 7 will present an analysis of the

relationships between observed sulphur dioxide monitored across the EMEP stations and countries' emissions data. Effects of the neighbouring countries will be also analyzed. General conclusions will be discussed in Chapter 8.

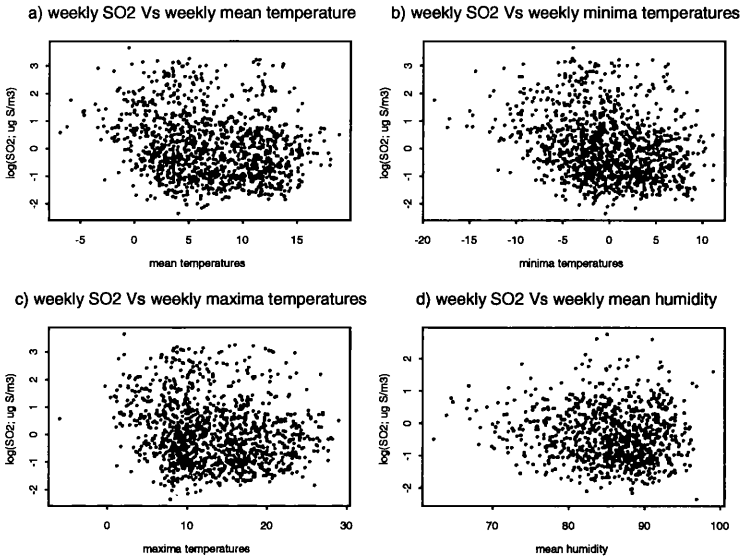

Figure 1.21: a)SO_2 Vs Mean Temperature, b)SO_2 Vs Minima Temperature, c)SO_2 Vs Maxima Temperature, d)SO_2 Vs Humidity, at Eskdalemuir (GB02).

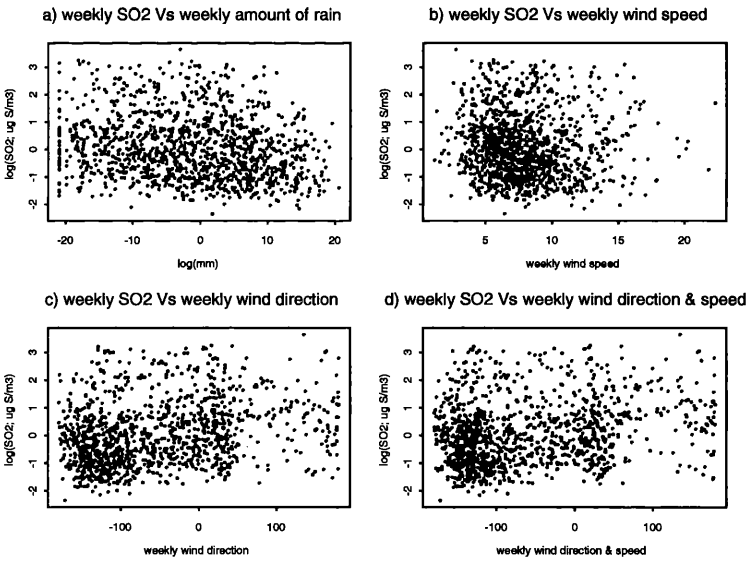

Figure 1.22: a) SO_2 Vs Rain, b) SO_2 Vs Wind speed, c)SO_2 Vs wind direction, d)SO_2 Vs wind direction weighted by speed, at Eskdalemuir (GB02).

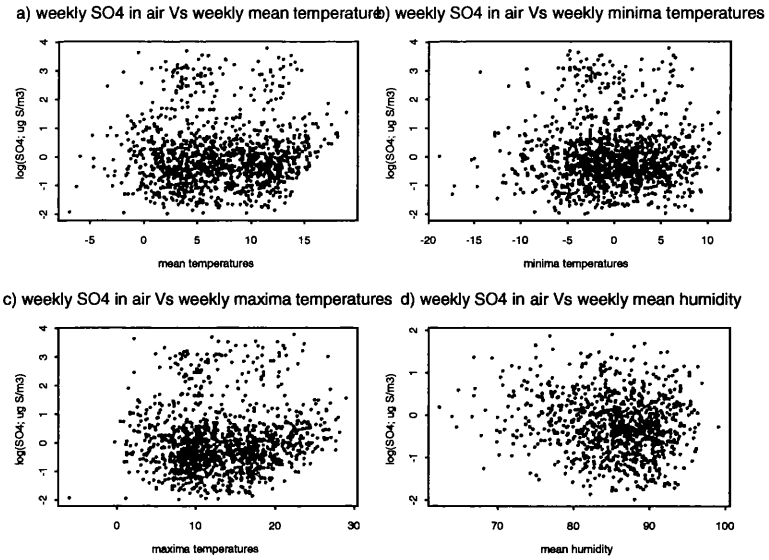

Figure 1.23: a)SO_4 in air Vs Mean Temperature, b)SO_4 in air Vs Minima Temperature, c)SO_4 in air Vs Maxima Temperature, d)SO_4 in air Vs Humidity, at Eskdalemuir (GB02).

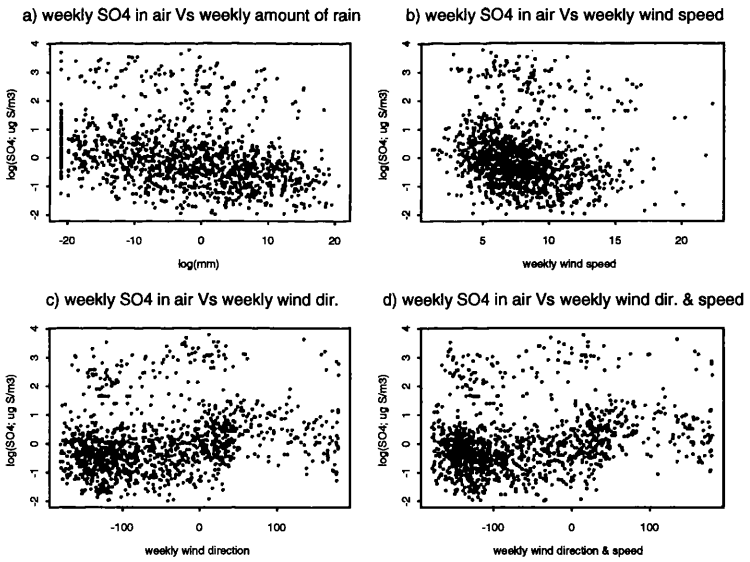

Figure 1.24: a) SO_4 in air Vs Rain, b) SO_4 in air Vs Wind speed, c)SO_4 in air Vs wind direction, d)SO_4 in air Vs wind direction weighted by speed, at Eskdalemuir (GB02).

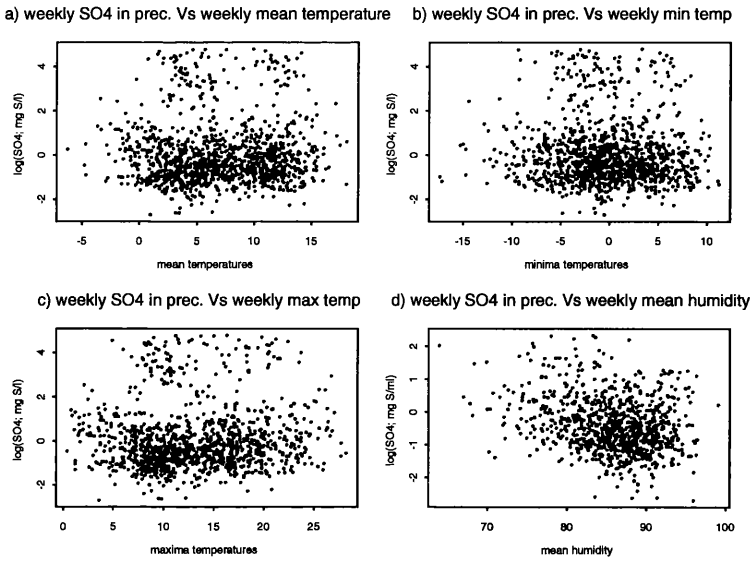

Figure 1.25: a)SO_4 in precipitation Vs Mean Temperature, b)SO_4 in precipitation Vs Minima Temperature, c)SO_4 in precipitation Vs Maxima Temperature, d)SO_4 in precipitation Vs Humidity, at Eskdalemuir (GB02).

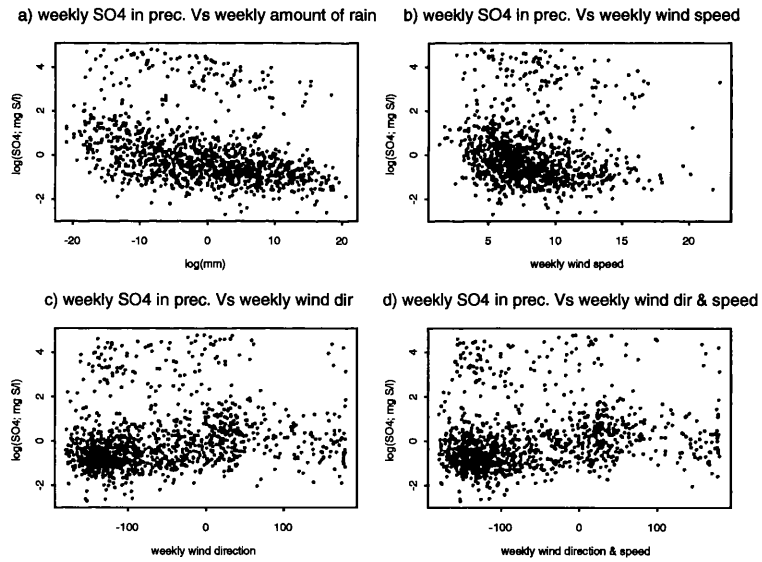

Figure 1.26: a) SO_4 in precipitation Vs Rain, b) SO_4 in precipitation Vs Wind speed, c)SO_4 in precipitation Vs wind direction, d)SO_4 in precipitation Vs wind direction weighted by speed, at Eskdalemuir (GB02).

Chapter 2

Modeling Pollutants With Independent Errors

This chapter presents a modeling approach when the data are assumed independent. Firstly, the theory of fitting and testing nonparametric additive models is presented and then some applications to pollutant concentrations monitored across Europe since 1970's are given.

2.1 Linear and Generalized Additive Models

One of the most commonly used modeling tools is the multiple linear regression model, where dependence of a response variable (y) on a set of covariates ($x_1, \ldots, x_p$), is modeled by:

$$y = a + x_1\beta_1 + \ldots + x_p\beta_p + \varepsilon$$

assuming $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$. As noted by Hastie and Tibshirani (1990), the advantages of this model are that:

- it provides a simple description of the data,
- it summarizes the contribution of each predictor with a single coefficient and,
- it provides a simple method for predicting new observations.

Obviously this model makes a strong assumption about the dependence of $E(y)$ on $x_1, \ldots, x_p$, namely that the dependence is linear in each of the predictors. There are cases where such models cannot be applied because of the intrinsic nonlinearity in the data. It is possible to extend the linear model very simply by adding or modifying terms (such as logarithms, quadratics, and so on), but often it is difficult to guess the most appropriate functional form just from looking at the data. Another powerful tool that generalizes the linear model is nonparametric regression models. Nonparametric regression models the data by letting the data show the appropriate functional form, without pre-specifying any particular shape. This is the idea behind the scatterplot smoother that highlights the functional dependence without imposing a rigid parametric assumption about that dependence.

A smoother is a tool for summarizing the trend of a response measurement y as a function of one or more predictor measurements $x_1, x_2, \ldots, x_p$. It produces an estimate of the trend (m) that is less variable than y itself, hence the name smoother. The notation $\hat{m}$ indicates the estimate produced by a smoother. The single predictor case is the most common and so-called scatterplot smoothing,

defined as a function of x and y, produces a function $\hat{m}$ with the same domain as the values in X. Smoothers have two main uses. The first use is description as through a smoother it is possible to enhance the visual appearance of the scatterplot of y against X. The second, but not less important, use is to estimate the dependence of the mean of y on the predictors. Most smoothers attempt to mimic category averaging through local averaging, that is, averaging the y-values of observations having predictor values close to a target value. The averaging is done in neighbourhoods around the target value.

There are two main decisions to be made in scatterplot smoothing:

1. how to average the response values in each neighbourhood, and

2. how big to take the neighbourhoods.

For the latter question it is necessary to define the size of the neighbourhood in terms of an adjustable smoothing parameter. Intuitively, large neighbourhoods will produce an estimate with low variance but potentially high bias, and conversely for small neighbourhoods. Thus there is a fundamental trade-off between bias and variance, governed by the smoothing parameter or bandwidth. This issue is exactly analogous to the question of how many predictors to put in a regression equation). The former question is really the question of which "brand" of smoother to use, because smoothers differ mainly in their method of averaging. Many approaches are available, as discussed by Green and Silverman (1994), Simonoff (1996), Bowman and Azzalini (1997) and many other authors. Although the methods differ in philosophy and style, the end results in terms of estimation are often similar. It is therefore acceptable to select a method of smoothing which is convenient to the problem at hand.

The local linear method of smoothing (Cleveland, 1979) is adopted here. It is a conceptually appealing way of constructing an estimate from observed data $\{(x_i, y_i);\ i = 1, \ldots, n\}$ by fitting a linear model in a local manner, using weights $w(x_i - x; h)$ to focus attention on the estimation point x of interest (where h is the smoothing parameter or bandwidth). Specifically, the estimator $\hat{m}(x)$ is taken as the least squares estimator $\hat{\alpha}$ which arises from the criterion 2.1.

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \{y_i - \alpha - \beta(x_i - x)\}^2 w(x_i - x; h) \tag{2.1}$$

The weight function $w(.; h)$ should be a smooth, symmetric, unimodal function which is here taken to be a normal density function with mean 0 and standard deviation h, so that $w(x_i - x; h) = e^{-\frac{(x_i - x)^2}{2h^2}}$. General expressions for the bias and variance effects are shown in Ruppert and Wand (1994), and are:

$$\mathbb{E}\{\hat{m}(x)\} \approx m(x) + \frac{h^2}{2}\sigma_w^2 m''(x) \tag{2.2}$$

$$var\{\hat{m}(x)\} \approx \frac{\sigma^2}{nh}\frac{\alpha(w)}{f(w)} \tag{2.3}$$

where σ_w^2 denotes $\int z^2 w(z)dz$, $\alpha(w)$ denotes $\int w(z)^2 dz$, and f denotes the local density of the design points. Expression 2.2 shows that the higher the smoothing parameters and the degrees of curvature (m''), the bigger the bias. Expression 2.3 shows instead that the higher the smoothing parameters and the density of the design points at values neighbouring of x, the smaller the variance. An approximate balance between bias and variance as a function of the smoothing parameter is therefore required.

Bowman and Azzalini (1997) summarize the methods available for estimating the variance σ^2. On the basis of the linear model, one of the most common estimator is given by:

$$\hat{\sigma}^2 = \frac{RSS}{df} \tag{2.4}$$

where df stands for degrees of freedom and their definitions will be shown later on.

An alternative approach was taken by Rice (1984) who proposed an estimator based on differences of the response variable, in order to remove the principal effects of the underlying mean curve. Specifically the estimator is given by:

$$\hat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{i=2}^{n} (y_i - y_{i-1})^2 \tag{2.5}$$

where for simplicity of notation, it is assumed that the observations (x_i, y_i) have been ordered by x_i.

Local linear regression has a number of attractive properties. Conceptually, it can be viewed as a relaxation of the usual linear regression model. As h becomes very large the weights attached to each observation by the kernel functions become similar and the curve estimate approaches the fitted least squares regression line. It is appealing to have this standard model within the nonparametric formulation. From a more theoretical perspective, Fan and Gijbels (1992) and Fan (1993) showed the excellent properties which this estimator possesses.

As said before, in the present work, the kernel mainly used is the normal density function. However, when dealing with particular variables, it has been necessary to amend the kernel function. In particular, variables such as "weeks of the year" and "wind direction" need to be modeled by circular smoothers. Since

there is no natural analogue of linear regression on a cyclical scale, a local mean estimator can be constructed as

$$\min_{\alpha} \sum_{i=1}^{n} \{y_i - \alpha\}^2 w(x_i - x; h), \tag{2.6}$$

where the weight function is now defined as $w(x_i - x; h) = e^{\frac{1}{h}\cos\left(2\pi\frac{x_i - x}{r}\right)}$, with r representing the range of the sample space for x. The use of this Von Mises weight function ensures that the estimate, taken again to be the least squares solution $\hat{\alpha}$, is adapted to the cyclical scale, with observations at one end influencing the estimate at the other end. Loader (1999) describes a similar approach using different weight functions, that will be briefly described here but it will not be used in the following applications. He uses a tricube weight function $W(d/h) = (1 - |d/h|^3)^3$, where h is the bandwidth and d is a periodic distance function, defined by:

$$d(x_1, x_2) = 2|\sin\{(x_1 - x_2)/(2s)\}| \tag{2.7}$$

where s is a scale parameter. A periodic component through a bivariate model (e.g. trend & seasonality), can be fitted using a bivariate distance function given by:

$$d[(x_1, y_1), (x_2, y_2)] = [2\sin\{(y_1 - y_2)\pi/r_y\}]^2 + \left(\frac{x_1 - x_2}{r_x}\right)^2 \tag{2.8}$$

where r_y and r_x are the width of the range of y and x respectively. $W(\cdot)$ produces the weights to give local parameter estimates of a circular local quadratic model:

$$\mu_x(x_i) = a_0 + a_1 s\sin\{(x_i - x)/s\} + a_2 s^2[1 - \cos\{(x_i - x)/s\}] \tag{2.9}$$

As local regression is based on a weighted least squares criterion, the resulting estimates can be expressed as linear combinations of the elements of the vector of response data y. It is convenient to define a *smoothing matrix* S whose rows contain the weights which are appropriate for estimation points on the scale of the covariate. The vector of estimated values $\hat{m}$ at these points then has the convenient representation $\hat{m} = Sy$. This is particularly useful for the construction of standard errors and methods of model comparison.

An alternative formulation in the case of independent data arises by first rewriting the local least squares criterion (2.1) in vector–matrix form as

$$\{y - \alpha 1_n - X\beta\}^T W \{y - \alpha 1_n - X\beta\} \tag{2.10}$$

where X denotes a vector with ith element $(x_i - x)$ and the matrix W has the elements $w(x_i - x; h)$ down the diagonal and 0's elsewhere. Explicit solution of criterion (2.1) for the standard local linear estimator is given in expression (2.11), and details are given in Wand & Jones (1995), Bowman and Azzalini (1997) and other authors.

$$\alpha_x = \frac{(\sum_i^n w_i y_i)(\sum_i^n x_i^2 w_i) - (\sum_i^n w_i x_i y_i)(\sum_i^n w_i x_i)}{(\sum_i^n w_i x_i^2)(\sum_i^n w_i) - (\sum_i^n w_i x_i)^2} \tag{2.11}$$

The local constant estimator is a special case of the local linear estimator, and its explicit representation follows as:

$$\hat{m}(x) = \frac{\sum_j w_j y_j}{\sum_j w_j} \tag{2.12}$$

Bowman and Azzalini (1997) suggested that standard error bands for the smooth estimate $\hat{y}$ can be built by computing standard errors, given by:

$$s.e.(\hat{y}) = \sqrt{var(Sy)} = \sqrt{diag(SS^T)\sigma^2} \tag{2.13}$$

where $var(y) = \sigma^2 I$, and I is the identity matrix.

The local linear regression smoother given in expression (2.1) can be extended to the case of two covariates. The bivariate local linear regression smoother, is defined as the value of $\hat{\alpha}$ from the weighted least squares problem:

$$\min_{\alpha,\beta,\gamma} \sum_{i=1}^{n} \{y_i - \alpha - \beta(x_{1i} - x_1) - \gamma(x_{2i} - x_2)\}^2 w_1(x_{1i} - x_1; h_1) w_2(x_{2i} - x_2; h_2) \tag{2.14}$$

Denoting by X the $n \times 3$ matrix whose ith row is $\{1, (x_{1i} - x_1), (x_{2i} - x_2)\}$, and with W the diagonal matrix whose (i,i)th element is $w_i = w_1(x_{1i} - x_1; h_1) w_2(x_{2i} - x_2; h_2)$, then the solution of the weighted least-squares problem (2.14), is

$$(X^T W X)^{-1} X^T W y \tag{2.15}$$

The local linear estimate is defined by the first element of this vector of length 3. The elements of the 3×3 matrix $A = (a_{ij}) = (X^T W X)$ are all of the form $\sum_i w_i (x_{1i} - x_1; h_1)^r (x_{2i} - x_2; h_2)^s$, where $r + s \leq 2$. To obtain the first element of the least squares solution, we need only the first row of $(X^T W X)^{-1}$, denoted by (b_1, b_2, b_3). By applying standard linear algebra results, reported for instance by

Healy (1986, section 3.4) these can be written as:

$$\begin{aligned} b_1 &= 1\Big/\Big(a_{11} - \frac{1}{d}\{(a_{12}a_{33} - a_{13}a_{23})a_{12} + (a_{13}a_{22} - a_{12}a_{23})a_{13}\}\Big) \\ b_2 &= \frac{b_1}{d}(a_{13}a_{23} - a_{12}a_{33}) \\ b_3 &= \frac{b_1}{d}(a_{12}a_{23} - a_{13}a_{22}) \end{aligned}$$

where $d = a_{22}a_{33} - a_{23}^2$. When the vector (b_1, b_2, b_3) is post-multiplied by X^TW, the result is a vector of length n, whose ith element is $\{b_1w_i + b_2(x_{1i} - x_1)w_i + b_3(x_{2i} - x_2)w_i\}$. The inner product of this vector with y produces the local linear estimates at (z_1, z_2).

Local linear regression smoothing presents some problems when more than two predictors are present, namely:

- with neighbourhoods of two or more dimensions there is usually some metric assumptions made which are difficult to justify when the variables are measured in different units or are highly correlated,
- the "curse of dimensionality", i.e. as the number of independent variables, p, increase, a fixed number of points, n, rapidly becomes sparse. When p is large then, the number of neighbourhoods is less local for a fixed span than a single variable smoother and large bias will result,
- Multivariate versions are computationally expensive to compute.

It is with these problems in mind that Hastie and Tibshirani (1990) took a different approach and used the one dimensional smoother as a building block for a restricted class of nonparametric multiple regression models. In fact the additive

model that they proposed has the following form:

$$y_l = \alpha_l + \sum_{j=1}^{p} m_j(x_{jl}) + \varepsilon_l, \quad l = 1, \ldots, n \tag{2.16}$$

where the errors ε are independent of the $x_j, E(\varepsilon) = 0, Var(\varepsilon) = \sigma^2$ and the m_j are arbitrary univariate functions, one for each predictor, and it is easy to imagine them as smooth functions that are individually estimated by a scatterplot smoother. In fact additive models retain the important feature that they are additive in the predictor effects which can be examined separately. That means that the nature of the effects of a variable on the response does not depend on the values of the other variables, thus the smoothing is always one dimensional and consequently no dimensionality problems occur, at the cost obviously of an approximation of the errors. So the variation of the fitted response surface holding all but one predictor fixed does not depend on the values of the other predictors. In practice this means that, once the additive model is fitted to data, it is possible to plot the coordinate functions separately to examine the roles of the predictors in modeling the response. Such simplicity does not come free; the additive model is almost always an approximation to the true regression surface, but hopefully a useful one. Additive models are more general approximations than linear models.

There are many ways to approach the formulation and estimation of additive models. Typically they differ in the way the smoothness constraints are imposed on the functions in the model. The backfitting algorithm (Hastie and Tibshirani, 1990) is a general algorithm that enables one to fit an additive model using any regression-type fitting mechanisms. It is an iterative fitting procedure, and this is the price one pays for the added generality. Conditional expectations provide

a simple intuitive motivation for the backfitting algorithm. If the additive model is correct, then for any covariate k, $E\{y - \alpha - \sum_{j \neq k} m_j(x_j)|x_k\} = m_k(x_k)$. This immediately suggests an iterative algorithm for computing all the m_j, which in terms of data and arbitrary scatterplot smoothers S_j, can be summarized in the following steps:

1. Initialize: $\alpha = \hat{y}, \hat{m}_1^{(1)} = S_1(y - \hat{\alpha}), \hat{m}_j^{(1)} = S_j(y - \hat{\alpha} - \sum_{k<j} \hat{m}_k^{(1)}), j = 1, \ldots, p$
2. Cycle: $\hat{m}_j^{(i)} = S_j(y - \hat{\alpha} - \sum_{k>j} \hat{m}_k^{(i-1)} - \sum_{k<j} \hat{m}_k^{(i)}),\ j = 1, \ldots, p,$
3. Continue 2) until the individual functions don't change,

where $\hat{m}_j^{(i)}$ indicates the smoother of variable j at iteration i. In order to ensure unique definitions of the estimators, the intercept term can be held at $\hat{\alpha} = \bar{y}$, the sample mean, throughout and additional adjustment to ensure that $\sum_l \hat{m}_j^{(i)}(x_{jl}) = 0$ for each j can be applied at each step.

The main idea is to fit the functions simultaneously, so the individual smoothing steps make sense. When readjusting $\hat{m}_j$, the effects of all the other variables from y are removed, before smoothing this partial residual against x_j.

Applications are also given in Hastie and Tibshirani (1987). Hastie and Tibshirani (2000) proposed a general procedure for posterior sampling from additive and generalized additive models. They proposed a Bayesian backfitting procedure that smooths the same partial residual that the usual backfitting algorithm does, and then adds appropriate noise to obtain a new realization of the current function. This is equivalent to Gibbs sampling for an appropriately defined Bayesian model.

Opsomer and Ruppert (1997) explored the sufficient conditions guaranteeing convergence of the backfitting algorithm for the bivariate additive model, using

local polynomial regression. Opsomer and Ruppert (1997) also showed the asymptotic properties of the estimators. Opsomer and Ruppert (1997) provided the theoretical framework that Opsomer and Ruppert (1998) applied in developing a plug-in bandwidth selection method for additive models. Opsomer (2000) derived explicit expressions for the backfitting estimators of the component functions of $\mathcal{D}$-dimensional additive models for general linear smoothers.

Hastie and Tibshirani (1990) have also proposed another class of models called Generalized Additive Models, that allow an extension of the generalized linear model in the same way that the additive model extends the linear regression model. The Generalized Linear Model assumes that the expectation of y, denoted by μ, is related to the set of covariates $x_1, \ldots, x_p$ by $g(\mu) = \eta$ where $\eta = \alpha + x_1\beta_1 + \ldots + x_p\beta_p$, and y is assumed to have exponential family density as follow:

$$\rho_y(y;\theta;\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

where θ is called the natural parameter, ϕ is the dispersion parameter, η is the systematic component, called the linear predictor, and $g(.)$ is the link function. Generalized Additive Models differ from Generalized Linear Models in that an additive predictor replaces the linear predictor. Specifically, we assume that the response y has an exponential family density, with mean $\mu = E(y|x_1, \ldots, x_p)$ linked to the predictors via

$$g(\mu) = \alpha + \sum_{j=1}^{p} m_j(x_j)$$

The estimation of α and $m_1, \ldots, m_p$ is accomplished by replacing the weighted linear regression in the adjusted dependent variable regression by an appropriate algorithm for fitting a weighted additive model, proposed by Hastie and Tibshirani (1990) and called a local scoring procedure.

2.2 Testing Models

Once the models have been fitted, the next step is to compare them. The following sections will present some of the most common tests for model selection in both linear and nonparametric cases. The theory behind these tests is explained in order to have a better understanding, when their generalized versions that account for correlation will be presented.

2.2.1 Testing Linear Models with Uncorrelated Errors

In fitting a linear model $y = Xb + \varepsilon$, a general hypothesis that sometimes is of interest is $H : R^T b = L$, where y is the n-order response vector, X is the $n \times p$ design matrix of rank p, b is the p-order parameter vector, R^T is any matrix of s rows and p columns, L is a vector of order s of specified constant value. The only limitation on R^T, is that it must have full row rank ($r(R^T) = s$). The F-statistic that is usually formulated to test the hypothesis $H : R^T b = L$, is based on the following assumptions:

$$\begin{aligned} y &\sim N(Xb, \sigma^2 I) \\ \hat{b} &= (X^T X)^{-1} X^T y \\ \hat{b} &\sim N[b, (X^T X)^{-1} \sigma^2] \end{aligned} \tag{2.17}$$

and therefore $R^T\hat{b} - L \sim N[R^Tb - L, R^T(X^TX)^{-1}R\sigma^2]$.

The F test is also based on three theorems that are worth recalling (Searle, 1971):

- Theorem 1: When x is $N(\mu, V)$, then $E(x^TAx) = tr(AV) + \mu^TA\mu$.
- Theorem 2: When x is $N(\mu, V)$, then $x^TAx \sim \chi^2[r(A), \mu^TA\mu/2]$ if and only if AV is idempotent.
- Theorem 3: When $x \sim N(\mu, V)$, the quadratic forms x^TAx and x^TBx are distributed independently if and only if $AVB = 0$ (or equivalently $BVA = 0$).

Recalling Theorem 2, it is possible to see that, setting Q to:

$$Q = (R^T\hat{b} - L)^T[R^T(X^TX)^{-1}R]^{-1}(R^T\hat{b} - L)$$

then

$$Q/\sigma^2 \sim \chi^2\{s, (R^Tb - L)^T[R^T(X^TX)^{-1}R]^{-1}(R^Tb - L)/2\sigma^2\}$$

Replacing $\hat{b}$ by $(X^TX)^{-1}X^Ty$, and knowing that $(R^TR)^{-1}$ exists, because R^T has full row rank, it is possible to express Q in the following way:

$$Q = [y - XR(R^TR)^{-1}L]^TX(X^TX)^{-1}R[R^T(X^TX)^{-1}R]^{-1}R^T(X^TX)^{-1}X^T[y - XR(R^TR)^{-1}L]$$

Noticing that $[I - X(X^TX)^{-1}X^T]$ is symmetric and idempotent, and $X^T[I - X(X^TX)^{-1}X^T] = [I - X(X^TX)^{-1}X^T]X = 0$, it is possible to write the residual

sum of squares:

$$\begin{aligned} RSS &= (y-\hat{y})^T(y-\hat{y}) = y^T[I - X(X^TX)^{-1}X^T]y \\ &= [y - XR(R^TR)^{-1}L]^T[I - X(X^TX)^{-1}X^T][y - XR(R^TR)^{-1}L] \end{aligned}$$

Therefore recalling Theorem 2, it is possible to write

$$RSS/\sigma^2 \sim \chi^2\{r[I - X(X^TX)^{-1}X^T], b^TX^T[I - X(X^TX)^{-1}X^T]Xb/2\sigma^2\}$$

Both Q and RSS are also expressed as quadratics in the vector $y - XR(R^TR)^{-1}L$, which is a normally distributed vector, and the matrix in each quadratic is idempotent. The product of the two matrices is null:

$$[I - X(X^TX)^{-1}X^T]X(X^TX)^{-1}R[R^T(X^TX)^{-1}R]^{-1}R^T(X^TX)^{-1}X^T = 0$$

and recalling Theorem 3, hence:

$$\begin{aligned} F(H) &= \frac{Q/s}{RSS/[n-r(X)]} = \frac{Q/s}{\hat{\sigma}^2} \\ &\sim F\{s, n-r(X), (R^Tb - L)^T[R^T(X^TX)^{-1}R]^{-1}(R^Tb - L)/2\sigma^2\} \end{aligned}$$

and under the null hypothesis $H : R^Tb = L, F(H) \sim F_{s,n-r(X)}$.

For the degrees of freedom, a possible interpretation is obtained looking at the expected value of $RSS = y^T[I - X(X^TX)^{-1}X^T]y$. In fact recalling Theorem 1, it is possible to see that:

$$E[RSS] = tr[I - X(X^TX)^{-1}X^T]I\sigma^2 + b^TX^T[I - X(X^TX)^{-1}X^T]Xb$$

$$= r[I - X(X^T X)^{-1} X^T]\sigma^2$$
$$= [n - r(X)]\sigma^2$$

so the degrees of freedom can be seen as the ratio of the expected value of the residual sum of square over the residual error variance ($df = E[RSS]/\sigma^2$).

At this point the F-statistic can be interpreted as the proportional increase in the residual sum of squares that is obtained moving from the full to the reduced model (multiplied by the ratio of the degrees of freedom).

2.2.2 Testing non linear Models with Uncorrelated Errors

2.2.2.1 Approximate F test

All the theory that has been explained up to now is related to the inference for linear models. This framework can be expanded to the nonparametric case, when smoothing is used. In fact, using the definition of the residual sum of squares and of the degrees of freedom of section 2.2.1, it is possible to derive the analogous quantities for the nonparametric case. A general nonparametric model can be formulated as $y = m(X) + \varepsilon$, where $m(X)$ is the true function and the errors ε are independent and identically distributed with mean zero and variance σ^2. Most of the nonparametric estimators of m are characterized by linearity, in such a way that they can be expressed as $\hat{m} = Sy$, where S denotes the smoothing matrix (also called the projection or hat matrix) whose rows consist of the weights appropriate to estimation at each evaluation point. Both $X\hat{b}$ and Sy are estimators of the same quantity, m, for the linear and nonparametric cases respectively. Therefore it is possible to define the residual sum of squares and the

degrees of freedom for the nonparametric case, by analogy with the linear one, in order to define an "approximate F-statistic".

In fact expressing the residual sum of squares as follows:

$$RSS = (y - \hat{y})^T(y - \hat{y}) = (y - Sy)^T(y - Sy) = y^T(I - S)^T(I - S)y$$

and recalling Theorem 1 of 2.2.1, it is possible to write:

$$\begin{aligned} E[RSS] &= tr[(I - S)^T(I - S)\sigma^2] + \mu^T S^T[(I - S)^T(I - S)]S\mu \\ &= [n - \text{tr}(2S - S^T S)]\sigma^2 + \mu^T S^T[(I - S)^T(I - S)]S\mu \end{aligned}$$

where $\mu^T S^T[(I - S)^T(I - S)]S\mu$ is the bias of the estimator when using the smoothing matrix S. If we then assume that the bias is zero, then $n - \text{tr}(2S - SS^T)$ is a good estimator of df. This definition of degrees of freedom has been given by Hastie and Tibshirani (1990). They proposed two other definitions of degrees of freedom for a linear smoother, namely $tr(S)$ and $\text{tr}(SS^T)$. The first definition $tr(S)$, is interpreted as the "effective number of parameters" of a smoother and is the equivalent to the linear regression case where the degrees of freedom are defined as the trace of the hat matrix $(X(X^T X)^{-1}X^T)$. In fact df becomes the sum of the eigenvalues of S, and gives an indication of the amount of smoothing done. The second definition of degrees of freedom $\text{tr}(SS^T)$, is relative to degrees of freedom for variance. If s_i is the ith row of the smoothing matrix S then it follows that the summed variances of the fitted values are given by:

$$\sum_i var(\hat{m}(x_i)) = \sum_i var(s_i y)$$

$$
\begin{aligned}
&= \sum_i var(s_i m + s_i \varepsilon) \\
&= \sum_i s_i (s_i)^T \sigma^2 \\
&= \mathrm{tr}(SS^T)\sigma^2 \\
&= \mathrm{tr}(S^T S)\sigma^2
\end{aligned}
$$

Therefore the definitions of degrees of freedom proposed by Hastie and Tibshirani (1987) are:

$$
\begin{aligned}
df_{par.} &= tr(S) \\
df_{var.} &= tr(SS^T) \\
df_{err.} &= tr(2S - SS^T)
\end{aligned}
$$

Having defined the residual sum of squares and the degrees of freedom, the approximate F test can be formulated similarly to expression (2.18), as follows:

$$
F = \frac{(RSS_0 - RSS_1)/(df_1 - df_0)}{RSS_1/(n - df_1)} \tag{2.18}
$$

where RSS_0, df_0 and RSS_1, df_1 indicate respectively the residual sum of squares and the degrees of freedom for the reduced model and for the full model. Hastie and Tibshirani (1990) proposed that this statistic should be compared to an F distribution with $df_1 - df_0$ and df_1 degrees of freedom, by analogy with linear models. It is necessary to point out however that the presence of bias in the residual sums-of-squares and the absence of the required properties in the underlying projection matrices mean that the test statistic will not follow an F distribution under the null hypothesis. However, this distribution does provide a

helpful benchmark.

It is important to note that the numerator of the approximate F test consists of the difference of the RSS of the reduced minus the full model, that are both affected by the bias expressed in 2.2. Therefore using the same smoothing parameters for both, reduced and full models, the bias of the test will be canceled.

2.2.2.2 Pseudo Likelihood Ratio Test

An alternative method of testing nonparametric models is based on quadratic form calculations. In fact, a tool for quantifying the difference between the residual sums of squares is provided by a statistic such as:

$$F = \frac{RSS_0 - RSS_1}{RSS_1} \tag{2.19}$$

which is proportional to the usual F statistic and whose construction as a ratio scales out the effect of the error variance σ^2 (RSS_0 and RSS_1 are the residual sum of squares of the reduced model and the full one respectively). As in the approximate F test, in formulation (2.19), by using the same smoothing parameter for the reduced and for the full model it will be possible to have a reduction in bias. Formulation (2.19) is essentially the same test statistic as the approximate F test of expression 2.18. The difference concerns the distributional calculations, and a suitable name for expression (2.19) is *Pseudo Likelihood Ratio Test* (PLRT) statistic, as discussed by Bowman and Azzalini (2003). To implement the test, it is necessary to find the distribution of the F statistic under the reduced model

H_0. Expressing the F statistic in terms of quadratic forms:

$$RSS_0 = y^T(I - S_0)^T(I - S_0)y; \quad RSS_1 = y^T(I - S_1)^T(I - S_1)y$$

where S_0 and S_1 denote the smoothing matrix for the reduced and for the full model respectively, it is possible to express the F statistic as the ratio of quadratic forms in Normal random variables with means approximately zero and the same variance, as follow:

$$F = \frac{y^T B y}{y^T A y}$$

where A is the matrix $(I - S_1)^T(I - S_1)$ and B is the matrix $(I - S_0)^T(I - S_0) - (I - S_1)^T(I - S_1)$. Unfortunately, standard results from linear models do not apply because the matrices A and B do not have the necessary properties, such as positive definiteness. Fortunately there are results which can be used to handle statistics of this kind under more general conditions. These results require only that the matrices from which the quadratic forms are created are symmetric, which is the case here. As a first step it is helpful to reformulate the problem, and to focus on the significance of the statistic F as expressed in its p value. This is:

$$p = \mathbb{P}\left\{\frac{y^T B y}{y^T A y} > F_{obs}\right\} = \mathbb{P}\{y^T C y > 0\} \tag{2.20}$$

where F_{obs} denotes the value calculated from the observed data, and C is the matrix given by $(B - F_{obs}A)$. Johnson and Kotz (1972) summarize general results about the distribution of a quadratic form in normal variables, such as $y^T C y$, for any symmetric matrix C. These results can be applied most easily when the normal random variables have mean zero. In the present setting y has mean μ

under the null hypothesis. However, it is easy to see from the form of the residual sum of squares RSS_0 and RSS_1, which are the building blocks of the test statistic F, that μ cancels out because of the differences involved. Indeed the numerator of expression (2.20) can be expressed in the following way:

$$\begin{aligned}\mathbb{E}(y^T By) &= \mathbb{E}[(\mu+\varepsilon)^T(I-S_0)^T(I-S_0)(\mu+\varepsilon) - \\ &\qquad (\mu+\varepsilon)^T(I-S_1)^T(I-S_1)(\mu+\varepsilon)] \\ &= \mathbb{E}[\mu^T(I-S_0)^T(I-S_0)\mu + \varepsilon^T(I-S_0)^T(I-S_0)\varepsilon - \\ &\qquad \mu^T(I-S_1)^T(I-S_1)\mu - \varepsilon^T(I-S_1)^T(I-S_1)\varepsilon] \qquad (2.21)\end{aligned}$$

where using the same smoothing parameter for S_0 and S_1, it is not unreasonable to assume $\mathbb{E}[\mu^T(I-S_0)^T(I-S_0)\mu] \simeq \mathbb{E}[\mu^T(I-S_1)^T(I-S_1)\mu]$ that means that the expected values of the estimates under the reduced (Model 0) and the full (Model 1) model are assumed the same. In this way the bias is canceled out and the numerator of expression (2.20) can be expressed in the following way:

$$\mathbb{E}(y^T By) = \mathbb{E}[\varepsilon^T B\varepsilon] \qquad (2.22)$$

In the denominator the bias will be still present, but since it will make the denominator bigger, and the test statistic smaller, its effect is conservative. The quadratic form $y^T Cy$ is equivalent to the quadratic form $Q = \varepsilon^T C\varepsilon$. The results of Johnson and Kotz (1972) then allow the probability p defined above to be calculated exactly, although in numerical rather than analytical form. Their results show that it is possible to obtain an approximation to the p value, by replacing the distribution of Q by another more convenient distribution, with

the same first three or four moments. This approach is known to work well in a number of similar situations. By matching the moments of an $aX_b^2 + c$ distribution, that is the shifted and the scaled χ^2 distribution, with the moments of Q, it is possible to define $a = |K_3|/(4K_2), b = (8K_2^3)/K_3^2$ and $c = K_1 - ab$, where $K_j = 2^{j-1}(j-1)!tr\{(VQ)^j\}$, where V is the correlation matrix of y (that in the present setting is equal to the identity matrix I). The p value of interest can then be accurately approximated as $1 - q$, where q is the probability of lying below the point $-c/a$ in a χ^2 distribution with b degrees of freedom.

2.2.3 Comparing components of Additive Models with Uncorrelated Errors

In this section two tests are presented to compare components of two different additive models. Fitting two additive models of the following form:

$$y = \alpha_1 + m_x(x) + m_z(z) + \varepsilon_1 \tag{2.23}$$

$$y = \alpha_2 + m_x(x) + \varepsilon_2 \tag{2.24}$$

indicating with $\hat{m}_{x,1}$ and $\hat{m}_{x,2}$, the estimates for m_x of model (2.23) and model (2.24) respectively, interest could be in testing the hypothesis that the estimate $\hat{m}_{x,1}$ is equal to the estimate $\hat{m}_{x,2}$. This problem arises, for the air pollution application, when we want to compare estimates of trends or seasonal components with and without the effect of other covariates, such as meteorological variables. Bowman and Azzalini (1997) proposed a statistic for comparing regression curves,

based on:

$$\tilde{F} = \frac{(\hat{m}_{x,1} - \hat{m}_{x,2})^T(\hat{m}_{x,1} - \hat{m}_{x,2})}{\hat{\sigma}^2} \tag{2.25}$$

where $\hat{\sigma}^2$ denotes an estimate of the error variance σ^2 that can be obtained from equation (2.4) or (2.5). The test statistic (2.25) can be expressed in two different formulations, one similar to the approximate F test (it will be indicated with $\tilde{F}_A$), and another one similar to the Pseudo Likelihood Ratio test (it will be indicated with $\tilde{F}_L$). The one similar to the approximate F test has the following expression:

$$\tilde{F}_A = \frac{y^T(P_{x,1} - P_{x,2})^T(P_{x,1} - P_{x,2})y/df^*}{y^T(I - S_1)^T(I - S_1)y/df} \tag{2.26}$$

where $P_{x,1}$ is the smoothing matrix that gives the smooth estimates $\hat{m}_{x,1} = P_{x,1}y$, similarly $P_{x,2}$ is the smoothing matrix that gives the smooth estimates $\hat{m}_{x,2} = P_{x,2}y$, and S_1 is the smoothing matrix of model (2.23), that produce the estimates $\hat{y} = S_1 y$. It is therefore clear that the numerator of expression (2.26) consists of the sum of squares of the differences between the estimates of the same component (x) of two different models. The denominator of expression (2.26), consists of the estimate of the variance of y. Indeed $y^T(I - S_1)^T(I - S_1)y$ is the residual sum of squares and the degrees of freedom of the denominator (df) are:

$$df = n - tr(2S_1 + S_1^T S_1) \tag{2.27}$$

while the degrees of freedom of the numerator (df^*) are:

$$df^* = tr[(P_{x,1} - P_{x,2})^T(P_{x,1} - P_{x,2})] \tag{2.28}$$

Expression (2.25) can also be written in a similar expression to the Pseudo Likelihood Ratio Test introduced in section 2.2.2.2, and given by:

$$\tilde{F}_L = \frac{y^T(P_{x,1} - P_{x,2})^T(P_{x,1} - P_{x,2})y}{y^T(I - S_1)^T(I - S_1)y} = \frac{y^T Q y}{y^T B y} \tag{2.29}$$

and its p values will be given by

$$p = \mathbb{P}\left\{\frac{y^T Q y}{y^T B y} > F_{L.obs}\right\} = \mathbb{P}\{y^T C y > 0\} \tag{2.30}$$

where $F_{L.obs}$ denotes the value calculated from the observed data, and C is the matrix given by $(Q - F_{L.obs}B)$. Results from this formulation of the test can be obtained from Johnson and Kotz (1972) analysis, summarized in section 2.2.2.1.

For both ways of testing the statistic (2.25), a graphical display can be obtained by drawing standard error reference bands for the difference of the smoothers $\hat{m}_{x,1}$ and $\hat{m}_{x,2}$, given by:

$$s.e.(\hat{m}_{x,1} - \hat{m}_{x,2}) = \sqrt{var\{(P_{x,1} - P_{x,2})y\}} = \sqrt{diag\{(P_{x,1} - P_{x,2})(P_{x,1} - P_{x,2})^T\}\sigma^2} \tag{2.31}$$

where $var(y) = \sigma^2 I$, and an estimate of σ can be obtained by equation (2.4).

2.2.4 Tests for no effect with uncorrelated errors

This section presents two tests for assessing the presence of trends. Fitting a model of the following form:

$$Model\ 1: \ y = m(x) + \varepsilon \tag{2.32}$$

interest could be in testing if there is any effect of x over y. This could be done by comparing *Model* 1 versus *Model* 0, where y is just seen as function of its mean value μ.

$$Model\ 0: \ y = \mu + \varepsilon \tag{2.33}$$

In order to compare *Model* 1 and *Model* 0, Bowman and Azzalini (1997) proposed two tests, based on the ones introduced in sections 2.2.2.1, and shown below.

$$F = \frac{(RSS_0 - RSS_1)/(df_1 - df_0)}{RSS_1/(n - df_1)} \tag{2.34}$$

$$F = \frac{RSS_0 - RSS_1}{RSS_1} \tag{2.35}$$

where RSS_0, df_0 and RSS_1, df_1 indicate respectively the residual sum of squares and the degrees of freedom for the reduced model (*Model* 0), and for the full model (*Model* 1). Both tests are based on the assumption of independent errors, and the RSS and df used are the ones listed below:

$$\begin{aligned} df_i &= n - tr(2S_i + S_i^T S_i); \quad i = 0, 1 \\ RSS_i &= y^T (I - S_i)^T (I - S_i) y; \quad i = 0, 1 \end{aligned}$$

where S_0 and S_1 are the smoothing matrices for *Model* 0 and *Model* 1 respectively. Because of the nature of *Model* 0, S_0 is a matrix whose elements are $\frac{1}{n}$.

Bowman and Azzalini (1997) also proposed standard errors band around *Model* 0, within which *Model* 1 should lie if *Model* 0 and *Model* 1 are less than 2 standard errors apart. The standard errors proposed by Bowman and

Azzalini (1997) are given in equation (2.36)

$$s.e. = \sqrt{var\{(S_0 - S_1)y\}} = \sqrt{diag\{(S_0 - S_1)(S_0 - S_1)^T\}\sigma^2} \tag{2.36}$$

where $var(y) = \sigma^2 I$, and an estimate of σ can be obtained by equation (2.4) or equation (2.5).

2.3 Application of modeling trend and seasonality in pollutants

In this section the trend and the seasonality of the pollutants will be modeled by fitting some parametric, semiparametric and nonparametric models that assume independent errors.

The models that will be fitted are the ones proposed by McMullan (2004), according to the general formula shown in expression (2.37).

$$y_i = \alpha(x_i) + \beta(x_i)cos(2\pi x_i - \theta(x_i)) + \varepsilon_i \tag{2.37}$$

In expression (2.37), y_i corresponds to the concentration of each compound, x_i is the time (week of each year), $\alpha(x_i)$ is the mean level for each compound, $\beta(x_i)$ is the amplitude of the cosine curve and $\theta(x_i)$ is the phase of cosine curve. $\alpha(x_i)$, $\beta(x_i)$ and $\theta(x_i)$ are allowed to vary smoothly over time x_i. The assumptions on ε_i are that they have mean zero and constant variance. From Model (2.37), deciding to fix some parameters and leaving others to vary smoothly over time x_i, it is possible to obtain different models. In particular, the models fitted here

were the following ones:

- *Model* 1 : $y_i = \alpha + \beta \cos(2\pi x_i - \theta) + \varepsilon_i$,
- *Model* 2 : $y_i = \alpha(x_i) + \beta \cos(2\pi x_i - \theta) + \varepsilon_i$,
- *Model* 3 : $y_i = \alpha(x_i) + \beta(x_i) \cos(2\pi x_i - \theta) + \varepsilon_i$,
- *Model* 4 : $y_i = \alpha(x_i) + \beta(x_i) \cos(2\pi x_i - \theta(x_i)) + \varepsilon_i$.

The simple non linear Model 1 is the most basic model with all three terms fixed, and may be fitted using the non linear regression modeling techniques adapted from Venables and Ripley (1994). In Figure 2.1 there is an example of its fit to the logarithm of SO_2 monitored at Eskdalemuir (GB02).

The semiparametric Model 2 allows the mean level to vary smoothly over time while keeping the amplitude and phase of the seasonal variation fixed. It can be expressed in the form:

$$y_i = \beta_0 + \beta_1 z_i + m(x_i) + \varepsilon_i \tag{2.38}$$

where z_i denotes $\cos(2\pi x_i - \theta)$, β_0 is the overall mean level of the pollutant, and $m(x_i)$ allows the mean level to vary smoothly over time while keeping the amplitude and phase of the seasonal variation fixed. To fit this model it is possible to use the results from Green et al. (1985) who considered a penalized least squares approach for this problem. Express Model 2 as follows:

$$y = \mathcal{D}\rho + m(x) + \varepsilon \tag{2.39}$$

where $\mathcal{D}\rho$ represents the linear component. Explicit solutions for the estimates

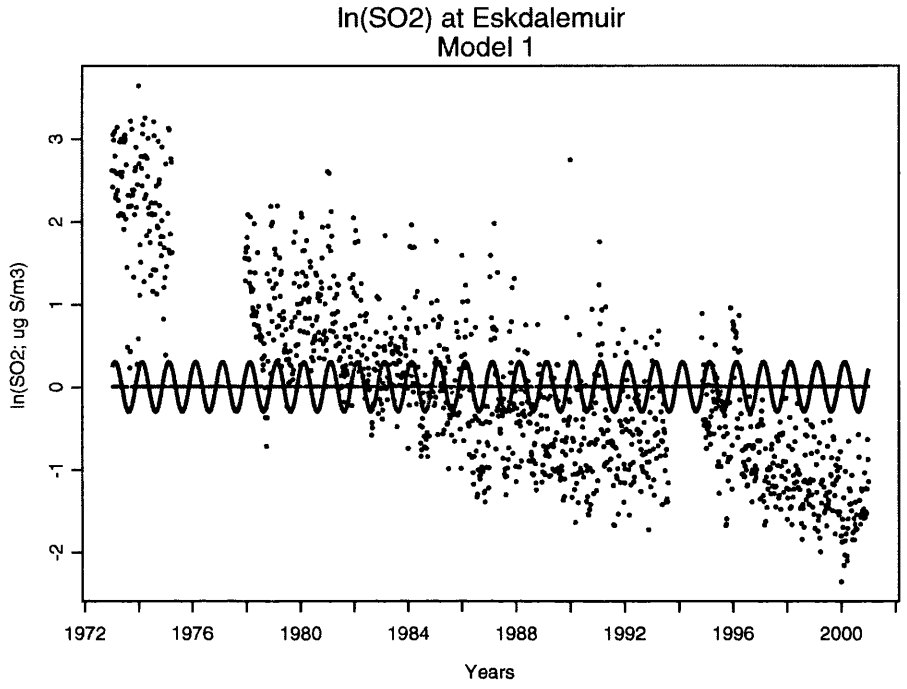

Figure 2.1: Fit of Model 1 for SO_2 monitored at Eskdalemuir (GB02).

can be derived as:

$$\hat{\rho} = (\mathcal{D}^T(I-S)\mathcal{D})^{-1}\mathcal{D}^T(I-S)y \tag{2.40}$$

$$\hat{m}(x) = S(y-\mathcal{D}\hat{\rho}) \tag{2.41}$$

where S is the smoothing matrix. This holds for a linear smoothing technique when there is only one nonparametric term in the model. The `sm.weight` function from the `sm` library is used to construct the smoothing matrix S (Bowman and Azzalini, 1997) using the local linear approach. Figure 2.2 is an example of the model's fit to the logarithm of SO_2 monitored at Eskdalemuir (GB02).

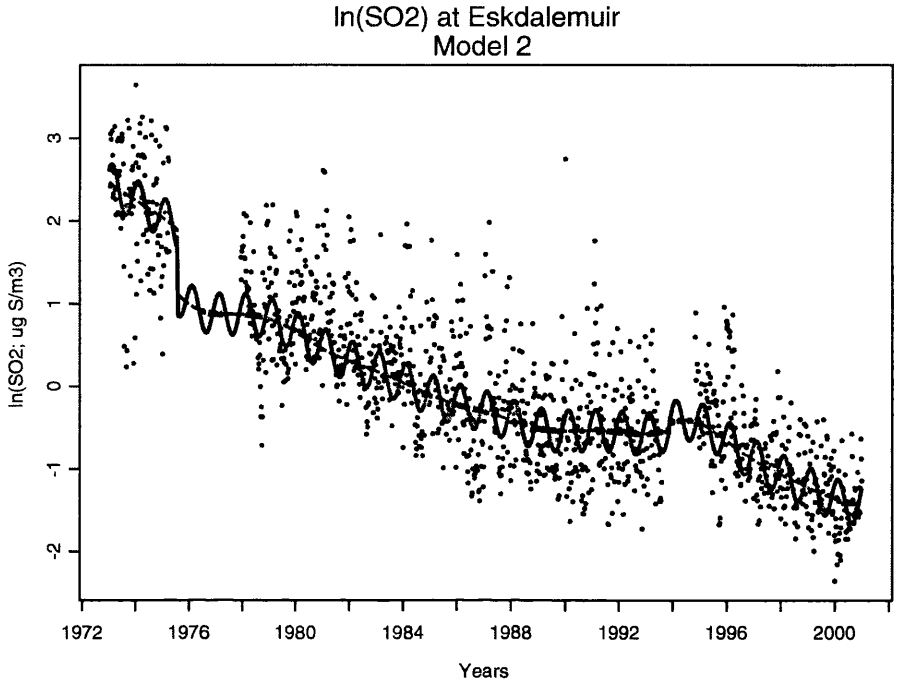

Figure 2.2: Fit of Model 2 for SO_2 monitored at Eskdalemuir (GB02).

Model 3 belongs to the class of models called "varying coefficient model", where the mean level of each compound and the amplitude of the seasonal variation vary smoothly over time, while the phase of the seasonal variation is kept fixed. The additive varying coefficient model can be fitted by applying two dimensional nonparametric regression, using the local linear method with a very large smoothing parameter for the z component, where z again denotes $\cos(2\pi x - \theta)$. This creates an estimator which is linear in z but whose coefficient varies as different neighborhoods of x are used to define the data to which this linear regression is applied. In Figure 2.3 there is an example of its fitting to the logarithm of SO_2 monitored at Eskdalemuir (GB02). It is possible to analyze the three-dimensional

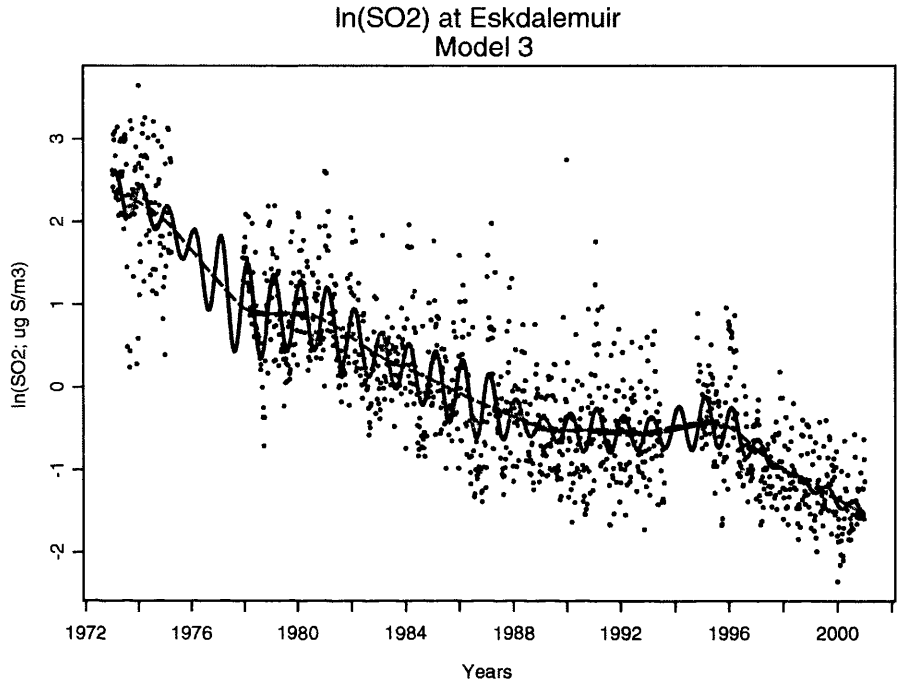

Figure 2.3: Fit of Model 3 for SO_2 monitored at Eskdalemuir (GB02).

nature of this model plotting a graph, whose x-axis is the term $\cos(2\pi x_i - \theta)$, y-axis is time and z-axis is the compound under analysis. It can be seen that the effect of z is fixed to be linear, but the slopes and the intercepts are allowed to vary smoothly over the year (x). Figure 2.4 shows the changing amplitude of the seasonal variation, in other words examining the slope of the various lines along the z-axis, it is possible to identify the effect of amplitude.

The fourth model allows all three terms of the model to vary smoothly over time and so is the most flexible. This non linear varying coefficient model is the most complex to fit as it requires a non linear model to be estimated in a local manner. As suggested by McMullan et al. (2005), the fitting could be achieved

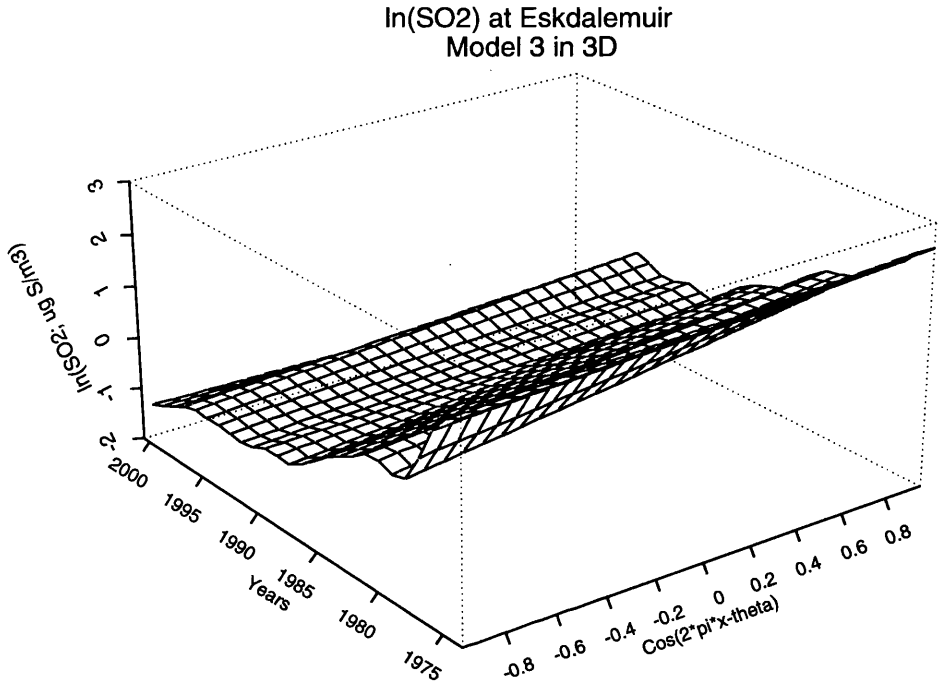

Figure 2.4: Fit of Model 3 for SO_2 monitored at Eskdalemuir (GB02).

by using a technique introduced by Venables and Ripley (1994), who point out that known weights may be handled in non linear regression models by writing the formula as $sqrt(W) * (y - M)$ instead of $y - M$ in Splus, where W is a weight vector at each point of interest x, weights across the corresponding row of W are of the form $exp(-0.5(\frac{x-x_i}{h}))^2$, y is the response and M is the model $(\beta_0(x_i)+\beta_1(x_i)\cos(2\pi x_i-\theta(x_i))$. Hence, estimates for the mean level of pollutants β_0, the amplitude of seasonal variation β_1 and the phase of seasonal variation θ can be obtained at every point of interest x_i. The varying phase in this model allows us to identify seasonal changes. In Figure 2.5 an example of its fit to the logarithm of SO_2 monitored at Eskdalemuir (GB02) is shown.

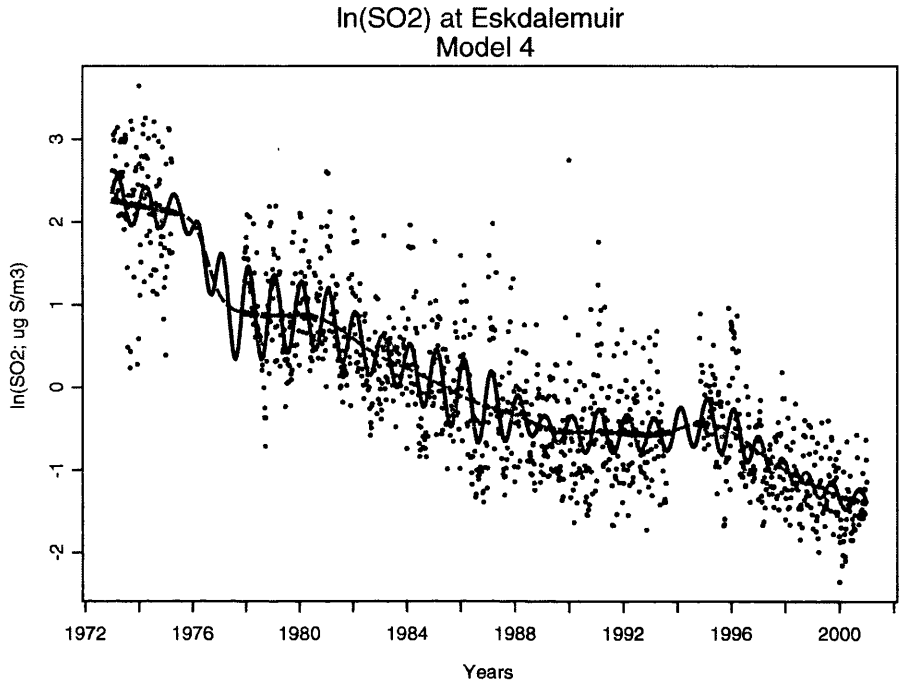

Figure 2.5: Fit of Model 4 for SO_2 monitored at Eskdalemuir (GB02).

These models seems to work quite well. However it is possible to note that there is a large amount of variation around the fitted curves of the four models, indicating that there could be other covariates not included in the model that may improve the fitting of the model. Therefore there is interest in fitting models that can account for meteorological variables. In the following section the fit of Additive Models, that allow several covariates to be included, will be shown.

2.4 Application of Additive Models

In this section we apply Additive Models (AMs) to the daily and then the weekly means of the natural logarithm of SO_2, SO_4 in air and SO_4 in precipitation monitored from 1970's up to 2000 at 11 sites across Europe: Eskdalemuir (GB02, Scotland), Westerland (DE01, Germany), Waldhof (DE02, Germany), Schauinsland (DE03, Germany), Deuselbach (DE04, Germany), Brotjacklriegel (DE05, Germany), Kosetice (CZ03, Czech Republic), Rörvik (SE02, Sweden), Bredkälen (SE05, Sweden), Hoburg (SE08, Sweden) and Payerne (CH02, Switzerland). Meteorological variables have been used as covariates. In particular, the meteorological variables that have been used in this analysis are: Temperature, in terms of Mean, Minimum and Maximum (degrees Celsius), Humidity (%), Precipitation (mm) and Wind Speed (knots), Wind Direction (degrees), and average Wind Direction weighted by Wind Speed.

In the previous chapter, the pollutants and the meteorological variables have been explored. Now a more detailed analysis, assuming independent errors, will give a better understanding of the relations between pollutants and meteorology.

It is necessary to note that some of the variables could have a statistically significant non-linear effect. So AM with a loess smooth for each variable will be fitted using the `gam` function available in Splus or R. Tests of the significance of non-linear effects of the predictors will be computed.

Fitting a full AM to SO_2 at GB02 and testing it against the full linear model, the F statistic gives a significant p value ($1.6e^{-11}$), indicating the presence of some significant non-linear components.

Proceeding with the AMs, the nonparametric fit has been tested against the

linear fit on each term, using the "approximate partial tests". This is a test automatically implemented by Splus and R, and is used for screening variables' inclusion in the model as non linear terms. These tests, in other words, give us information of the importance of the smooth component of each term in the model (the approximate nature of these tests is discussed in Hastie and Tibshirani (1990)). For each variable in the model, this is equivalent to testing for a difference between a linear fit and a smooth fit, which includes a linear term along with the smooth term.

Once the tests have been performed, the best semiparametric model for SO_2 at GB02 is given by:

$$\begin{aligned} \ln(SO_2) &= \beta_0 + lo_1(Y) + lo_2(W) + \beta_1 Mean.T + lo_3(Max.T) + \beta_2 W.S. \\ &+ lo_4(W.D.) + lo_5(W.D.S.) + \varepsilon \end{aligned} \tag{2.42}$$

where Y are the years, W are the weeks, W.D.S. is the Wind direction average weighted by the wind speed, W.D. is the Wind direction average defined by Mardia and Goodall (1993), W.S. is the Wind speed average, Mean.T is the mean temperature, and Max.T is the maximum temperature. This means that for $Y, W, Max.T, W.D., W.D.S.$, there is a statistically significant nonlinear effect, while $Mean.T$ and $W.S$ have a statistically significant linear effect but not a nonlinear one. A visual inspection of the model just built, can be obtained by plotting each of the selected terms in the model. In particular the graphs in Figure 2.6 show all main-effect functions of each predictor, with upper and lower pointwise twice-standard-error curves.

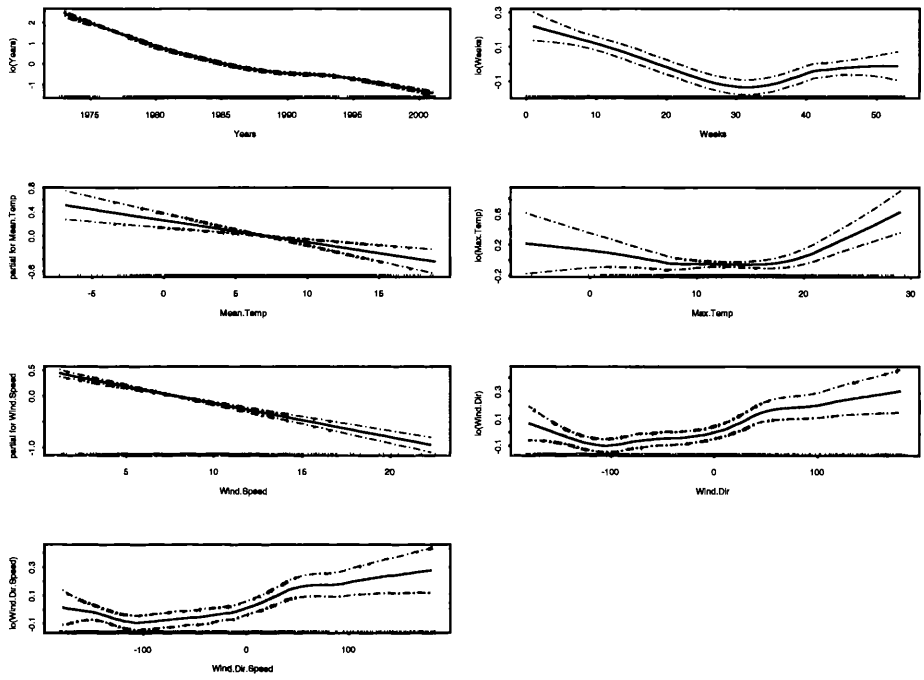

Figure 2.6: Semiparametric fit for SO_2 monitored at Eskdalemuir (GB02).

Applying and fitting semiparametric models to those sites that have meteorology as well, it is not possible to define a common model for all the compounds across all the stations. It really seems that each compound at each station behaves quite differently. However, almost always, highly significant results for the trend and seasonal component are found.

It is necessary to specify that the models presented here are based on the assumption of independence of the errors, an assumption that is definitely not respected by our data. In addition, the `gam` function of Splus/R does not allow circular smoothers to be fitted, and these are required for components such as weeks (*W.*) and wind direction (*W.D.* and *W.D.S.*). Moreover the `gam` function

doesn't release as output the projection matrix, that would be useful for testing purposes. So the need of more general and flexible smoothers to fit each component, and the need of redefining the algorithm for fitting an Additive Model, will represent the main tasks of the following chapters. Before starting the analysis of the Additive Model with correlated data, a diagnostic for detecting discontinuities in correlated regression settings will be shown in the next chapter.

Chapter 3

Detecting Discontinuities

3.1 Introduction

One of the main purposes in analyzing time series in general and the application discussed in this thesis in particular, is to identify the presence of a trend. However the ability to model and detect trends can be affected by a number of features, which need further study. In particular one of the most relevant pieces of information is the presence of any change point in the trend. If a discontinuity is present and is detected, it is necessary first of all to know its cause, and secondly to decide what to do with it. With respect to air pollution data, a discontinuity could be due to several reasons, such as a change in emissions, a change in laboratory or instrumentation used or a particular climatic condition.

The presence of such a discontinuity or change will have an impact on the overall detection of a trend in the pollutant level. It may be necessary to adjust the data series, or if this is not possible, to treat each sub-period separately. A "change point" in this research means a change in the mean level, where the

change can be permanent or temporary. With the EMEP data, the presence of discontinuities and whether any discontinuities are common to stations within each country and across countries are questions of scientific interest. In this section, first some of the methodologies present in the literature are presented, then a methodology for detecting discontinuities proposed by Bowman et al. (2004), and finally an amended version to account for the correlation of the data will be described.

3.2 Detecting Discontinuities: literature review

There is a variety of ways of approaching the problem of detecting discontinuities, and in this section these are classified into two broad classes. The methodologies can be classified as nonparametric or likelihood based.

3.2.1 Literature Review of Discontinuity Detection Methodologies: nonparametric methods.

Starting with the nonparametric approach, there has been considerable development in the use of these methods since the 1990's.

One of the first nonparametric procedures for detecting a change in a distribution was proposed by Bhattacharya and Frierson (1981). This model arises in the context where a machine produces items from which random samples are taken at frequent intervals. This methodology aims to detect small disorders in the machine that change the cumulative distribution function (cdf) of the random samples. A one-sided stopping rule based on cumulative sums of sequential

ranks is considered.

One paper that represents a common reference in many of the recent studies among the nonparametric methodologies to model discontinuities is by McDonald and Owen (1986). They proposed the "split linear smoothers" that are obtained by computing at each "evaluation point" a weighted average of linear fits obtained by windows of various sizes and orientations (windows could be entirely centered, entirely to the left or to the right of the evaluation points). The weights of the "split linear smoother" are obtained by the goodness of fit of the linear models for each evaluation point, window size and orientation.

On the basis of the work of McDonald and Owen (1986), Hall and Titterington (1992) proposed an alternative, edge-preserving, smoothing algorithm with specific analytical properties that is less complicated to implement than the method of McDonald and Owen (1986). Hall and Titterington (1992) based their algorithm loosely on kernel-type smoothing, whereas McDonald and Owen (1986) used ordinary least squares fitting. Bowman et al. (2004) followed a similar approach building an overall test to detect discontinuities comparing one sided local linear regression smoothers. This approach will be described in more detail in the following sections.

Another important paper that gave the insight to many other researchers, was the one proposed by Lombard (1988) who showed how Fourier analysis of the cumulative sum (CUMSUM) statistic can be used in the analysis of change-point detection.

A few years later, Müller (1992) proposed a methodology for detecting change-points based on nonparametric regression analysis. In this article, estimators for

the location and size of discontinuities are obtained as the solution of a maximization problem involving the difference of left and right one-sided kernel smoothers. Müller (1992) mainly focuses on the properties of convergence of the estimators of the locations. Müller and Song (1997) showed that the asymptotic rate of convergence can be improved by adding a second step to the procedure proposed by Müller (1992). Once an initial estimate of the change point and the associated confidence intervals are obtained, the second step consists in maximizing a weighted mean difference within these intervals.

Inspired by the article of Müller (1992), Loader (1996) proposed an estimate of the location of the discontinuity based on a one-sided nonparametric local polynomial model fitted by weighted least squares. The author pointed out that the estimate is similar in principle to that studied by Müller (1992), but by imposing different conditions on the weights assigned to the observations the estimates proposed in this article show higher rates of convergence and also show the same asymptotic distribution as maximum likelihood estimates considered by other authors under parametric regression models.

Qiu and Yandell (1998) focused on the detection of jumps in derivatives of one-dimensional functions. They suggested that if jumps exist in the mth order derivative, then the coefficient of order $m + 1$ of a local polynomial function fitted to the underlying regression function, should show an abrupt change. This method can be applied to the regression function itself simply considering the case $m = 0$

Previously, Qiu and Yandell (1997) use local least squares (LS) to analyze the presence of jumps in regression surfaces (JRS). They suggested to fit a LS plane at each "evaluation point", and using the coefficients of this plane to get

an approximated value of the gradient direction of the JRS. The jump detection criterion is then based on the comparison of two neighbours along the gradient of the LS plane.

Local polynomials have also been used recently by Horvath and Kokoszka (2002), who compared the estimates of the coefficients of local polynomials fitted from left and from right.

Linear splines to estimate discontinuous regression functions, was the approach of Koo (1997). They noted that wherever there is a discontinuity, the fit of two continuous splines is improved by fitting a discontinuous spline and a continuous one.

Local linear kernel regression with long-range dependent errors was analyzed by Anh et al. (1999). In this paper, the authors apply local linear (LL) kernel estimation to test whether the mean function of a sequence of Long Range Dependent (LRD) processes has change-points and they construct nonparametric estimates both for the locations of change-points and for the corresponding jump-sizes. They also establish asymptotic distributions of the constructed estimates.

The change-point problem for dependent observations has been the object of study by Giraitis et al. (1996). This paper proposes an approach for detecting discontinuities of the marginal distribution function on the basis of the asymptotic behavior of Kolmogorov-Smirnov type tests.

Horvath (2001) proposed a change point detection technique for dependent observations assuming that they have a parametric form described by the parameters $(k_k, \lambda_k) = [(k_{k,1}, \lambda_{k,1}), (k_{k,2}, \lambda_{k,2}), \ldots, (k_{k,p}, \lambda_{k,p})]$. They formulated a test

based on the limit distribution that tested the null hypothesis:

$$H_0 : (k_1, \lambda_1) = (k_2, \lambda_2) = \ldots = (k_n, \lambda_n)$$

against the alternative:

H_1 : there is an integer $k^*, 1 \leq k^* < n$, such that

$$(k_1, \lambda_1) = \ldots = (k_{k^*}, \lambda_{k^*}) \neq (k_{k^*+1}, \lambda_{k^*+1}) = \ldots = (k_n, \lambda_n)$$

.

3.2.2 Literature Review of Discontinuity Detection Methodologies: likelihood based methods.

Another large part of the discontinuity literature is concerned with likelihood based methodologies. Gombay and Horvath (1997) presented an application of likelihood to change-point detection. In their paper, they reanalyze the log-transformed data of water discharges from Nacetinsky from 1951 to 1990 analyzed previously by Jaruskova (1997). Her analysis, based on the assumption that the monthly averages follow a log-normal distribution with different means and variances and that the transformed series is an auto-regressive sequence with no changes at the end of the sequence, showed a change in the mean of the transformed variables, but no changes in the variance of the transformed variables could be detected, so there was no change in the shape factor. Their analysis instead demonstrates that a likelihood method can be used to detect possible changes in the parameters of the distributions of the observations.

Jaruskova (1998) presented a review of change detection in the behaviour of meteorological and hydrological series. She illustrated practical problems that can often be encountered and have not yet been solved in using some of the new results present in the study of change-point detection. In particular Jaruskova, after distinguishing two kinds of discontinuity (Sudden Change and Continuous Change), presents, for each of these two, tests to assess the presence of discontinuities, all based on the idea of "test of maximum type". That is, given an interval of n observations, the objective is to look for the time k where the test statistic achieves a maximum value (using a max function). The approximate critical values for these tests, and the description of different types of max function that could be used are presented.

Later on, Hawkins (2001), taking the idea from regression trees, provided an exact and reasonably fast algorithm for performing a multi-way split on the basis of continuous or ordinal predictors. In contrast to the regression trees procedure that refers only to the case of the normal mean, this algorithm is suitable for an arbitrary parameter in an exponential family model.

Another large part of discontinuity studies involved the use of wavelets. Work in this area was suggested by Wang (1995) that used wavelet transformation of one-dimensional functions, in order to detect jumps and sharp cusps. Asymptotic theory is established and practical implementation is discussed in his work. The wavelets used are the Daubechies form with 1, 2 and 3 vanishing moments, and the threshold used is the universal threshold. Later on, Wang (1998) took the same approach in order to extend his technique (Wang, 1995) to detect discontinuities in a two-dimensional function.

Odgen and Parzen (1996) proposed a data dependent technique for selecting a

threshold that divides the "large" coefficients from the "small" ones. The former should describe the significant signal, while the latter are due to noise and are shrunk to zero.

3.3 Methodology

Distinct from most of the methodologies proposed in the previous literature review, Bowman et al. (2004) proposed a global test for detecting discontinuities that would be interesting to use in this study. However Bowman et al. (2004) based their test on the assumption of independent errors, that is not clearly the case for air pollution concentrations. Therefore after presenting the methodology proposed by Bowman et al. (2004), its generalization for correlated data will be shown. Later sections will also show the properties of the proposed test through simulation and applications to air pollution data.

3.3.1 Test for independent data

The situation considered in Bowman et al. (2004) is the one in which the data are noisy observations of a function defined on an interval, and the form of the function is not specified, except that it may have a finite (but unknown) number of discontinuities, and between discontinuities it is smooth. It is important to say that the objective of this analysis is not to estimate the function but to discover the presence of discontinuities. In this research the word "discontinuity" or "change point " means a change in the mean level, where the change can be permanent or temporary. Observing $y_1, y_2, \ldots, y_n$ where $y_i = m(x_i) + \varepsilon_i$ for $i = 1, \ldots, n$, it is assumed that the ε_i are identically normally distributed with

mean 0 and finite variance σ^2.

The smoother used here to estimate $m(z)$ at $z \in (0,1)$, is a local linear regression smoother defined in expression (2.11) of Section 2.1 and here recalled

$$m(z) = \sum_{i=1}^{n} \left\{ \frac{w_i(\sum_{i=1}^{n} x_i^2 w_i) - (w_i x_i)(\sum_{i=1}^{n} w_i x_i)}{(\sum_{i=1}^{n} w_i x_i^2)(\sum_{i=1}^{n} w_i) - (\sum_{i=1}^{n} w_i x_i)^2} \right\} y_i = \sum_{i=1}^{n} s_i y_i \qquad (3.1)$$

where $w_i = w(x_i - z; h)$ is defined by a normal density function with mean 0 and standard deviation h. The true regression function m is assumed to be a piecewise C^r function on the interval [a,b], by which we mean:

- there are $a < t_1 < \ldots < t_d < b$ points at which g has discontinuities;
- at each discontinuity point t_j the left $m(t_j-)$ and the right $m(t_j+)$ limits exist and are different;
- m has at least $r \geq 2$ derivatives at each $x \in [a,b] \setminus \{a, t_1, \ldots, t_d, b\}$.

The case $d = 0$ is allowed and that means that m has no discontinuities. Smoothing through discontinuities gives very poor results in terms of mean integrated squared error, or mean squared error in the neighbourhood of discontinuities. The situation is often much worse than this: the effect of discontinuities may be felt twice by typical nonparametric smoothers which use a global bandwidth selector (h). A bandwidth which is reasonable for the smooth part of the function between discontinuities will cause over-smoothing at the discontinuities, which has an inflating effect on the mean integrated square error. On the other hand, a global bandwidth selector which trades estimated curvature against estimated residual variance may be perturbed by discontinuities in the direction of under-smoothing,

because the function will appear to the selector to be globally rougher than it is away from the discontinuities, or may be perturbed in the opposite direction because the discontinuities may inflate the estimated variance. The overall effect is hard to predict, but one would generally expect the combination to exaggerate any tendency to under-smooth where the function is smooth, and to over-smooth where it is rough or discontinuous. The idea on which this diagnostic is based, is to compare at each point two linear smooths of the data. Each smooth is "one-sided" in that it is defined in terms of data lying entirely to the right or entirely to the left of the point at which we wish to test for a discontinuity. It is important to underline the fact that, among the above assumptions, no information about the derivatives of m is transmitted across a discontinuity: there are no assumptions about the relationship between $m'(t_j-)$ and $m'(t_j+)$. The left-smooth and the right-smooth are given, respectively by:

$$\widehat{m}_l(z) = \sum_{i=1}^{n} s_i I\{x_i < z\} y_i \tag{3.2}$$

$$\widehat{m}_r(z) = \sum_{i=1}^{n} s_i I\{x_i > z\} y_i \tag{3.3}$$

where I denotes the indicator function in the usual way. The main fact that it is necessary to note is that the estimators are linear in the observations y_i. These last two smooths are separate estimates of the value of $m(z)$, using the data lying on opposite sides of the point of interest z. If we ignore the possibility of discontinuities, we would expect the two smooths to have similar values, but if there is a discontinuity then we might hope to detect a difference between the two. The estimate $\widehat{m}_l(z)$ is easily recognized as the value that would have been

obtained if the estimator $\widehat{m}$ had been applied with the data truncated so that all the design points to the right of z were missing. Precisely this situation arises when nonparametric regression estimates are made under standard conditions at the boundary of the data. Thus in general the accuracy of $\widehat{m}_l(z)$ as an estimate of $m(z)$ will depend on how well the smoother m performs at the boundary of the data. For testing purposes, all points z between two adjacent but distinct design points should be effectively equivalent so we need to evaluate the smooths $\widehat{m}_l$ and $\widehat{m}_r$ only at the midpoints of the intervals between distinct design points. So assuming that the design space is equi-spaced, indicated by $x_i = i/n$, for $i = 1, 2, \ldots, n$, the midpoints will be $z_i = (x_i + x_{i+1})/2$. Thus the observations are made in the interval (0,1), and similarly to expression 3.1, the estimates can be written in the follow way:

$$\widehat{m}(z_j) = \sum_{i=1}^{n} s_{ij} y_i = S_j^T y, \quad j = 1, \ldots, n$$

where $y = (y_1, y_2, \ldots, y_n)^T$ is the vector of observations, and although it is not clear from the notation, S_j depends on the smoothing parameter h. Obviously, it is possible, similarly, to re-write the left and right smooths, whose difference is written as:

$$r_j = \widehat{m}_r(z_j) - \widehat{m}_l(z_j) = (-s_{lj}^T, s_{rj}^T) y \tag{3.4}$$

where clearly $S_j^T = (s_{lj}^T, s_{rj}^T)$. Since s_j may be regarded as an estimate of the size of the jump at z_j, it is natural to consider the test statistic:

$$\sum_{j=1}^{n} r_j^2 = y^T D^T D y \tag{3.5}$$

where D is the matrix whose jth row is given by $(-s_{lj}^T, s_{rj}^T)$. It is natural to expect $\sum_{j=1}^{n} r_j^2$ to be large if there are discontinuities and small otherwise. In order to use (3.5) as a test statistic, it will clearly be necessary to standardize it by the variance of r_j. Because $r = Dy$, and its variance is given by $var(r) = diag(DD^T)\sigma^2$, expression 3.5 can be reformulated in the following way:

$$F(h) = \sum_{j=1}^{n} \frac{r_j^2}{\widehat{var}(r_j)} = \frac{y^T D^T \Lambda^{-1} D y}{\hat{\sigma}^2} \tag{3.6}$$

where Λ is a matrix whose diagonal elements are given by $diag(DD^T)$ and zero elsewhere. Assuming independent errors, an alternative estimate $\hat{\sigma}^2$ of the noise variance σ^2 is given by:

$$\hat{\sigma}^2 = \frac{y^T \Delta y}{\mathrm{tr}\Delta} = \frac{\sum_{i=2}^{n}(y_i - y_{i-1})^2}{2(n-1)} \tag{3.7}$$

where

$$\Delta = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & -1 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 2 & -1 \\ 0 & \cdots & 0 & -1 & 1 \end{bmatrix}$$

is the second difference operator.

The null hypothesis that we want to test is absence of discontinuities in the interval $[a, b]$, and so if the observed value of $F(h)$ is bigger than some critical value C_α, chosen so that when H_0 is true $P\{F(h) > C_\alpha\} = \alpha$, the null hypothesis

will be rejected. The problem of defining the distribution under the null hypothesis is solved using results of section 2.2.2.1 by writing this function as the ratio of quadratic forms in Normal random variables with means approximately zero and the same variance, as follow:

$$P\left(\frac{y^TAy}{y^TBy} > t\right) = P(y^TAy - ty^TBy > 0) = P(y^TQy > 0) \tag{3.8}$$

where, t is the observed value of the statistic $F = \frac{y^TAy}{y^TBy}$, where $A = D^T\Lambda^{-1}D$, $B = \frac{\Delta}{\text{tr}(\Delta)}$.

3.3.2 Test for correlated data

This section presents a generalization of the test proposed by Bowman et al. (2004) in order to account for serial correlation of the data, based on the idea of Yap (2004). Indicating with V the estimated correlation matrix, the variance of r_j, given in expression (3.4), now becomes $var(r_j) = var(Dy) = diag(DVD^T)\sigma^2$. Expression (3.6), then becomes:

$$F(h) = \sum_{j=1}^{n} \frac{r_j^2}{\widehat{var}(r_j)} = \frac{y^TD^T\Omega^{-1}Dy}{\hat{\sigma}^2} \tag{3.9}$$

where Ω is a matrix whose diagonal elements are given by $diag(DVD^T)$ and zero elsewhere.

In expression (3.9), it has been assumed that the correlation matrix V is known. In practice, it will often be required to estimate this. As in the case of linear models, an effective strategy is to fit an independence model and use the residuals from this to identify a suitable structure for the error component. This

follows the approach of Niu (1996). In this work, the correlation matrix will be indicated with V, whose $[i,j]$ element is given by $\widehat{\rho}^{|x_i - x_j|}$, and $\hat{\rho}$ is the estimated correlation coefficient at lag 1 of an $AR(1)$ model. Obviously the choice of this formulation is based on the assumption that residuals follow an AR(1) model.

While we would not expect this to hold exactly, it should nevertheless absorb the majority of the structure of the correlation. This aspect of the model is a "nuisance" feature and so it is not the principal focus.

In the nonparametric case an additional issue arises as a result of the bias which is inevitably present in the estimation of the regression function, as discussed in Section 2.1. However, this bias will be transferred to the residuals, leading to inflation of the estimates of the error correlation and variance parameters. Indeed on the basis of expression (2.2), it is possible to write:

$$\begin{aligned} r_j &= y_j - \hat{m}(x_j) \\ \mathbb{E}\{r_j\} &\approx m(x_j) - m(x_j) - \frac{h^2}{2}\sigma_w^2 m''(x) \end{aligned} \tag{3.10}$$

The inflation of the r_j estimates will increase both numerator and denominator of expression (3.9), and an overall reduction of the bias in the test statistic (3.9) is then expected.

It has to be noted in addition that, in the case of correlated errors, the variance estimate of y ($\hat{\sigma}^2$) given by equation (3.7) is not an unbiased estimate. An estimate of the variance of the detrended y, can be written as follow:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{m}_i)^2}{n-1} = \frac{y^T(I-S)^T V^{-1}(I-S)y}{n-1} = y^T \Gamma y \tag{3.11}$$

where $\hat{m} = Sy$ is a nonparametric estimate of the trend, S is the smoothing matrix, and $\Gamma = \frac{(I-S)^T V^{-1}(I-S)}{n-1}$. The expected value of (3.11) is given by:

$$\mathbb{E}(y^T \Gamma y) \simeq \sigma^2 \text{tr}(\Gamma V) \tag{3.12}$$

and therefore from expression (3.12), it follows that:

$$\mathbb{E}\left\{\frac{y^T \Gamma y}{tr(\Gamma V)}\right\} \simeq \sigma^2 \tag{3.13}$$

An appropriate estimate of σ^2 is therefore given by:

$$\hat{\sigma}^2 = \frac{y^T \Gamma y}{\text{tr}(\Gamma V)} \tag{3.14}$$

Following the terminology of Hastie and Tibshirani (1987), the normalizing constant $tr(\Gamma V)$ is referred to as the *approximate degrees of freedom.* When V is replaced by the identity matrix this reduces to one of the standard definitions of approximate degrees of freedom used in the independent errors case.

For the case of correlated errors expression (3.8) still holds, but A and B are now written as:

$$A = D^T \Omega^{-1} D; \quad B = \frac{\Gamma}{\text{tr}(\Gamma V)} \tag{3.15}$$

and $Q = D^T \Omega^{-1} D - t\frac{\Gamma}{\text{tr}(\Gamma V)}$, where t is the observed value of the F statistic.

In the following sections, simulations of the test will be performed and applications to pollutant data will be shown.

3.4 Simulation Study

In this section the size and the power of the discontinuity test have been studied by simulation. Data from four kinds of models have been simulated.

$$\text{flat, } y = c\, I(x > 0.5) + \varepsilon \tag{3.16}$$

$$\text{linear, } y = x + c\, I(x > 0.5) + \varepsilon \tag{3.17}$$

$$\text{quadratic, } y = 4x^2 + c\, I(x > 0.5) + \varepsilon \tag{3.18}$$

$$\text{sine, } y = \sin(2\pi x) + c\, I(x > 0.5) + \varepsilon \tag{3.19}$$

These functions have been chosen because they cover a wide range of situations that are usually present in nature. Simulations consisted in generating 200 data sets of 100 data points equally spaced between 0 and 1 from each of the models listed above. Setting the significance level of the test at $\alpha = 0.05$, we would expect that the number of significant p values under the null hypothesis follows a Binomial distribution, that means that percentage of significant p is in the range $n\ \alpha \pm 2\sqrt{n\ \alpha\ (1-\alpha)} = 5\% \pm 3.1\%$ (i.e. 10 ± 6.2 over the 200 simulated data sets). Simulations have been carried out using the values listed below:

- the ε have been generated from an AR(1) model with different correlation parameters $(0, 0.2, 0.4, 0.8)$ and standard deviation equal to 1;
- $h = 0.08, 0.12, 0.16, 0.20, 0.24, 0.28$, h is the smoothing parameter used for the test;
- $c = 0, 1, 2, 3$.

Some simulation results are presented in Tables 3.1 and 3.2 and these show that the proportions of significant p values from 200 simulated sets of data for correlation values of 0 and 0.4. It is clear that when c is equal to 0, no jump is generated and the results refer to the size of the test. Whenever c is different from 0, a jump is generated and the results therefore refer to the power of the test.

Table 3.1: proportions of significant p values from testing for discontinuities with data simulated from models (3.16) (3.17) (3.18) (3.19) with correlation parameter of 0.

	size ($c = 0$)				power ($c = 1$)			
	flat	linear	quadratic	sin	flat	linear	quadratic	sin
$h = 0.08$	0.070	0.070	0.070	0.065	0.110	0.110	0.100	0.100
$h = 0.12$	0.020	0.040	0.045	0.030	0.150	0.150	0.105	0.090
$h = 0.16$	0.020	0.070	0.065	0.090	0.180	0.180	0.190	0.120
$h = 0.20$	0.035	0.035	0.035	0.285	0.205	0.075	0.240	0.145
$h = 0.24$	0.035	0.035	0.040	0.510	0.305	0.170	0.335	0.150
$h = 0.28$	0.045	0.045	0.065	0.700	0.250	0.270	0.290	0.170

	power ($c = 2$)				power ($c = 3$)			
	flat	linear	quadratic	sin	flat	linear	quadratic	sin
$h = 0.08$	0.220	0.220	0.225	0.265	0.515	0.510	0.525	0.515
$h = 0.12$	0.470	0.470	0.475	0.475	0.785	0.790	0.830	0.760
$h = 0.16$	0.635	0.635	0.625	0.410	0.940	0.945	0.925	0.820
$h = 0.20$	0.725	0.725	0.715	0.265	0.995	0.990	0.970	0.760
$h = 0.24$	0.805	0.805	0.795	0.255	0.990	0.995	0.990	0.675
$h = 0.28$	0.855	0.855	0.830	0.185	0.975	0.975	0.990	0.640

The proportions for the flat, the linear, and the quadratic model are approximately 5% for all smoothing parameters and for all correlation values, except 0.8, while the same cannot be said for the sine trend, where the size for all correlation values does not seem to work with smoothing parameters bigger than 0.16. This

Table 3.2: proportions of significant p values from testing for discontinuities with data simulated from models (3.16) (3.17) (3.18) (3.19) with correlation parameter of 0.4.

	size ($c = 0$)				power ($c = 1$)			
	flat	linear	quadratic	sin	flat	linear	quadratic	sin
$h = 0.08$	0.145	0.145	0.145	0.130	0.185	0.185	0.175	0.170
$h = 0.12$	0.075	0.075	0.075	0.075	0.135	0.135	0.125	0.085
$h = 0.16$	0.060	0.060	0.060	0.065	0.105	0.105	0.125	0.060
$h = 0.20$	0.055	0.055	0.060	0.110	0.140	0.140	0.115	0.060
$h = 0.24$	0.060	0.060	0.065	0.165	0.115	0.115	0.150	0.085
$h = 0.28$	0.040	0.040	0.035	0.245	0.115	0.115	0.140	0.100

	power ($c = 2$)				power ($c = 3$)			
	flat	linear	quadratic	sin	flat	linear	quadratic	sin
$h = 0.08$	0.280	0.280	0.275	0.240	0.380	0.320	0.360	0.395
$h = 0.12$	0.245	0.245	0.235	0.225	0.385	0.395	0.430	0.385
$h = 0.16$	0.270	0.270	0.255	0.150	0.440	0.470	0.425	0.335
$h = 0.20$	0.275	0.275	0.265	0.135	0.490	0.495	0.570	0.250
$h = 0.24$	0.335	0.335	0.330	0.080	0.585	0.605	0.600	0.225
$h = 0.28$	0.360	0.360	0.315	0.055	0.615	0.570	0.555	0.200

is mainly due to the fact that the sine is a cyclical function and therefore the amount of smoothing applied should not be any bigger than $h = 0.16$. Figures 3.1 and 3.2 show the size/power as a function of the smoothing parameter for each type of trend. It is possible to note how the huge amount of smoothing for the sine trend functions affects the results of both size/power. For the flat, the linear, and the quadratic trend, the power of the test increases as soon as the smoothing parameter increases; for the sine trend, the power has maximum value for a smoothing parameter around 0.12, and then it reduces as soon as the smoothing parameter increases.

It is possible to note from figures 3.2, that for a correlation of 0.8 the size of the test does not seem to work for all four kinds of trend, and for all smoothing parameters. Indeed it is expected that with the higher correlations, the simulated data will be less smooth and therefore the detection of discontinuities will be more difficult.

The power of the test has been further analyzed by looking at the location of the discontinuities detected. Up to now the power simulation study has analyzed how powerful the test is in detecting the presence of a simulated discontinuity. In order to obtain an indication of the locations of the discontinuities identified, the standardized difference between the left and the right smooths at each point has been computed. Indeed it is expected that if the p values are significant, the point with the highest standardized difference of the left and the right smooth is a possible candidate for locating the discontinuity. Therefore for each of the four correlation parameters (0, 0.2, 0.4, 0.8), for each of the three jumps ($c = 1$, $c = 2$, $c = 3$), and for each of the smoothing parameter values ($h = 0.08, 0.12, 0.16, 0.20, 0.24, 0.28$) the proportions of significant discontinuities detected with the highest standardized difference of left minus right smoother, whose location was between 0.4 and 0.6 have been computed and displayed in Figures 3.3, 3.4. Comparing Figures 3.3, 3.4 with Figures 3.1, 3.2, it is possible to note that nearly always, whenever the p value is significant, the discontinuity detected with the highest standardized difference of left - right smoother, is located between 0.4 and 0.6. Indeed for all the correlation values, for all the trend types, and for all the size of jumps, the graphs of the power as functions of the smoothing parameters are very similar to the graphs of the proportions of significant discontinuities detected with the highest standardized difference of

left-right smoother, whose location is between 0.4 and 0.6.

Figures 3.5, 3.6 show the histograms of the locations of the significant discontinuities detected with the highest standardized difference of left - right smoothers, for different values of smoothing parameter ($h = 0.08, 0.12, 0.16, 0.20, 0.24, 0.28$), for flat, linear, quadratic and sine trends, with correlation parameter of 0.2 and a jump of $c = 2$. The figures show that most of the discontinuities detected are located between 0.4 and 0.6, confirming that the discontinuities detected with the highest standardized difference of left - right smoother represent a useful guideline in determining the location of the discontinuity.

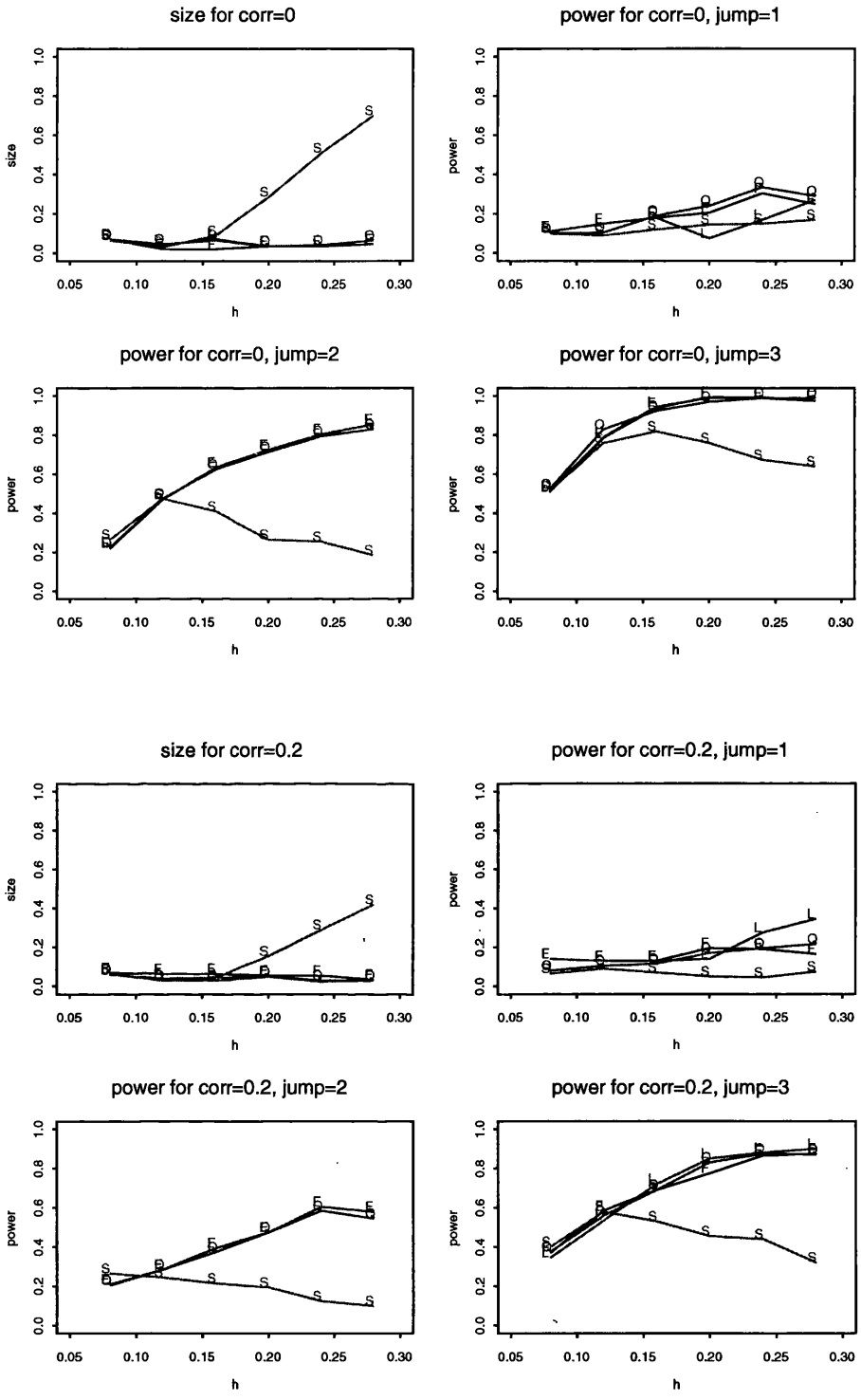

Figure 3.1: size (jump=0) and power (jump=1, 2 and 3) of the discontinuity test as function of the smoothing parameter, for flat (F), linear (L), quadratic (Q) and sine (S) trend, and with correlation = 0, 0.2.

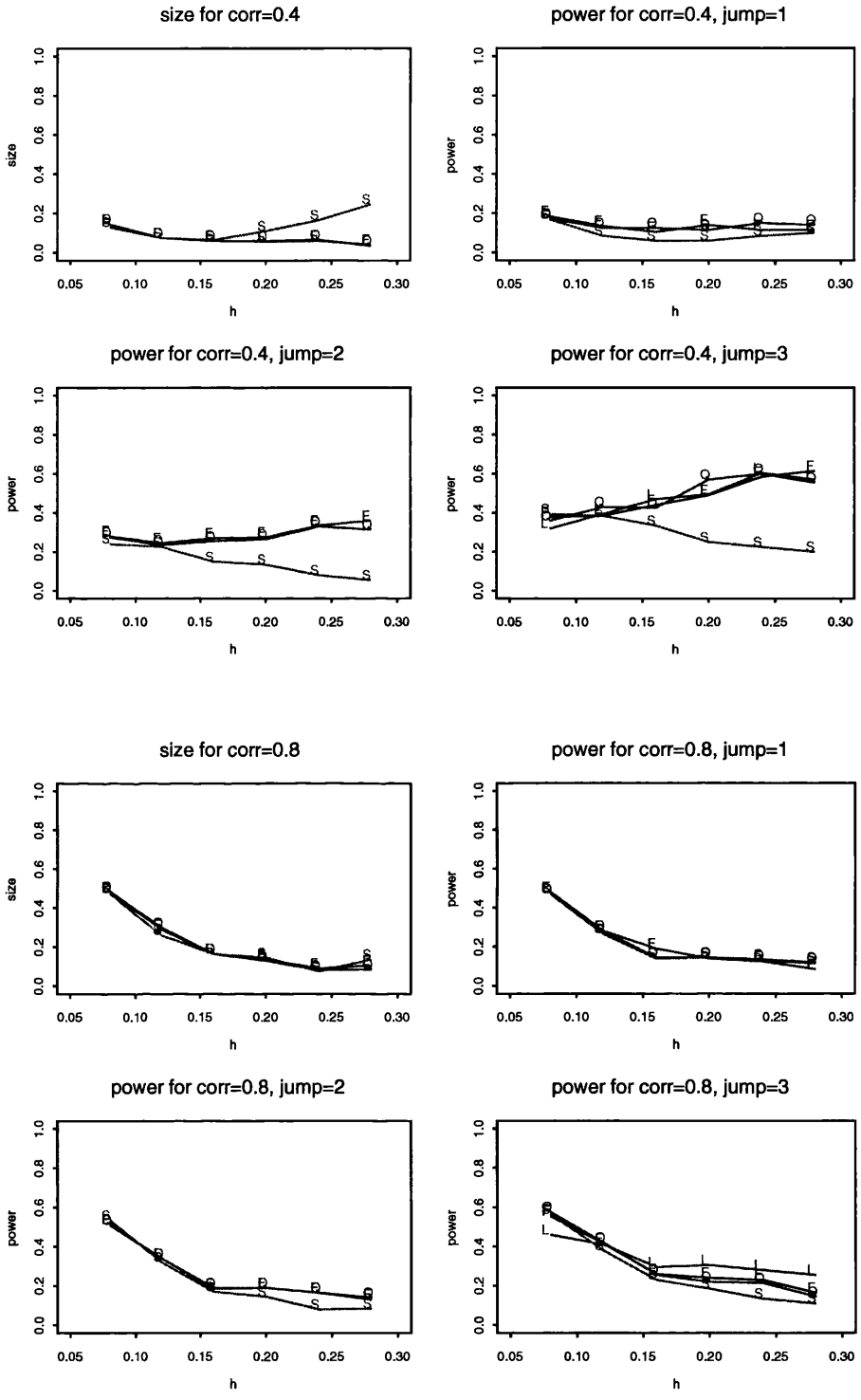

Figure 3.2: size (jump=0) and power (jump=1, 2 and 3) of the discontinuity test as function of the smoothing parameter, for flat (F), linear (L), quadratic (Q) and sine (S) trend, and with correlation = 0.4, 0.8.

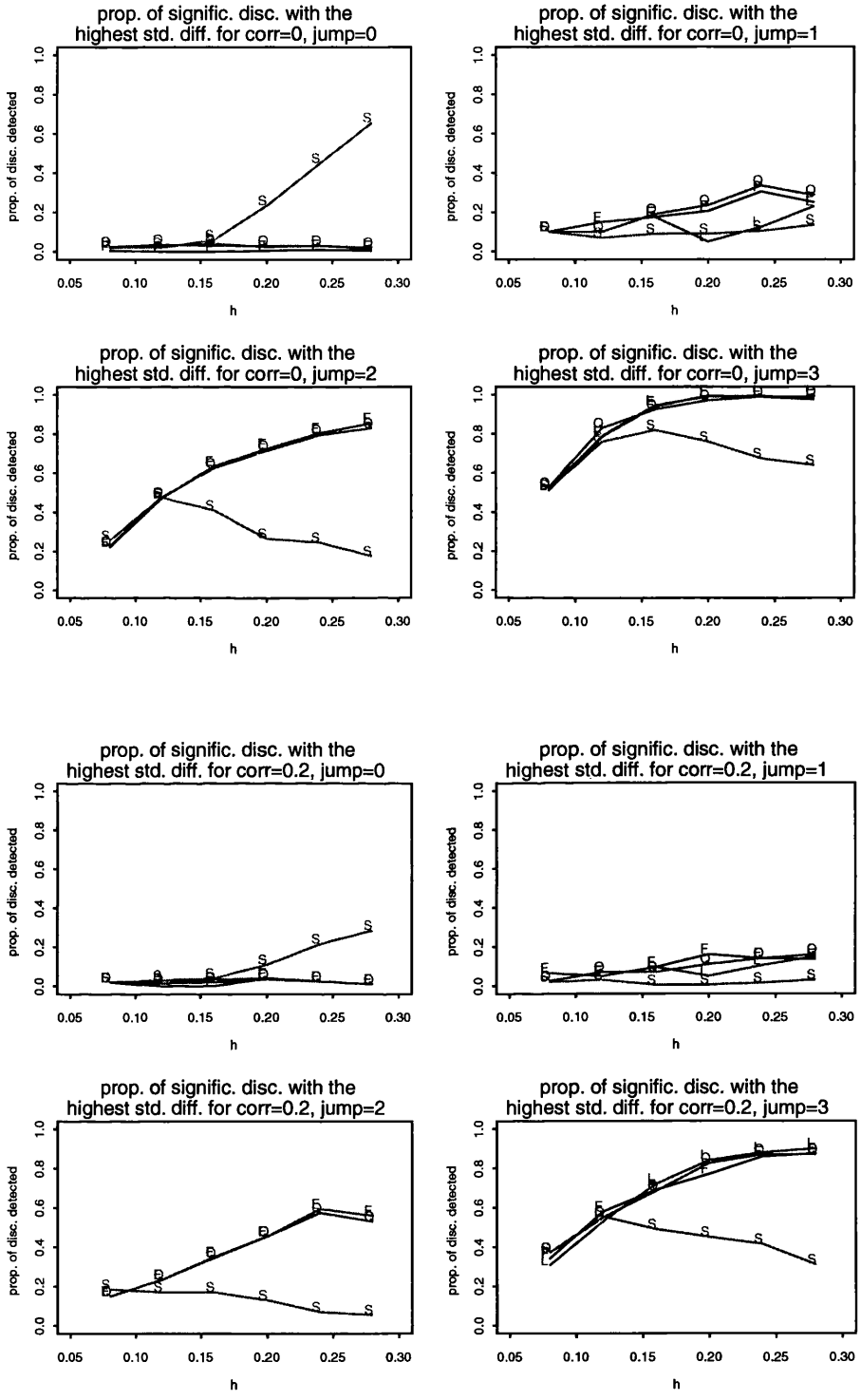

Figure 3.3: proportions of significant discontinuities detected with the highest standardized difference as function of the smoothing parameter for each type of trend (flat (F), linear (L), quadratic (Q) and sine (S)) with jump = 0, 1, 2 and 3, and for correlation = 0, 0.2.

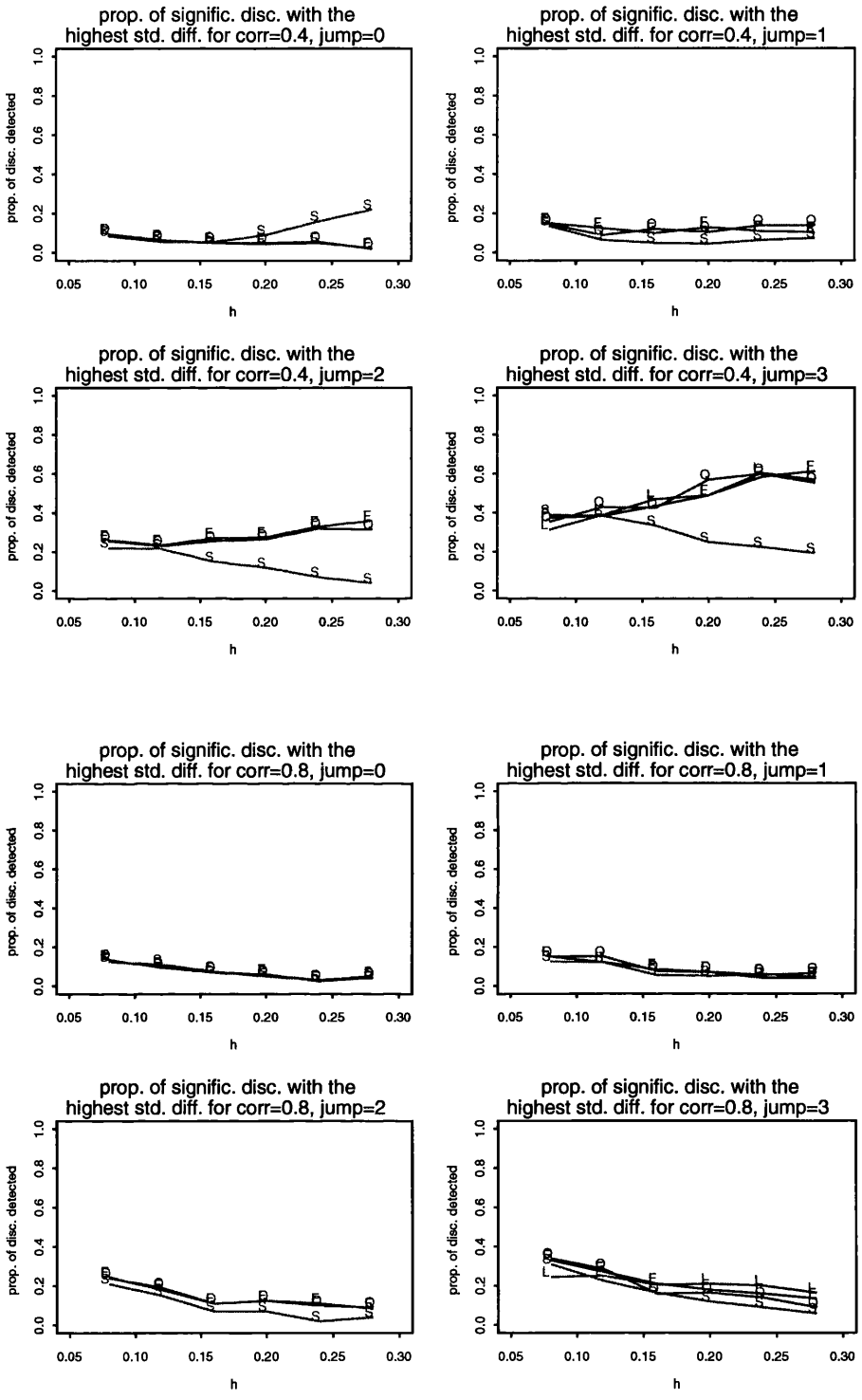

Figure 3.4: proportions of significant discontinuities detected with the highest standardized difference as function of the smoothing parameter for each type of trend (flat (F), linear (L), quadratic (Q) and sine (S)) with jump = 0, 1, 2 and 3, and for correlation = 0.4, 0.8.

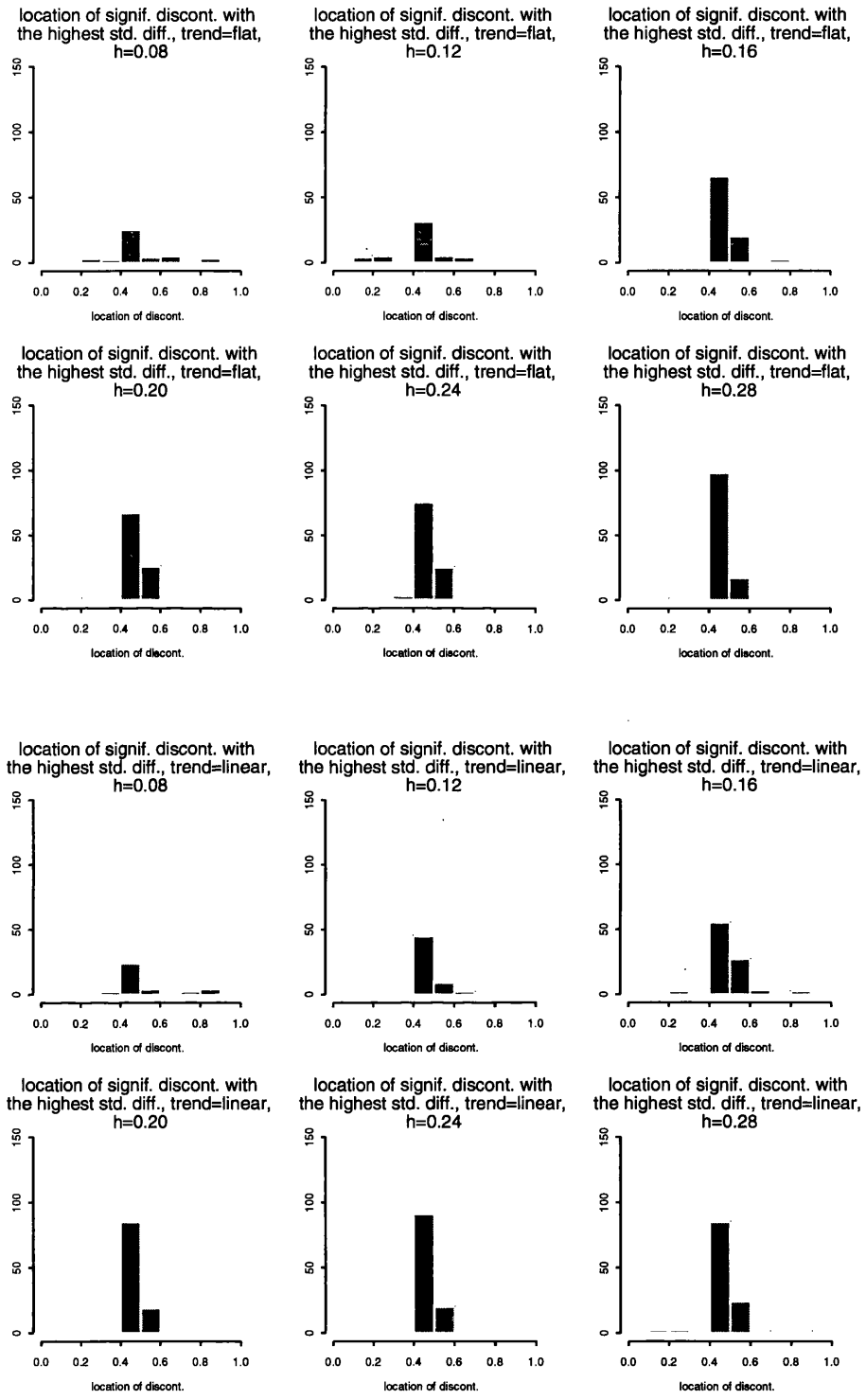

Figure 3.5: locations of significant discontinuities detected with the highest standardized difference of left minus right smoothers, for different values of smoothing parameter ($h = 0.08, 0.12, 0.16, 0.20, 0.24, 0.28$) and with flat and linear trends.

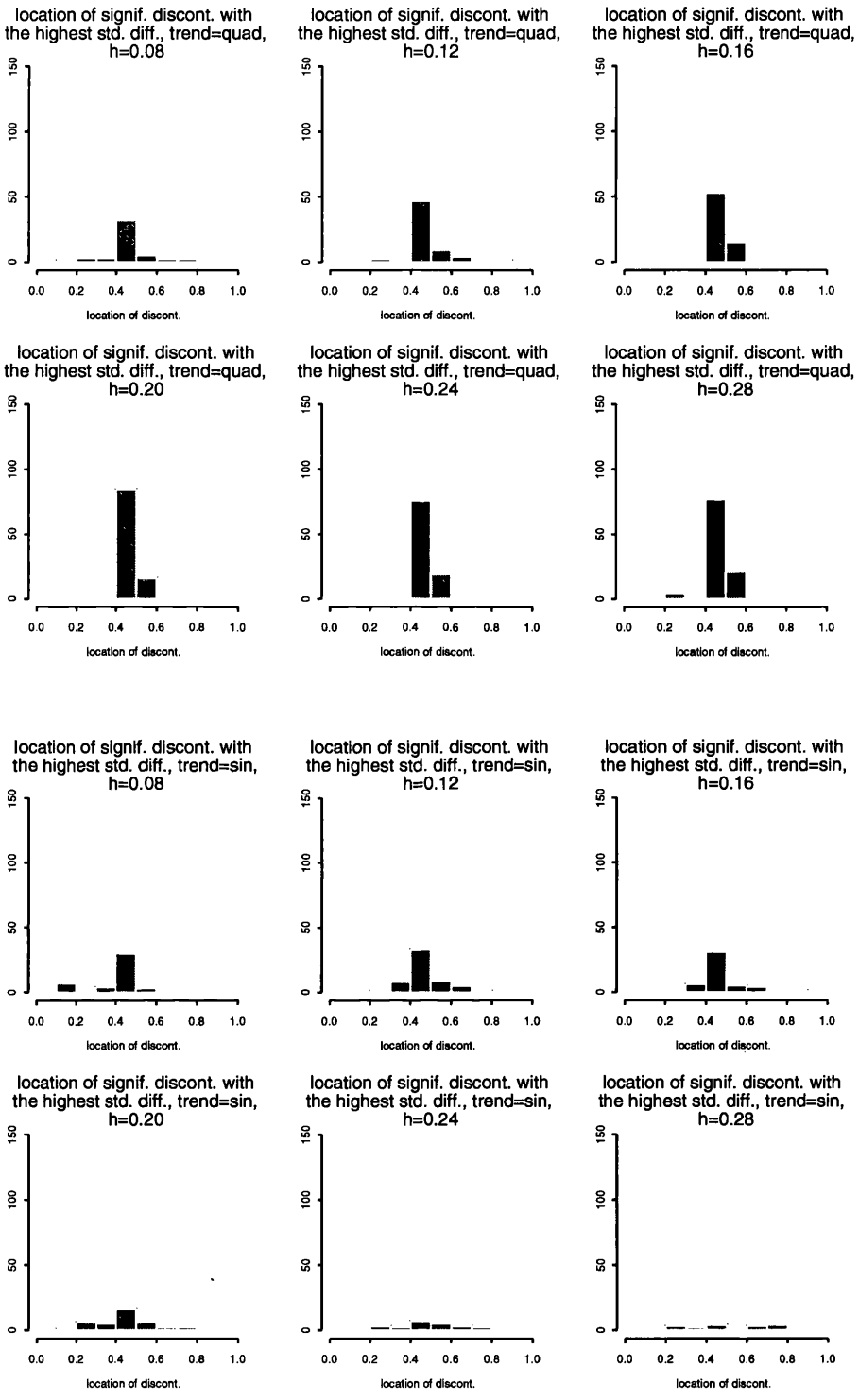

Figure 3.6: locations of significant discontinuities detected with the highest standardized difference of left minus right smoothers, for different values of smoothing parameter ($h = 0.08, 0.12, 0.16, 0.20, 0.24, 0.28$) and with quadratic and sine trends.

3.5 Applications and Results

The data that have been analyzed are the weekly means of the natural logarithm of the daily data for SO_2, SO_4 in air and precipitation, across Europe. As was shown in Section 1.4, the daily data are clearly skewed and show considerable variation, making the detection of discontinuities difficult. Therefore the tests have been applied to the weekly means of the logarithm.

In particular the analysis in 1.4, showed the presence of two kinds of seasonality, "days within the week" and "weeks within the year". As explained before, the first kind of seasonality has to be removed before computing the weekly means, because of the missing values, in order to avoid highly biased weekly means. The second kind of seasonality, is also removed because it could affect the results of the test. In fact, if a cycle is present in a time series, the test could detect as a discontinuity a change that is just due to the seasonal cycle.

Therefore, from the daily data, both kinds of seasonality were removed before computing the weekly means, and this has been done by applying a linear model, fitting the days of the week and the weeks within the year as factors, and then the de-seasonalised data have been used to obtain the weekly means. Fig.3.7 shows, the steps that have been followed before computing the weekly means.

In particular, Figure 3.7 a) shows the logarithm of the daily data, fitting a trend and the linear model that accounts for seasonality. Figure 3.7 b) shows the logarithm of the daily data after removing both seasonal components; the trend of the deseasonalised data is also plotted. Figure 3.7 c) and Figure 3.7 d) show the estimates of the "day within week" and "week within year" parameters. The days of the week and the weeks of the year are considered as factors and their

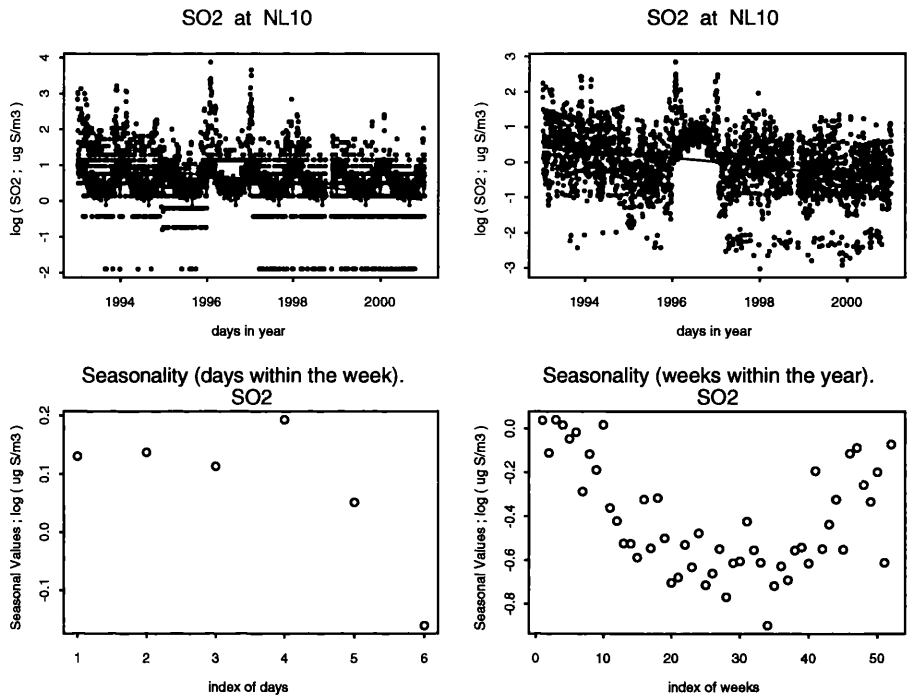

Figure 3.7: Analysis of seasonality for SO2 at Vreedepeel (NL10). a)log of the data; b)log of the deseasonalised data; c)estimates of the "day within week" parameters (1=Tue, 2=Wed,...,6=Sun); d)estimates of the "week within year" parameters.

contrast matrix includes each level as a dummy variable, excluding the first one. So the 6 values that are presented in the plot of Figure 3.7 c), represent the values for Tuesday, Wednesday, Thursday, Friday, Saturday and Sunday compared with Monday. Similarly, Figure 3.7 d), presents the values of the weeks of the year compared to week number 1. From these plots, the presence of a daily and of a weekly seasonality is apparent.

The data used for the discontinuity test were the de-seasonalised weekly means without removal of trend.

As a result of the edge bias in smoothers, fifty "testing points" at the start and fifty at the end of the series have been excluded. This means that all the points of the series are used to estimate the left and the right smoother, but only the observations in the central part of the series are used in the discontinuity test. The identification of discontinuities is based on several pieces of information. Firstly extremely helpful is the scatterplot of the left and the right smooth and of a shaded region which is bounded by the curves obtained by adding to the average of g_l and g_r plus or minus one and half (estimated) standard deviations of $g_l - g_r$. This is referred to as a "reference band" and it gives a guide to places where discontinuities may be found, because if both left and right smooth leave the shaded region, then they are separated by more than three standard deviations, suggesting the possible presence of a "change point". Since the reference bands are pointwise bands, the choice of three standard deviations gives an informal protection against the multiple comparisons problem. Figure 3.8 shows the same case of Figure 3.7 (SO_2 Vreedepeel, NL10), with the reference band and relative dates (expressed in terms of "week/year") of where the discontinuities have been detected. The bold line marks the most significant discontinuity detected.

For a better understanding of the change in the mean level, another kind of graph is presented in Figure 3.9. The entire time series is divided into sections lying between the identified change-points, and the trend in each separate section is plotted. In diagnosing discontinuities, a third output of the analysis is represented by the list of the positions of the points whose left and right smooths are more than three standard deviations apart, and the difference between left minus right smoothers.

From Table 3.3 and from the pictures introduced before, it is possible to

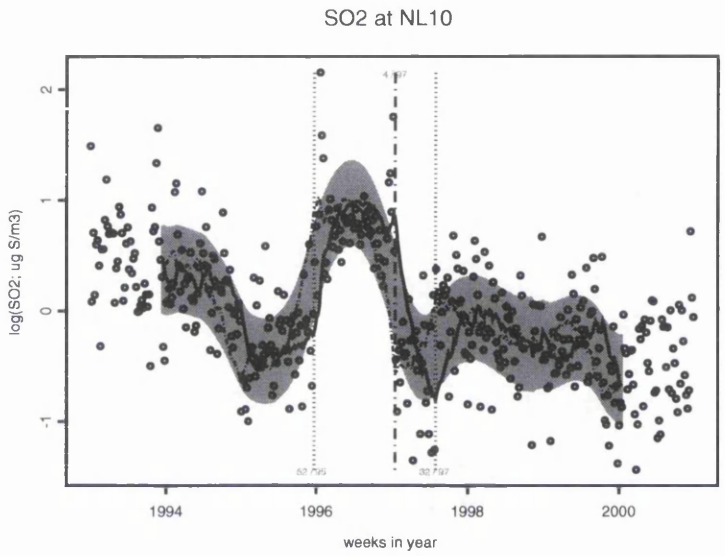

Figure 3.8: Discontinuities for SO_2 at Vreedepeel (NL10).

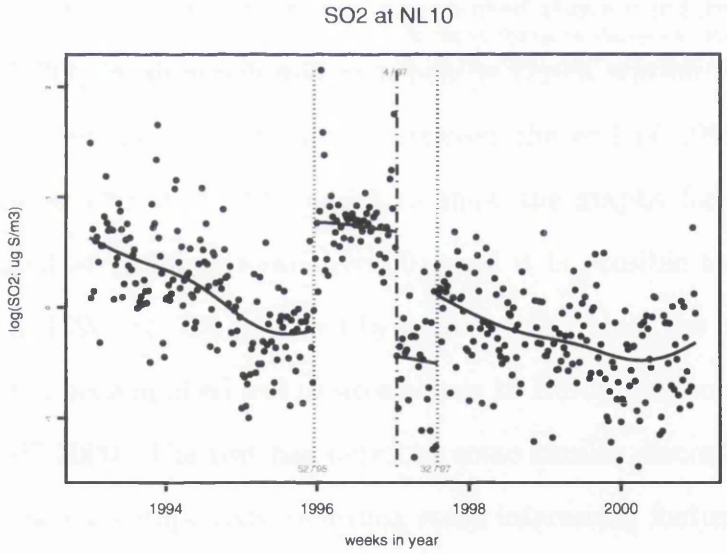

Figure 3.9: smoothing the sub-trends at the discontinuities detected for SO_2 at Vreedepeel (NL10).

Site	Compound	Week	Year	Jump
NL09	SO2	52	1995	-1.850
NL09	SO2	53	1996	1.561
NL09	SO4air	40	1995	-0.671
NL09	SO4air	17	1996	0.626
NL09	SO4prec.	15	1994	-1.012
NL09	SO4prec.	53	1994	-0.995
NL10	SO2	52	1995	-1.111
NL10	SO2	4	1997	1.402
NL10	SO2	32	1997	-0.928
NL10	SO4air	6	1997	0.855

Table 3.3: Discontinuities detected at Kollumerwaard (NL09) Vreedepeel (NL10).

have an idea of the presence of discontinuities across the stations in a country, and on the basis of this analysis it is possible to compare the "common national discontinuities". For example, for the case presented above it is interesting to note how SO_2 and SO_4 in air monitored at a nearby Dutch station Kollumerwaard (NL09), showed similar discontinuities, between the end of 1995 and the end of 1996. Figures 3.10, 3.11, 3.12 and 3.13 show the graphs for SO_2 and SO_4 in air monitored at Kollumerwaard (NL09), and it is possible to note how the observations in 1996 are characterized by higher values than the others.

This test has been applied to 113 sites across 16 European countries, covering the period 1977-2000. The test has detected some similar discontinuities across countries and across compounds, revealing some interesting features of the data. Experts suggested that some of the discontinuities may be due to change in meteorology. The idea is that modeling pollutants accounting for meteorology would eliminate those changes in trend.

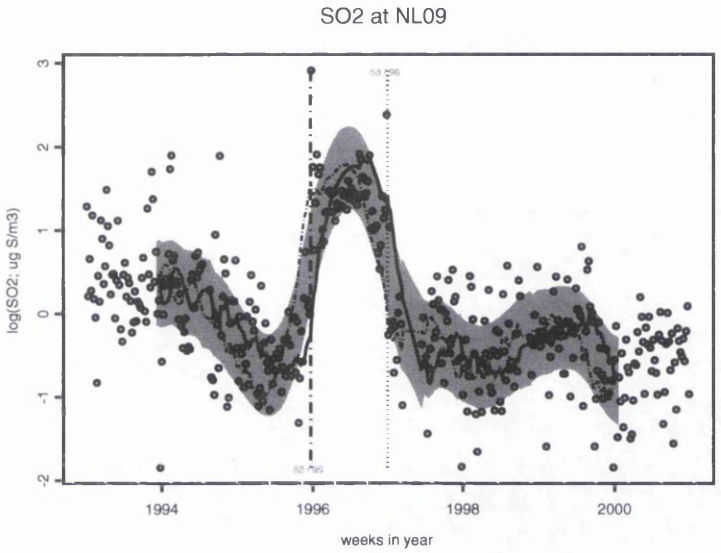

Figure 3.10: Discontinuities for SO_2 at Kollumerwaard (NL09).

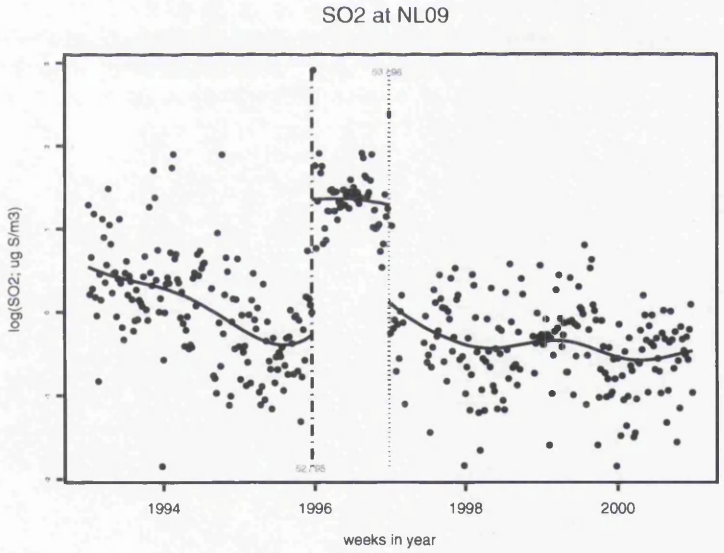

Figure 3.11: smoothing the sub-trends at the discontinuities detected for SO_2 at Kollumerwaard (NL09).

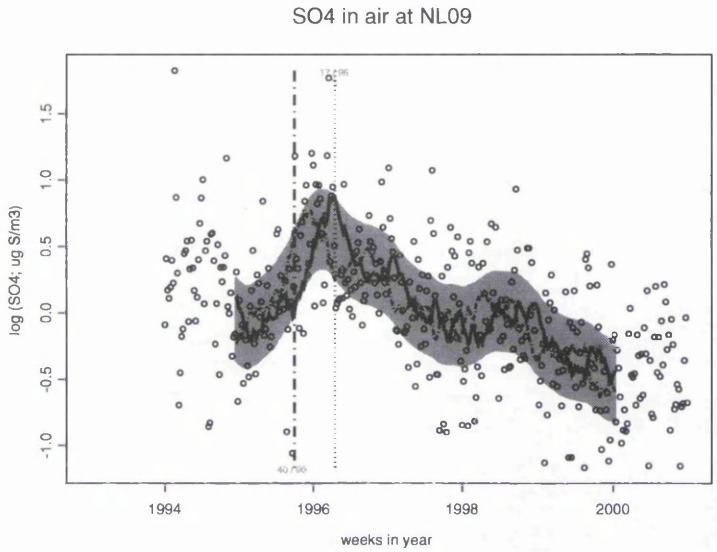

Figure 3.12: Discontinuities for SO_4 in air at Kollumerwaard (NL09).

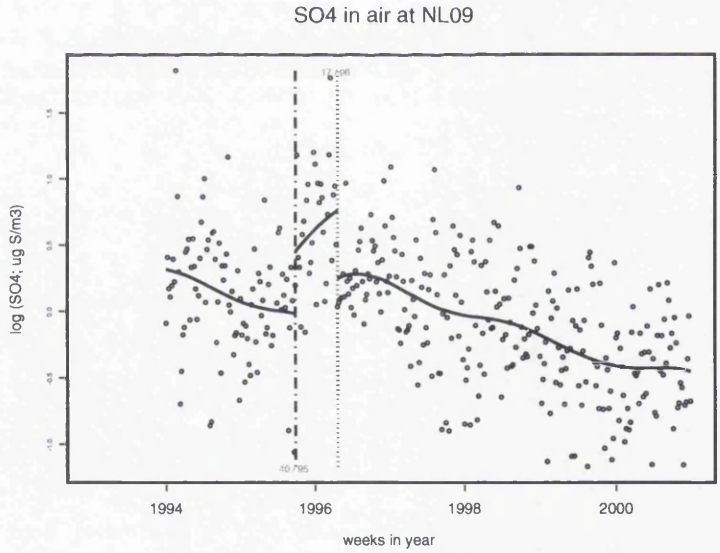

Figure 3.13: smoothing the sub-trends at the discontinuities detected for SO_4 in air at Kollumerwaard (NL09).

Chapter 4

Modeling With Correlated Errors

This chapter presents an extension of methodologies proposed in Chapter 2 in order to account for correlated error. Air pollution concentrations are typically affected by correlation, and if a model is fitted assuming independent errors, serious dangers could affect the models' estimation and selection. Univariate and bivariate smoothers that account for correlations will be introduced. A reformulation of the backfitting algorithm (Hastie and Tibshirani, 1990), that has as output the projection matrix of the entire additive model at convergence will be presented. Tests for additive models' selection, and for comparing components across models, are also presented. Simulation studies will show the performances of the proposed methodologies.

4.1 Univariate smoothing with correlated errors

Local linear regression, as presented in section 2.1, is based on the assumption that the errors are independent. It is clear that this assumption does not always

apply, especially for air pollution time series data. Opsomer et al. (2001) reviewed the existing literature in kernel regression, smoothing splines, and wavelet regression in the presence of correlation, both for short-range and long-range dependence. They address the problems that the presence of correlation can create with smoothing parameter selection methods, such as cross-validation or plug-in. They propose data-driven smoothing parameter selection techniques which apply to correlated data. However in their work, Opsomer et al. (2001) used the standard focus of nonparametric regression and did not consider the possibility of defining new smoothers that account for correlation.

McMullan et al. (2005) considered the problem of nonparametric estimation in additive models when the errors terms are correlated. In the context of the simple model

$$y_i = m(x_i) + \varepsilon_i, \tag{4.1}$$

the vector of errors ε is assumed to have variance matrix $\sigma^2 V$, where V is a correlation matrix. McMullan et al. (2005) suggested moving to a scale where a model with independent errors could be applied, through the transformation $z = K^{-1}y$, where the correlation matrix V has the Cholesky decomposition $V = KK^T$. Model (4.1) can then be written as $z = \tilde{m}(x) + \eta$, where the elements of the error vector η are now independent. The regression function $\tilde{m}$ is equivalent to $K^{-1}m$. Examination of the structure of this term, and specifically the hypothesis $m = 0$, can be examined on the new model scale. However, estimates of m by back-transformation from $\tilde{m}$ can be problematic in the absence of conditions on K^{-1} which guarantee smoothness. This procedure seems to behave well with a

small number of covariates (around 2 or 3). However when the number of covariates increases, problems appear. In fact, when the estimates of each component are produced, and multiplied back by the K matrix obtained from the Cholesky decomposition, the resulting estimates do not have a smooth pattern, and are characterized by high variability.

The methodology proposed here instead resolves the problem of correlation at the first stage of analysis. It produces local linear regression smoothers that do not assume independence of the errors. To obtain these smoothers it is necessary to recall equation (2.10). This immediately suggests a local least squares criterion which incorporates the correlation structure directly as:

$$\min_{\alpha,\beta}\{y-\alpha 1_n - x\beta\}^T\{K^{-1}\}^T W\{K^{-1}\}\{y-\alpha 1_n - x\beta\} \tag{4.2}$$

where K is obtained from the Cholesky decomposition, $V = KK^T$, where V indicates the correlation matrix. The form (4.2) emphasises the connection with the transformation approach of McMullan et al. (2005).

In expression (4.2), it has been assumed that the correlation matrix V is known. In practice, it will often be required to estimate this. As described in section 3.3.2, the approach of Niu (1996) will be followed. This consists of fitting an independence model and using the residuals from this to identify a suitable structure for the error component.

An explicit expression can be derived from criterion (4.2) as

$$\alpha_x = \frac{2[(\sum_{i,j}^{n} v_{ij}w_j(y_i+y_j)](\sum_{i,j}^{n} x_i v_{ij} w_j x_j) - [\sum_{i,j}^{n} v_{ij}w_j(x_i y_j + x_j y_i)][\sum_{i,j}^{n} v_{ij}w_j(x_j+x_i)]}{4(\sum_{i,j}^{n} v_{ij}w_j x_i x_j)(\sum_{i,j}^{n} v_{ij}w_j) - [\sum_{i,j}^{n} v_{ij}w_j(x_j+x_i)]^2} \tag{4.3}$$

where v_{ij} indicates the (i,j)th element of the inverse correlation matrix V^{-1}.

The local constant estimator is a special case of the local linear estimator, and its explicit representation follows as:

$$\hat{m}(x) = \frac{\sum_j \sum_i v_{ij}w_j(y_i+y_j)}{2\sum_j \sum_i v_{ij}w_j}. \tag{4.4}$$

Standard error bands for the smoother estimates $\hat{y}$ can be obtained by computing their standard errors, given by:

$$s.e.(\hat{y}) = \sqrt{var(Sy)} = \sqrt{diag(SVS^T)\sigma^2} \tag{4.5}$$

where $var(y) = \sigma^2 V$. A suitable estimate of the variance of the detrended y is obtained by expression (3.14), and recalled here as

$$\hat{\sigma}^2 = \frac{y^T \Gamma y}{\mathrm{tr}(\Gamma V)} \tag{4.6}$$

where $\hat{m} = Sy$ is a nonparametric estimate of the trend, S is the smoothing matrix, and $\Gamma = \frac{(I-S)^T V^{-1}(I-S)}{n-1}$.

Notice that the effect of bias in the estimation of the mean component of the model has a conservative effect by inflating the estimate of σ^2.

4.2 Bivariate smoothing with correlated errors

In the case of two predictors (x_1, x_2), the least squares problem (4.2) can be extended to the bivariate case by solving the following weighted least squares problem:

$$\min_{\alpha,\beta,\gamma}\{y - \alpha 1_n - x_1\beta - x_2\gamma\}^T\{K^{-1}\}^T W\{K^{-1}\}\{y - \alpha - x_1\beta - x_2\gamma\} \qquad (4.7)$$

The solution of equation (4.7), is given by:

$$(X^T V^{-1} W X)^{-1} X^T V^{-1} W y \qquad (4.8)$$

where X is the $n \times 3$ matrix whose ith row is $\{1, (x_{1i} - z_1), (x_{2i} - z_2)\}$, W is the diagonal matrix whose (i,i)th element is $w_i = w_1(x_{1i} - z_1; h_1)w_2(x_{2i} - z_2; h_2)$, and z_1 and z_2 are the points where the estimates are computed.

The local linear estimate is defined by the first element of the vector (4.8). The elements of the 3×3 matrix $A = (a_{ij}) = (X^T V^{-1} W X)$ are all of the form

$$\sum_{ij} v_{ji} w_{1j}(x_{1j} - z_1; h_1) w_{2j}(x_{2j} - z_2; h_2)(x_{1i} - z_1)^{r_1}(x_{2i} - z_2)^{r_2}(x_{1j} - z_1)^{s_1}(x_{2j} - z_2)^{s_2}$$

where $r_1 + s_1 + r_2 + s_2 \leq 2$, where v_{ji} are the elements of the correlation matrix V^{-1}. To obtain the first element of the least squares solution, we need only the first row of $(X^T V^{-1} W X)^{-1}$, denoted by (b_1, b_2, b_3). By applying standard linear algebra results, reported for instance by Healy (1986) (section 3.4) these can be

written as:

$$\begin{aligned} b_1 &= 1\Big/\Big(a_{11} - \frac{1}{d}\{(a_{12}a_{33} - a_{13}a_{23})a_{12} + (a_{13}a_{22} - a_{12}a_{23})a_{13}\}\Big) \\ b_2 &= \frac{b_1}{d}(a_{13}a_{23} - a_{12}a_{33}) \\ b_3 &= \frac{b_1}{d}(a_{12}a_{23} - a_{13}a_{22}) \end{aligned}$$

where $d = a_{22}a_{33} - a_{23}^2$. Multiplying the vector (b_1, b_2, b_3) by $(X^T V^{-1} W)$, the result is a vector of length n, whose ith element is

$$\begin{aligned} & b_1 \sum_j v_{ij} w_{1i}(x_{1i} - z_1; h_1) w_{2i}(x_{2i} - z_2; h_2) \\ + \; & b_2 \sum_j v_{ij} w_{1i}(x_{1i} - z_1; h_1) w_{2i}(x_{2i} - z_2; h_2)(x_{1j} - z_1) \\ + \; & b_3 \sum_j v_{ij} w_{1i}(x_{1i} - z_1; h_1) w_{2i}(x_{2i} - z_2; h_2)(x_{2j} - z_2) \end{aligned}$$

The inner product of this vector with y produces the local linear estimates at (x_1, x_2).

The local constant estimator is a special case of the local linear estimator, and as for the univariate case is defined by:

$$\hat{m}(x_1, x_2) = \frac{\sum_j \sum_i v_{ij} w_j (y_i + y_j)}{2 \sum_j \sum_i v_{ij} w_j} \tag{4.9}$$

4.3 Deriving the smoothing matrix in the backfitting algorithm

As has been explained in Chapter 2, the fit of a model with more than two covariates can be tackled by additive models (Hastie and Tibshirani, 1990), expressed in equation (2.16), and recalled here as

$$y_l = \alpha_l + \sum_{j=1}^{p} m_j(x_{jl}) + \varepsilon_l, \quad l = 1, \ldots, n. \tag{4.10}$$

On the basis of the results of Sections 4.1 and 4.2, it is now possible to fit an additive model whose building blocks are the univariate and bivariate local linear regression smoothers that account for correlation of the errors ε.

Recalling the definition of the backfitting algorithm (Hastie and Tibshirani, 1990) presented in Chapter 2, it is possible to note that the estimated $\hat{y}$ are updated in such a way that at step (i) the $\hat{y}^{(i)}$ are "projected" from the observations y using a hat matrix $S^{(i)}$ that differs from the one used at the previous step $S^{(i-1)}$. This means that, when the algorithm converges, the definition of a projection matrix S is not straightforward.

Hastie and Tibshirani (1990) proposed an algorithm for computing approximate versions of the projection matrix P_j (using notation of Hastie and Tibshirani (1990), section 5.4.4, P_j corresponds to $\mathbf{R}_j$) that at convergence give $\hat{f}_j = P_j y$. This algorithm consists in applying the backfitting procedure to each of the n unit n-vectors that are the columns of I_n, the $n \times n$ identity matrix. The result of backfitting applied to the ith unit vector produces fitted vectors $\hat{f}_j^i, j = 1, \ldots, p$

where $\hat{f}_j^i$ is the ith column of P_j. Similarly, $\hat{f}_+^i$ is the ith column of P. The degrees of freedom for error are $df^{err} = n - tr(2P - PP^T)$. For model comparisons, the change in the error degrees of freedom $\triangle df^{err}$ due to an individual term is required. Let $P_{(j)}$ denote the operator that produces the additive fit with the jth term removed, then it is possible to define df_j^{err}, the degrees of freedom for error due to the jth term:

$$df_j^{err} = tr(2P - PP^T) - tr(2P_{(j)} - P_{(j)}P_{(j)}^T)$$

This is the expected increase in the residual sum of squares (up to a scale factor) if the jth predictor is excluded from the model, assuming its exclusion does not increase the bias. It is immediately understandable that these definitions are not attractive from a computational point of view. Each $P_{(j)}$ is obtained by applying the backfitting algorithm $n \times n$ times, each of which needs an undefined number of iterations before converging.

A different methodology is proposed here to compute an approximate version of the P_j matrix. This methodology is computationally less expensive, because it consists in updating and storing the projection matrix at each step of the backfitting algorithm. For the simplest case of two variables, it is possible to write the first two steps of the backfitting algorithm in matrix form as follows:

1. first step:

$$\begin{aligned} \hat{m}_1^{(1)} &= (I - N)S_1 y \\ \hat{m}_2^{(1)} &= (I - N)S_2[I - (I - N)S_1]y \\ &= (I - N)S_2[I - P_2^{(1)}]y \end{aligned}$$

2. second step:

$$\begin{aligned}
\hat{m}_1^{(2)} &= (I-N)S_1[I-(I-N)S_2(I-P_2^{(1)})]y \\
&= (I-N)S_1[I-P_1^{(2)}]y \\
\hat{m}_2^{(2)} &= (I-N)S_2[I-(I-N)S_1(I-P_1^{(2)})]y \\
&= (I-N)S_2[I-P_2^{(2)}]y
\end{aligned}$$

where the N matrix is an $n \times n$ matrix whose elements are $\frac{1}{n}$. Using induction, it is possible to note that the projection matrix for variable j, "updated" at iteration (i) is obtained from the following formula:

$$P_j^{(i)} = (I-N)S_j \left[I - \sum_{k>j} P_k^{(i-1)} + \sum_{k<j} P_k^{(i)} \right]$$

where, initializing $P_j^{(0)} = I$, at convergency step (i), it is possible to obtain $\hat{m}_j^{(i)} = P_j^{(i)}y$. In order to ensure unique definitions of the estimators, the intercept term can be held at $\hat{\alpha} = \bar{y}$, the sample mean, throughout and additional adjustment to ensure that $\sum_l \hat{m}_j^{(i)}(x_{jl}) = 0$ for each j, can be applied at each step.

Therefore, once the algorithm has converged at iteration (i), it is simply necessary to sum together the estimates of each component $\hat{m}_j^{(i)}, j = 1, \ldots, p$, and $\hat{\alpha}$ to obtain $\hat{y}$. In other words

$$\hat{y} = \bar{y} + (P_1 + P_2 + \ldots + P_p)y = (N + P_1 + P_2 + \ldots + P_p)y = Py. \qquad (4.11)$$

Having obtained the hat matrix P for the additive model, it is now possible to compute the residual sum of squares and degrees of freedom, in order to compute

the F test statistic. However the RSS and df proposed in chapter 2 are based on the assumption of uncorrelated data. Therefore, in the next section their generalizations to the correlated case are presented.

4.4 Testing models with correlated errors

4.4.1 Testing linear models with correlated errors

Given the linear model $y = Xb + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2 V)$, where V is a known $n \times n$ positive definite matrix, it is possible to write V, using the Cholesky decomposition, as a function of a non-singular $n \times n$ matrix K, such that $V = KK^T$. Therefore setting $z = K^{-1}y, W = K^{-1}X, \eta = K^{-1}\varepsilon$, it is possible to write the model $y = Xb + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 V)$, as $z = Wb + \eta$, $\eta \sim N(0, \sigma^2 I)$, where W is a $n \times p$ matrix of rank p (Seber, 1977). At this point, it is possible to write the results obtained in Section 2.2.1 firstly for W and z, and then in terms of X and y. So the residual sum of squares can be written as follows:

$$\begin{aligned} RSS &= z^T[I - W(W^TW)^{-1}W^T]z \\ &= (K^{-1}y)^T[I - K^{-1}X[X^T(KK^T)^{-1}X]^{-1}(K^{-1}X)^T]K^{-1}y \\ &= y^T[V^{-1} - V^{-1}X[X^TV^{-1}X]^{-1}X^TV^{-1}]y \\ &= y^TV^{-1}[I - H]y \end{aligned}$$

Regarding the degrees of freedom, indicating with $r(.)$ the rank of a matrix, it is possible to use the expression:

$$E[RSS] = tr[I - W(W^TW)^{-1}W^T]I\sigma^2 + b^TW^T[I - W(W^TW)^{-1}W^T]Wb$$

$$
\begin{aligned}
&= r[I - W(W^TW)^{-1}W^T]\sigma^2 \\
&= [N - r(W)]\sigma^2 \\
&= [N - r(K^{-1}X)]\sigma^2 \\
&= [N - r(X)]\sigma^2
\end{aligned}
$$

and therefore, $df = E[RSS]/\sigma^2 = [N - r(X)]$. At this point, all the components that are needed to compute the F statistic for linear models have been defined. In the same way that this section extends the RSS and df to the correlated case for the linear models, the next section will extend the RSS and the df definitions of the nonparametric models to the correlated case.

4.4.2 Testing nonlinear models with correlated errors

4.4.2.1 Approximate F test with correlated errors

On the basis of the linear case with correlated errors it is possible to define the residual sum of squares and the degrees of freedom in order to compute an approximate F statistic that accounts for correlation. For the residual sum of squares, it is possible to write:

$$RSS = (y - Sy)^TV^{-1}(y - Sy) = y^T(I - S)^TV^{-1}(I - S)y \tag{4.12}$$

and bearing in mind Theorem 1 of Section (2.2.1), it is possible to write:

$$
\begin{aligned}
E[RSS] &= tr[(I - S)^TV^{-1}(I - S)V\sigma^2] + \mu^TS^T[(I - S)^TV^{-1}(I - S)]S\mu \\
&= tr[I - V^{-1}SV - S^T + S^TV^{-1}SV]\sigma^2 + \mu^TS^T[(I - S)^TV^{-1}(I - S)]S\mu
\end{aligned}
$$

and therefore, $df_{err.c.} = n - tr[V^{-1}SV + S^T - S^TV^{-1}SV]$. Always working in the context of correlated errors, the definition $tr(S)$, as the "effective number of parameters" of a smoother, still holds, but a small change is necessary for the degrees of freedom for variance. In fact it now becomes $\mathrm{tr}(SVS^T)$. If s_i is the ith row of the smoothing matrix S then it follows that the summed variances of the fitted values are given by:

$$\begin{aligned}
\sum_i var(\hat{m}(x_i)) &= \sum_i var(s_i y) \\
&= \sum_i var(s_i m + s_i \varepsilon) \\
&= \sum_i s_i V(s_i)^T \sigma^2 \\
&= \mathrm{tr}(SVS^T)\sigma^2 = \mathrm{tr}(S^TVS)\sigma^2.
\end{aligned}$$

All the definitions of degrees of freedom obtained so far, are listed below:

$$\begin{aligned}
df_{par.} &= tr(S) \\
df_{var.} &= tr(SS^T) \\
df_{var.c.} &= tr(SVS^T) \\
df_{err.} &= n - tr(2S - SS^T) \\
df_{err.c.} &= n - tr[V^{-1}SV + S^T - S^TV^{-1}SV] \qquad (4.13)
\end{aligned}$$

where $df_{par.}, df_{var.}, df_{err.}$ are the ones that don't account for correlation, while $df_{var.c.}, df_{err.c.}$ do.

4.4.2.2 Pseudo Likelihood Ratio test with correlated errors

The Pseudo Likelihood Ratio Test expressed in section 2.2.2.2 can account for correlation simply by amending the formulation for the RSS. In fact, given the expression

$$F = \frac{RSS_r - RSS_f}{RSS_f}$$

the F statistic can be expressed in terms of quadratic forms in a way that accounts for correlation using the following expression for the residual sums of squares:

$$RSS_r = y^T(I - S_r)^T V^{-1}(I - S_r)y; \quad RSS_f = y^T(I - S_f)^T V^{-1}(I - S_f)y$$

where S_r and S_f denote the smoothing matrices for the reduced and full models respectively. Expressing the F statistic as the ratio of quadratic forms in Normal random variables with means approximately zero and the same variance, it is possible to write:

$$F = \frac{y^T B y}{y^T A y}$$

where A is the matrix $(I-S_f)^T V^{-1}(I-S_f)$ and B is the matrix $(I-S_r)^T V^{-1}(I-S_r) - (I - S_f)^T V^{-1}(I - S_f)$. Results from Johnson and Kotz (1972) can now be applied in the same way as summarized in Section 2.2.2.1. The only difference is that the correlation matrix V is no longer the identity matrix I.

4.4.3 Comparing components of Additive Models with Correlated Errors

In this section two tests for comparing components of two different additive models, that account for correlation, are presented. Fitting two additive models of the following form:

$$y = \alpha_1 + m_x(x) + m_z(z) + \varepsilon_1 \tag{4.14}$$

$$y = \alpha_2 + m_x(x) + \varepsilon_2 \tag{4.15}$$

and indicating by $\hat{m}_{x,1}$ and $\hat{m}_{x,2}$ the estimates for m_x of model (4.14) and model (4.15) respectively, interest could be addressed in testing the hypothesis that the estimate $\hat{m}_{x,1}$ is equal to the estimate $\hat{m}_{x,2}$. Bowman and Azzalini (1997) proposed a statistic for comparing regression curves, based on:

$$\tilde{F} = \frac{(\hat{m}_{x,1} - \hat{m}_{x,2})^T(\hat{m}_{x,1} - \hat{m}_{x,2})}{\hat{\sigma}^2} \tag{4.16}$$

where $\hat{\sigma}^2$ denotes an estimate of the error variance σ^2. The formulation of the test proposed by Bowman and Azzalini (1997) is based on the assumption of independent errors and more details have been given in Section 2.2.3. Here two generalizations of the test statistic (4.16) that account for the correlation of the data are proposed.

A first formulation of the test is based on the idea of the approximate F test, and is given by the following formulation:

$$\tilde{F}_{A.corr} = \frac{y^T(P_{x,1} - P_{x,2})^T V^{-1}(P_{x,1} - P_{x,2})y/df^*_{corr}}{y^T(I - S_1)^T V^{-1}(I - S_1)y/df_{corr}} \tag{4.17}$$

where $P_{x,1}$ is the smoothing matrix that gives the smooth estimates $\hat{m}_{x,1} = P_{x,1}y$, similarly $P_{x,2}$ is the smoothing matrix that gives the smooth estimates $\hat{m}_{x,2} = P_{x,2}y$, and S_1 is the smoothing matrix of the overall model (4.14), that produces the estimates $\hat{y} = S_1 y$. The numerator of expression (4.17) consists of the sum of squares of the differences between the estimates $\hat{m}_{x,1}$ and $\hat{m}_{x,2}$. The denominator of expression (4.17) consists in the estimate of the variance of y.

The degrees of freedom used in the denominator (df_{corr}) of expression (4.17) are given in the following formula:

$$df_{corr} = tr[V^{-1}S_1V + S_1^T - S_1^TV^{-1}S_1V] \tag{4.18}$$

and the degrees of freedom used in the numerator (df^*_{corr}) are obtained by:

$$df^*_{corr} = tr[(P_{x,1} - P_{x,2})^TV^{-1}(P_{x,1} - P_{x,2})V] \tag{4.19}$$

The presence of bias in the residual sums of squares and the absence of the required properties in the underlying projection matrices mean that the test statistic will not follow an F distribution under the null hypothesis. However, comparing $\tilde{F}_{A.corr}$ to an F distribution with df^*_{corr} and df_{corr} degrees of freedom does provide a helpful benchmark.

A second formulation of the test is based on the quadratic form test, and is given by the following formula:

$$\tilde{F}_{L.corr} = \frac{y^T(P_{x,1} - P_{x,2})^TV^{-1}(P_{x,1} - P_{x,2})y}{y^T(I - S_1)^TV^{-1}(I - S_1)y} = \frac{y^TQy}{y^TBy} \tag{4.20}$$

Having expressed the $\tilde{F}_{L.corr}$ as a ratio of quadratic form, it is now possible to

apply the theory of the Pseudo Likelihood Ratio Test presented in Section 4.4.2.

Once both tests have been implemented, a graphical display of the testing can be obtained by drawing a standard error reference band for the difference of the smoothers $\hat{m}_{x1}$ and $\hat{m}_{x2}$, given by:

$$\begin{aligned} s.e.(\hat{m}_{x,1} - \hat{m}_{x,2}) &= \sqrt{var\{(P_{x,1} - P_{x,2})y\}} \\ &= \sqrt{diag\{(P_{x,1} - P_{x,2})V(P_{x,1} - P_{x,2})^T\}\sigma^2} \end{aligned} \quad (4.21)$$

where $var(y) = \sigma^2 V$, and an estimate for σ^2 can be obtained by expression (3.14).

4.4.4 Tests for no effect with correlated errors

This section presents two tests for assessing the presence of trends accounting for correlation of the errors, generalizing those presented in Section 2.2.4. Given *Model* 0 and *Model* 1, the purpose of the tests is in testing any effect of x on y accounting for any correlation present in y.

$$\begin{aligned} Model\ 0: \ y &= \mu + \varepsilon \\ Model\ 1: \ y &= m(x) + \varepsilon \end{aligned}$$

Model 1 and *Model* 0 will be tested using the same formulations expressed in equations (2.34) and (2.35) of section 2.2.4, and here recalled as:

$$F = \frac{(RSS_0 - RSS_1)/(df_1 - df_0)}{RSS_1/(n - df_1)} \quad (4.22)$$

$$F = \frac{RSS_0 - RSS_1}{RSS_1} \quad (4.23)$$

where RSS_0, df_0 and RSS_1, df_1 indicate respectively the residual sum of squares and the degrees of freedom for the reduced model, *Model* 0, and for the full model, *Model* 1. In order to account for correlation of the errors, the RSS and the df used will be the ones that account for correlation. The RSS is the generalized version given below:

$$RSS_i = y^T(I - S_i)^T V^{-1}(I - S_i)y; \quad i = 0, 1$$

where $var(y) = \sigma^2 V$, and S_0 and S_1 are the smoothing matrices for *Model* 0 and *Model* 1 respectively. Given the nature of *Model* 0, it can be seen that S_0 is a matrix whose elements are $\frac{1}{n}$. The df used here are the df for the error that account for correlation, obtained in section 4.4.2.1 and recalled below:

$$df_i = tr[V^{-1}S_iV + S_i^T - S_i^T V^{-1}S_iV]; \quad i = 0, 1$$

In this way it will be possible to test for a significant effect of x on y accounting for the correlation of the errors.

Using the generalized RSS, it is possible to obtain the standard errors band modified for correlation and given in equation (4.24):

$$s.e. = \sqrt{var\{(S_0 - S_1)y\}} = \sqrt{diag\{(S_0 - S_1)V(S_0 - S_1)^T\}\sigma^2} \qquad (4.24)$$

where $var(y) = \sigma^2 V$. The meaning of this band is that if *Model* 1 lies outside the band, then *Model* 1 differs by more than 2 standard errors from *Model* 0 at that point.

As described in section (4.1), the estimate of the correlation matrix will follow

the approach of Niu (1996), which consists in fitting an independence model and using the residuals to identify a suitable structure for the error component.

4.5 A simulation study

This section presents the results from a simulation study carried out to understand the performances of the approximate F test and the Pseudo Likelihood Ratio test, both accounting for and not accounting for correlation. For the approximate F test, the definitions of the degrees of freedom that have been used are those introduced in section 4.4.2.1, and listed in (4.13).

The study was formulated to match the general patterns of the SO_2 data. Given Models M_A, M_B, M_C:

$$\begin{aligned} M_A: \quad y &= \alpha + m_w(weeks) + \varepsilon \\ M_B: \quad y &= \alpha + m_y(years) + m_w(weeks) + \varepsilon \\ M_C: \quad y &= \alpha + m_{yw}(years, weeks) + \varepsilon \end{aligned}$$

two different situations have been considered for the present simulations:

- **Situation 1**: The tests' performances were analyzed when they are used in testing the presence of trend. In other words this means that M_A and M_B were tested, generating data from M_A (size study), and generating data from M_B (power study).
- **Situation 2**: The tests' performances were analyzed when used in testing for changes in seasonality. In other words this means that M_B and M_C were tested, generating data from M_B (size study), and generating data from M_C (power study).

The following subsections will analyze each of these two situations described above.

4.5.1 Test for trend

4.5.1.1 Size

Data were simulated from the additive model (4.25)

$$y = 2 - \frac{1}{2}\cos\left(2\pi\frac{weeks}{53}\right) + \frac{\varepsilon}{2} \tag{4.25}$$

where the errors ε were sampled from an AR(1) process with variance 1 and correlation parameter $\rho = 0, 0.2, 0.4$. Comparisons of models M_A and M_B were carried out to assess evidence for the presence of trend over the years.

The following steps provide a summary of the estimation and testing procedures used for this simulation study:

- Step 1: Values are simulated from Model (4.25).
- Step 2: Model M_B, assuming independent errors, has been fitted to the simulated data of Step 1 with smoothing parameter h.
- Step 3: The residuals from Step 2 are used to compute the autocorrelation function, assuming an AR(1), from which an estimate of the correlation coefficient at lag 1 is used to estimate $\hat{\rho}$.
- Step 4: Models M_A and M_B have been fitted to the simulated values of Step 1 with smoothing parameter h, and correlation parameter $\hat{\rho}$ and $\tilde{\rho}$: $\hat{\rho}$ is the estimated correlation parameter of Step 3; $\tilde{\rho}$ are plug-in values of the correlation parameter.
- Step 5: Models fitted in Step 4 will be tested using the approximate F test (with degrees of freedom given in expressions 4.13) and the quadratic form

method.

Simulations consisted in generating 200 data sets of 11 years of data (1990-2000), with 53 weeks per year. Therefore the test will present a reasonable size if the proportions of significant p values under the null hypothesis are $5\% \pm 3.1\%$. Simulations have been carried out using different values of $h = (h_1, h_2)$ and ρ. A reasonable choice for the smoothing parameters is $h_1 = 1.3$ and $h_2 = 0.4$. Other values have been used for simulations, multiplying $h = (h_1, h_2)$ by four different multipliers: 0.5, 0.67, 1.5, 2. In this way, the behavior of the tests will be analyzed when the smoothing parameters used, range from half to double the choice of a well-chosen smoothing parameter.

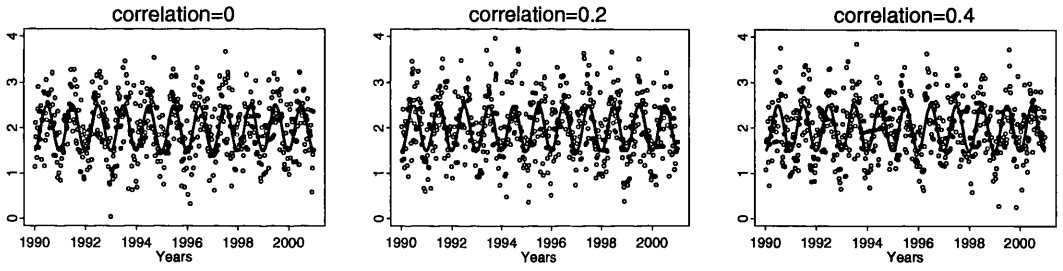

Figure 4.1: Simulations from Model (4.25) for correlation 0, 0.2, 0.4, with the seasonal component plotted by a line.

Figure 4.1 shows examples of simulated values from Model (4.25) for different correlation values with the seasonal component plotted by a line. Table 4.1 shows simulation results generated with correlation parameters $\rho = 0,\ 0.4$, and using the independent version of the Pseudo Likelihood Ratio test ($Q.F.$), and the approximate F test with degrees of freedom given by $df_{par.}$, $df_{var.}$, $df_{var.c.}$, $df_{err.}$ (see equation 4.13). The row ($\tilde{\rho}$) presents size results when the true correlation

value is plugged in, while the results in the row indicated with ($\hat{\rho}$) refer to the simulations when estimated correlations were used. From table 4.1, it is clear that the sizes across all the tests are not well controlled. The only two sensible size results are for the Pseudo Likelihood Ratio test ($Q.F.$) and the approximate F test with $df_{err.}$ when data are simulated with no correlation. However, even these increase substantially when correlation is present. Similar results have been obtained also for situation 2. Therefore from now on the simulation study shown will consist of the results concerning the Pseudo Likelihood Ratio test that accounts for correlation, and the approximate F test with $df_{err.c.}$ (for easier representation the $df_{err.c.}$ will be replaced by simply df). The results for the size study of situation 1, are shown in table 4.2. Each table refers to simulations generated with a different correlation parameter ($\rho = 0,\ 0.2,\ 0.4$). Each row presents size results when different plug in correlation values ($\tilde{\rho}$) or estimates ($\hat{\rho}$) are used.

Table 4.1: Empirical sizes results of the approximate F test with degrees of freedom defined by $df_{par.}$, $df_{var.}$, $df_{var.c.}$, $df_{err.}$ and the independent version of the Pseudo Likelihood Ratio test ($Q.F.$) from 200 data sets simulated with correlation parameters $\rho = 0,\ 0.4$, smoothing parameters $h = (h_1,\ h_2) = (1.3,\ 0.4)$, and using the true ($\tilde{\rho}$) and the estimated ($\hat{\rho}$) correlation.

	ρ=0					ρ=0.4				
	$df_{par.}$	$df_{var.}$	$df_{var.c.}$	$df_{err.}$	$Q.F.$	$df_{par.}$	$df_{var.}$	$df_{var.c.}$	$df_{err.}$	$Q.F.$
$\tilde{\rho}$	0.29	0.39	0.40	0.03	0.04	0.83	0.87	0.78	0.74	0.71
$\hat{\rho}$	0.31	0.37	0.39	0.04	0.06	0.93	0.96	0.91	0.83	0.79

Results of tables 4.2 show that whenever the correlation is known, the sizes of both tests, the approximate F test and Pseudo Likelihood Ratio test, perform

well across the smoothing parameters and the correlations used. There is some indication that the size of the approximate F test is consistently too low.

From Table 4.2, it is possible to note that whenever a correlation value of the test is smaller than the one used for simulating the data, the size of the test increases. This can be explained by thinking of correlation as generating irregular "waves", thus the estimated bivariate surface will consider a part of these "waves" to be due to correlation, since a small correlation parameter has been used. Therefore the "waves" left will be attributed to a bivariate term that can not be modeled just by an additive term. Therefore the smaller the correlation, the more frequent the rejection of the null hypothesis. On the other hand, if the correlation used for the estimates of the surfaces is greater than the one used for simulating the data, the size becomes smaller. This is due to the fact that a larger correlation parameter will model the true correlation generated, and also a proportion of the trend in the data due to the regression surface.

The results in Table 4.2 show that when the correlation has to be estimated, the choice of the smoothing parameter plays an important role. Low smoothing parameters cause underestimation of the correlation, increasing the size of the test, while high smoothing parameters overestimate the correlation, decreasing the size of the test. Therefore a conservative approach of adopting a fairly high smoothing parameter seems to be safer.

4.5.1.2 Power

Data were simulated from the additive model (4.26)

$$y = 2 - \frac{1}{40}(years - 1990) - \frac{1}{2}\cos\left(2\pi\frac{weeks}{53}\right) + \frac{\varepsilon}{2} \qquad (4.26)$$

where the errors ε are sampled from an AR(1) process with variance 1 and correlation parameter $\rho = 0, 0.2, 0.4$. Model (4.26) is characterized by a decreasing trend of $\frac{1}{40}$ over 11 years, i.e. a reduction of 2.5% over 11 years. Simulations have been also undertaken with different amounts of trend, but the results presented in this section will refer to $\frac{1}{40}$ because this gives the most interesting results in terms of the performances of both tests, the approximate F test and Pseudo Likelihood Ratio test. Figure 4.2 shows examples of simulated values from Model (4.26) for different correlation values with the seasonal component + trend plotted by a line. Comparisons of the models M_A and M_B were carried out to assess evidence

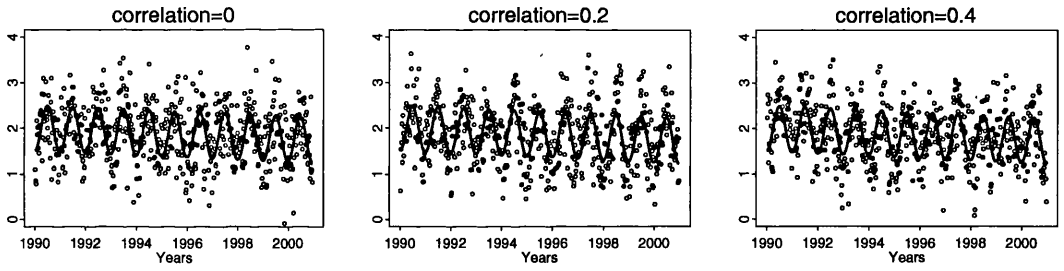

Figure 4.2: Simulations from Model (4.26) for correlation 0, 0.2, 0.4, with the seasonal component + trend plotted by a line.

for the presence of trend over the years.

The steps used in this power study are the same as in Section 4.5.1.1, apart from Step 1 where the values are simulated from model (4.26) rather than model (4.25).

Simulations consisted in generating 200 data sets of 11 years of data (1990-2000), with 53 weeks per year. Simulations have been carried out using different values of $h = (h_1, h_2)$ and ρ. On the basis of the results of section 4.5.1.1, a

reasonable choice for the smoothing parameters seem to be $h_1 = 1.3$ and $h_2 = 0.4$ or larger. The smoothing parameters values used are therefore $h_1 = 1.3$ and $h_2 = 0.4$ multiplied by 1, 1.5, 2. In this way we will be analyzing the power of the tests for larger smoothing parameters that produce effective size results.

The results are shown in Table 4.3. The rows for $\tilde{\rho}$ refer to those simulations where the true correlation was used. The rows $\hat{\rho}$ refer to those simulations where the estimated correlations were used.

Table 4.3 shows that the power performances are not substantially affected by the fact that the correlation is known or is estimated. The Pseudo Likelihood Ratio test seems to have slightly higher power compared to the approximate F test. Using different smoothing parameters does not have a large effect on the power results. However, high correlation reduces the power of both tests as expected. Overall, it can be said that both tests perform well when they are used to test for changes of trend of more than 2.5% in 11 years.

4.5.1.3 Test for trend: Conclusions

The simulation study just presented aimed to look at the performances of the approximate F test and Pseudo Likelihood Ratio test when they were used for testing the presence of trend (or what was called situation 1 in Section 4.5). Results show that both tests are characterized by sizes which are dramatically high when correlation is not accounted for. The Pseudo Likelihood Ratio test that accounts for correlation, and approximate F test with degrees of freedom defined by $df_{err.c.}$ in expression (4.13) are the only two formulations that show good size results, when the correlation is known, under all different settings analyzed. When the correlation has to be estimated, the size of both tests is dependent

on the choice of the smoothing parameters, which affects the estimates of the correlation parameters. However, the choice of a higher smoothing parameter is a safer approach in terms of its effect on the size of the tests. The power of both tests seems excellent for detecting changing of trends of at least 2.5% in 11 years. The next section will show the performances of the Pseudo Likelihood Ratio test accounting for correlation, and of the approximate F test with degrees of freedom defined by $df_{err.c.}$ when they are used for testing changes in seasonality (or what was called situation 2 in Section 4.5).

4.5.2 Test for changes in seasonality

As noted in Section 4.5, in the second situation, the performances of the methods were analyzed when testing for changes in seasonality. In other words, this means that M_B and M_C were tested on data generated from model M_B (size study), and from model M_C (power study).

4.5.2.1 Size

Data were simulated from the additive model (4.27)

$$y = 2 - \frac{year - 1990}{5} - \frac{1}{2}\cos\left(2\pi\frac{week}{53}\right) + \frac{\varepsilon}{2} \tag{4.27}$$

where the errors ε are sampled from an AR(1) process with variance 1 and correlation parameter $\rho = 0, 0.2, 0.4$. Comparisons of the models M_B and M_C were carried out to assess evidence for changes in seasonality across years.

The following steps provide a summary of the estimation and testing procedures used for this simulation study:

- Step 1: Values are simulated from Model (4.27).
- Step 2: Model M_C, assuming independent errors, has been fitted to the simulated data of Step 1 with smoothing parameter h.
- Step 3: The residuals from Step 2 are used to compute the autocorrelation function, assuming an AR(1), from which an estimate of the correlation coefficient at lag 1 is used to estimate $\hat{\rho}$.
- Step 4: Models M_B and M_C have been fitted to the simulated values of Step 1 with smoothing parameter h, and correlation parameter $\hat{\rho}$ and $\tilde{\rho}$: $\hat{\rho}$ is the estimated correlation parameter of Step 3; $\tilde{\rho}$ are plug-in values of the correlation parameter.
- Step 5: Models fitted in Step 4 will be tested using the approximate F test and the Pseudo Likelihood Ratio test.

Simulations consisted in generating 200 data sets of 11 years of data (1990-2000), with 53 weeks per year. The test will present a reasonable size if the proportion of significant p values under the null hypothesis is $5\% \pm 3.1\%$. As in situation 1, simulations have been carried out using $h = (1.3, 0.4)$, scaled by 0.5, 0.67, 1.5, 2. Figure 4.3 show examples of simulated values from Model (4.27) for different correlation values with the seasonal component + trend plotted by a line. Results are shown in Table 4.4. Each table refers to simulations generated with a different correlation parameter ($\rho = 0,\ 0.2,\ 0.4$). Each row of each table presents size results when different plug in correlation values ($\tilde{\rho}$) or estimated correlations ($\hat{\rho}$) are used.

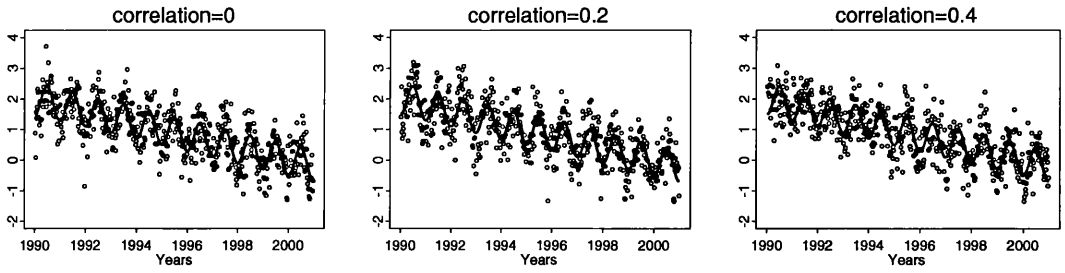

Figure 4.3: Simulations from Model (4.27) for correlation 0, 0.2, 0.4, with the seasonal component + trend plotted by a line.

The results in Table 4.4 show that whenever the correlation is known, the size of the Pseudo Likelihood Ratio test is well controlled, while the approximate F test results seem to have a smaller size.

As in situation 1, from Table 4.4, it is possible to note that whenever a correlation value of the test is smaller than the one used for simulating the data, the size of the test becomes bigger. On the other hand, if the correlation used for the estimates of the surfaces is larger than the one used for simulating the data, the size becomes smaller.

Even in this situation, the results show that when the correlation has to be estimated, the choice of the smoothing parameters plays an important role. Indeed, low smoothing parameters cause underestimation of the correlation, increasing the size of the test. On the other hand, high smoothing parameters overestimate the correlation, decreasing the size of the test. Therefore a conservative approach of choosing a fairly high smoothing parameter seems to be safer in this context.

4.5.2.2 Power

For the power study of situation 2, data need to be generated by a model whose seasonal component changes across years. Two different changes of seasonality across years have been considered: changes in amplitude, and changes in the phase of the seasonal cycles.

Data were therefore simulated from the additive models (4.28) and (4.29).

$$y = 2 - \frac{year - 1990}{5} - \frac{1}{2}\cos\left(2\pi\frac{week}{53} - \frac{year - 1990}{8}\right) + \frac{\varepsilon}{2} \quad (4.28)$$

$$y = 2 - \frac{year - 1990}{5} - \frac{1}{2}\cos\left(2\pi\frac{week}{53}\right)\frac{year - 2001}{10} + \frac{\varepsilon}{2} \quad (4.29)$$

where the errors ε are sampled from an AR(1) process with variance 1 and correlation parameter ρ. Comparisons of the models M_B and M_C were carried out to assess evidence for changes of seasonality over the years. Model (4.28) refers to a situation where the phase of the seasonal component changes across 11 years by about 2 months. Model (4.29) referred to a situation where the amplitude (peak - trough) of the seasonal component changes from 1.1 to 0 across 11 years.

Simulations have been also undertaken with other values of changes in amplitude and in the phase of the seasonal component, but the results presented in this section refer to the ones formulated in Model (4.28) and Model (4.29), because they show the most interesting results in terms of performances of both tests.

The steps used in this power study are the same as section 4.5.2.1, apart from Step 1 where the values are simulated from model (4.28) and from model (4.29), rather than model (4.27).

The simulations consisted in generating 200 data sets of 11 years of data (1990-2000), with 53 weeks per year. Simulations were carried out using different values of $h = (h_1, h_2)$ and ρ. In particular, a sensible choice for the smoothing parameters is $h_1 = 1.3$ and $h_2 = 0.4$. Then on the basis of results of section 4.5.2.1, other smoothing parameter values have been used for simulations by multiplying $h = (h_1, h_2)$ by two different multipliers: 1.5, 2. In this way we will be analyzing the power of the tests for large smoothing parameters that guarantee safer size results.

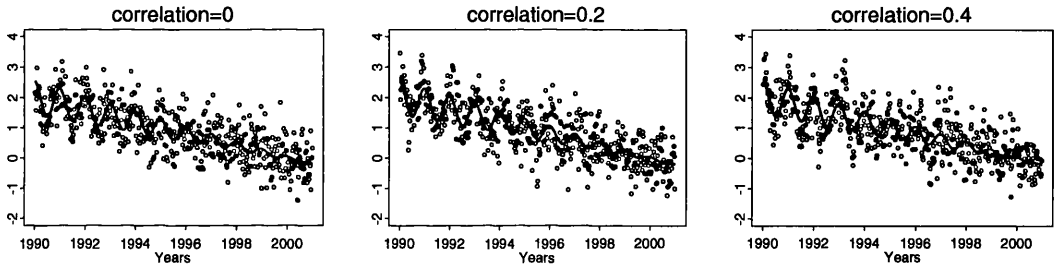

Figure 4.4: Simulations from Model (4.29) for correlation 0, 0.2, 0.4, with the seasonal component + trend plotted.

Figure 4.4 shows examples of simulated values from Model (4.29) for different correlation values with the seasonal component + trend plotted by a line. Figure 4.5 shows examples of simulated values from Model (4.28) for zero correlation, with the seasonal component + trend plotted by a line.

The results of the simulation are shown in Table 4.5. The results in the rows for $\tilde{\rho}$ refer to those simulations where the true correlation was used. Results in the rows for $\hat{\rho}$ refers to those simulations where the estimated correlations were used.

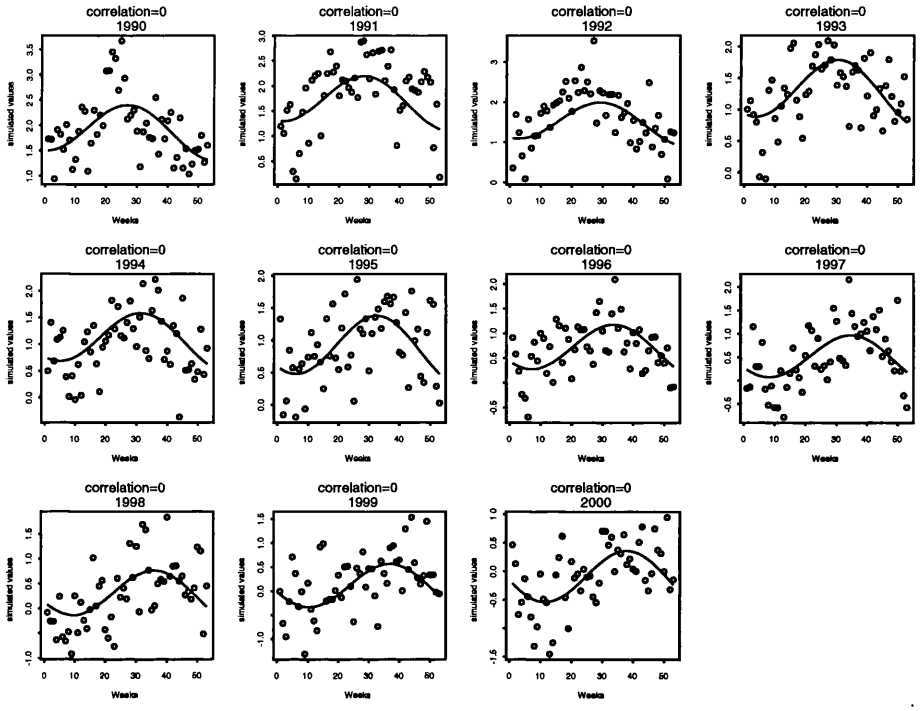

Figure 4.5: Simulations from Model (4.28) for correlation 0, with the seasonal component + trend plotted by a line.

Table 4.5 shows that the power performances reduce as the correlation increases. Both tests seem relatively unaffected by whether the correlation is known or estimated. The Pseudo Likelihood Ratio test seems to have slightly higher power compared to the approximate F test, especially with high correlations. When the correlation is known, the power is not substantially affected by the choice of smoothing parameter. When the correlation is estimated, different smoothing parameters have relatively little effect on the performance of the Pseudo Likelihood Ratio test, while the power of the approximate F test seems to reduce quite substantially.

Overall, it can be said that, both tests seem to have a high power of detecting

changes in phase of the seasonal component longer than 2 months across 11 years, and changes in amplitude (peak - trough) of the seasonal component of more than 1.1 across 11 years. There are some indications that the Pseudo Likelihood Ratio test is more effective.

4.5.2.3 Test for changes in seasonality: Conclusions

The simulation study just presented aimed to look at the performances of the approximate F test and Pseudo Likelihood Ratio test when they were used for testing changes in seasonality over time (or what was called situation 2 in Section 4.5). When the correlation is known, the size of the Pseudo Likelihood Ratio test seems to work very well under all different settings analyzed. However the sizes of the approximate F test seem rather low. As for "situation 1", when correlation has to be estimated, the size of both tests is dependent on the smoothing parameters, whose choice of higher values guarantee a safer approach in terms of better controlled size. The power of both tests seems excellent for detecting changes in amplitude (peak - trough) of more than 1.1 in 11 years, and changes in the phase of the seasonal component of about 2 months across 11 years.

4.6 Modeling with correlated errors: Conclusions

This chapter presents techniques that allow correlated data to be modeled as a function of a number of covariates. Tests for model selection and for components' comparison that account for correlation are also shown. The extension of well

established nonparametric techniques such as local linear regression, backfitting algorithm, approximate F test and Pseudo Likelihood Ratio test to account for correlation are shown to be flexible and powerful tools that can be used with a variety of data. Simulations show the importance of using these methodologies when correlation is present. Indeed not accounting for correlation, when it is present, would damage the size of the tests. The power of both tests seems excellent for detecting trends of at least 2.5% in 11 years, changes in amplitude (peak - trough) of more than 1.1 in 11 years, and changes in the phase of the seasonal component of about 2 months across 11 years. The following chapter will apply these methodologies to air pollution and meteorological data showing the importance of accounting for correlation. Besides, since the simulation study showed that the choice of the smoothing parameters affects the estimates of the correlation parameters, the next chapter will also present a sensitivity analysis of the results of the application with different correlation estimates.

Table 4.2: Empirical sizes of the test to compare models M_A and M_B. For each parameter setting, 200 datasets were simulated from Model (4.25), with smoothing parameters of $h_1 = 1.3$ and $h_2 = 0.4$ multiplied by the values indicated in the table. Within each cell the upper value refers to the approximate F test, while the lower value refers to the Pseudo Likelihood Ratio test. Each table is referred to simulations generated with a different correlation parameter (ρ = 0, 0.2, 0.4). Each row of each table presents size results when different plug in correlation values ($\tilde{\rho}$) or correlation's estimates ($\hat{\rho}$) are used.

	ρ=0				
h multipliers	0.5	0.67	1	1.5	2
$\tilde{\rho} = 0.4$	0.00	0.00	0.00	0.00	0.00
	0.00	0.00	0.00	0.00	0.00
$\tilde{\rho} = 0.2$	0.00	0.00	0.00	0.005	0.01
	0.00	0.00	0.00	0.01	0.01
$\tilde{\rho} = 0.0$	0.04	0.035	0.02	0.025	0.01
	0.055	0.06	0.03	0.03	0.01
$\hat{\rho}$	0.095	0.04	0.035	0.02	0.01
	0.13	0.06	0.055	0.02	0.01

	ρ=0.2				
h multipliers	0.5	0.67	1	1.5	2
$\tilde{\rho} = 0.4$	0.00	0.00	0.005	0.00	0.005
	0.00	0.00	0.01	0.00	0.005
$\tilde{\rho} = 0.2$	0.05	0.025	0.015	0.015	0.035
	0.065	0.04	0.05	0.04	0.04
$\tilde{\rho} = 0.0$	0.225	0.195	0.165	0.1	0.055
	0.32	0.25	0.225	0.11	0.09
$\hat{\rho}$	0.165	0.065	0.045	0.03	0.005
	0.19	0.09	0.06	0.045	0.01

	ρ=0.4				
h multipliers	0.5	0.67	1	1.5	2
$\tilde{\rho} = 0.4$	0.025	0.015	0.02	0.025	0.02
	0.04	0.03	0.035	0.025	0.045
$\tilde{\rho} = 0.2$	0.275	0.275	0.185	0.135	0.13
	0.365	0.36	0.23	0.17	0.155
$\tilde{\rho} = 0.0$	0.635	0.48	0.375	0.325	0.22
	0.665	0.565	0.445	0.375	0.3
$\hat{\rho}$	0.225	0.1	0.06	0.03	0.02
	0.29	0.145	0.075	0.03	0.035

Table 4.3: Empirical power of the tests to compare models M_A and M_B. For each parameter setting, 200 datasets were simulated from Model (4.26), with smoothing parameters of $h_1 = 1.3$ and $h_2 = 0.4$ multiplied by the values indicated in the table. Within each cell the upper value refers to the approximate F test, while the lower value refers to the Pseudo Likelihood Ratio test. Simulations generated with a different correlation parameter ($\rho = 0,\ 0.2,\ 0.4$). Results in row $\tilde{\rho}$ refer to those simulation where the true correlation was used. Results in row $\hat{\rho}$ refer to those simulations where the estimated correlations were used.

	$\rho = 0$			$\rho = 0.2$			$\rho = 0.4$		
h multiplier	1	1.5	2	1	1.5	2	1	1.5	2
$\tilde{\rho}$	0.85	0.905	0.925	0.695	0.695	0.725	0.465	0.47	0.515
	0.89	0.925	0.945	0.76	0.725	0.785	0.525	0.535	0.56
$\hat{\rho}$	0.885	0.895	0.9	0.69	0.705	0.67	0.46	0.47	0.41
	0.915	0.935	0.92	0.75	0.76	0.72	0.555	0.53	0.475

Table 4.4: Empirical sizes of the test to compare models M_B and M_C. For each parameter setting, 200 datasets were simulated from Model (4.27), with smoothing parameters of $h_1 = 1.3$ and $h_2 = 0.4$ multiplied by the values indicated in the table. Within each cell the upper value refers to the approximate F test, while the lower value refers to the Pseudo Likelihood Ratio test. Each table refers to simulations generated with a range of correlation parameter ($\rho = 0,\ 0.2,\ 0.4$). Each row of each table presents size results when different plug in correlation values ($\tilde{\rho}$) or correlation estimates ($\hat{\rho}$) are used.

	ρ=0				
h multipliers	0.5	0.67	1	1.5	2
$\tilde{\rho} = 0.4$	0.00	0.00	0.00	0.00	0.00
	0.00	0.00	0.00	0.00	0.00
$\tilde{\rho} = 0.2$	0.00	0.00	0.00	0.00	0.00
	0.00	0.00	0.00	0.005	0.00
$\tilde{\rho} = 0.0$	0.015	0.02	0.01	0.015	0.01
	0.035	0.075	0.05	0.035	0.045
$\hat{\rho}$	0.14	0.02	0.005	0.00	0.00
	0.295	0.09	0.025	0.01	0.005

	ρ=0.2				
h multipliers	0.5	0.67	1	1.5	2
$\tilde{\rho} = 0.4$	0.00	0.00	0.00	0.00	0.00
	0.00	0.00	0.00	0.00	0.005
$\tilde{\rho} = 0.2$	0.02	0.02	0.00	0.01	0.01
	0.06	0.04	0.04	0.045	0.045
$\tilde{\rho} = 0.0$	0.645	0.415	0.23	0.095	0.065
	0.845	0.625	0.395	0.255	0.19
$\hat{\rho}$	0.31	0.085	0.005	0.005	0.00
	0.49	0.18	0.055	0.015	0.00

	ρ=0.4				
h multipliers	0.5	0.67	1	1.5	2
$\tilde{\rho} = 0.4$	0.015	0.005	0.015	0.005	0.01
	0.04	0.03	0.07	0.035	0.085
$\tilde{\rho} = 0.2$	0.84	0.6	0.34	0.16	0.08
	0.94	0.795	0.495	0.33	0.275
$\tilde{\rho} = 0.0$	0.995	0.98	0.795	0.405	0.215
	1.00	1.00	0.905	0.595	0.41
$\hat{\rho}$	0.655	0.15	0.005	0.00	0.00
	0.84	0.285	0.095	0.03	0.05

Table 4.5: Empirical power of the tests to compare models M_B and M_C. For each parameter setting, 200 datasets were simulated from Model (4.28) and from Model (4.29), with smoothing parameters of $h_1 = 1.3$ and $h_2 = 0.4$ multiplied by the values indicated in the table. Within each cell the upper value refers to the approximate F test, while the lower value refers to the Pseudo Likelihood Ratio test. Simulations generated with a range of correlation parameters ($\rho =$ 0, 0.2, 0.4). Results in row $\tilde{\rho}$ refer to those simulations where the true correlation was used. Results in row $\hat{\rho}$ refer to those simulations where the estimated correlations were used.

Model (4.29)	$\rho = 0$			$\rho = 0.2$			$\rho = 0.4$		
h multiplier	1	1.5	2	1	1.5	2	1	1.5	2
$\tilde{\rho}$	0.94	0.99	0.965	0.81	0.905	0.84	0.50	0.575	0.56
	0.98	1.00	0.995	0.88	0.975	0.95	0.665	0.725	0.76
$\hat{\rho}$	0.96	0.965	0.915	0.925	0.79	0.68	0.65	0.49	0.415
	0.995	0.99	0.975	0.955	0.91	0.86	0.795	0.72	0.615

Model (4.28)	$\rho = 0$			$\rho = 0.2$			$\rho = 0.4$		
h multiplier	1	1.5	2	1	1.5	2	1	1.5	2
$\tilde{\rho}$	1.00	0.995	1.00	0.95	0.97	0.965	0.7	0.725	0.755
	1.00	1.00	1.00	0.985	0.985	0.995	0.83	0.865	0.87
$\hat{\rho}$	1.00	0.99	0.955	0.94	0.92	0.795	0.795	0.65	0.49
	1.00	1.00	0.995	0.965	0.975	0.96	0.86	0.815	0.75

Chapter 5

Applications Of Additive Models With Correlated Errors

This chapter presents some applications to air pollution data of the methodologies introduced in Chapter 4. Trend and seasonal cycles for SO_2 will be studied accounting for correlation and for meteorological effects at 11 stations across Europe. The sensitivity of the results to the correlation estimates will also be presented.

5.1 Modeling SO_2 accounting for meteorology

Additive models that account for meteorology have been applied, and the results are presented. The data analyzed in this section are the concentrations of SO_2 monitored at: Eskdalemuir (GB02, Scotland), Westerland (DE01, Germany), Waldhof (DE02, Germany), Schauinsland (DE03, Germany), Deuselbach (DE04, Germany), Brotjacklriegel (DE05, Germany), Kosetice (CZ03, Czech Republic),

Rörvik (SE02, Sweden), Bredkälen (SE05, Sweden), Hoburg (SE08, Sweden), Payerne (CH02, Switzerland), from 1973 up to 2001. The predictors used are: years (as fraction of weeks), weeks of the year, weekly amount of precipitation, weekly temperature mean, weekly humidity mean, weekly mean of wind direction weighted by wind speed (defined in equation (1.4) of section 1.5). Because of the skewness of SO_2 and of the amount of rainfall, it has been decided to work on the logarithm of SO_2 and on the logarithm of amount of rainfall. The models that have been fitted are:

$$\begin{aligned}
\ln(SO_2) &= \mu + m_{yw}(years, weeks) + m_r(rain) + m_t(temperature) \\
&+ m_h(humidity) + m_{w.d.s.}(wind.direction.speed) + \varepsilon \quad (Model\ a) \\
\ln(SO_2) &= \mu + m_y(years) + m_w(weeks) + m_r(rain) + m_t(temperature) \\
&+ m_h(humidity) + m_{w.d.s.}(wind.direction.speed) + \varepsilon \quad (Model\ b) \\
\ln(SO_2) &= \mu + m_{yw}(years, weeks) + \varepsilon \quad (Model\ c) \\
\ln(SO_2) &= \mu + m_y(years) + m_w(weeks) + \varepsilon \quad (Model\ d)
\end{aligned}$$

All these models are fitted accounting for temporal correlation and for circular smoothers where needed (e.g. $m_w(weeks)$, $m_{w.d.s.}(wind.direction.speed)$), using the methodology explained in the previous sections.

Model a - *Model d* have been fitted because the purpose of this application is answer the following questions of interest:

- Is the meteorology significant in explaining the variability of SO_2?
- Do the trend and the seasonal cycle estimates change if meteorology is

accounted for?

- Is the seasonality changing significantly across years?

The following subsections will tackle each of the above questions.

5.1.1 Testing the significance of meteorological variables.

Model b includes all the meteorological variables and each of them was fitted in an additive model, using univariate smoothers that account for correlation. Examples of the fits of each component of *Model b* at Deuselbach (DE04), and at Rörvik (SE02) are shown in Figure 5.1 and 5.2. The estimated components are red dashed lines, and the standard error bands are shown with continuous red lines. The dashed black lines and green bands are the estimates and the standard error bands that are obtained if correlation is not accounted for. Apart from the trend estimates for the meteorological variables, inclusion of correlation produces much wider bands.

At both sites, there are decreasing trends and a seasonality characterized by lower values in summer time. At Deuselbach as the amount of rain, the temperature and humidity increase, the concentration of SO_2 decreases. Moreover, the highest concentration of SO_2 corresponds to the wind values coming from the east and north-east. At Rörvik the relationships between the response variable and all the meteorological variables are almost flat, giving an indication that the meteorological variables may be not significant. In order to analyze if the meteorological variables give significant extra information in explaining the SO_2 variability, *Model d*, which includes only trend and seasonality, has been fitted, and compared with *Model b*. The overall fits of *Model b* and *Model d* across

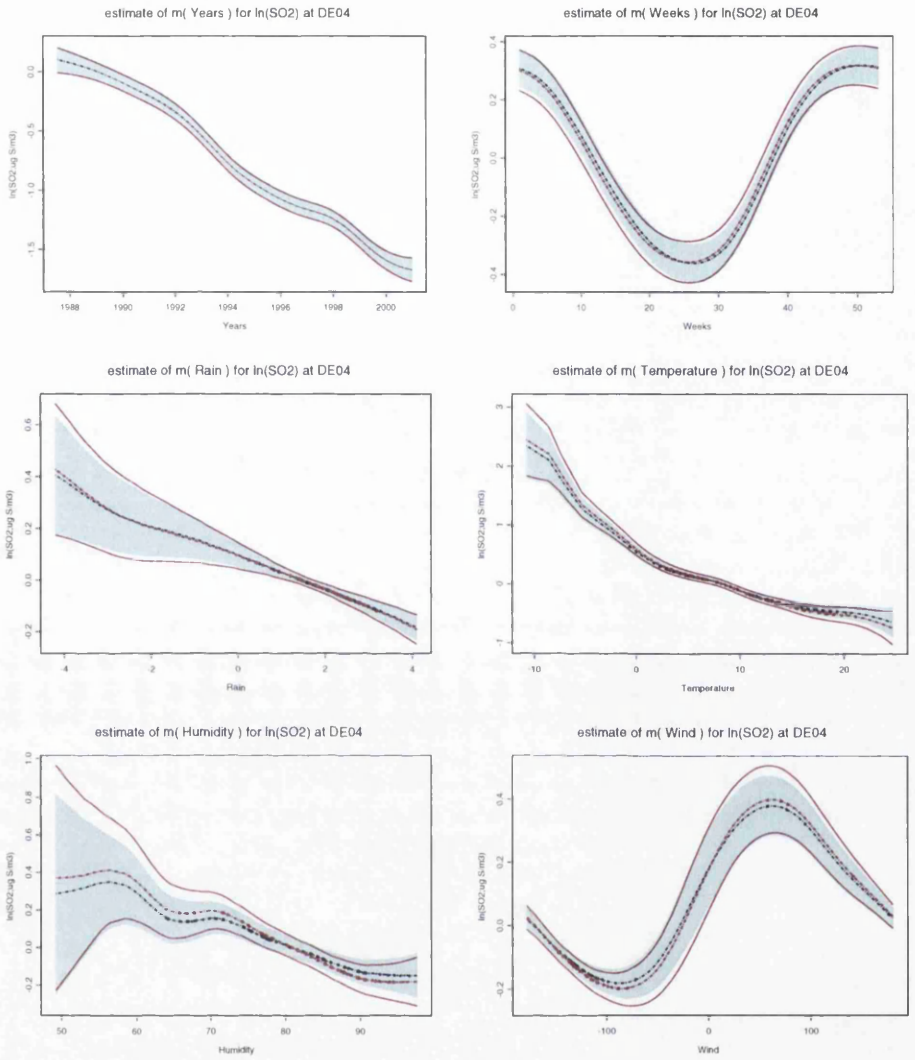

Figure 5.1: a) fit of $m_y(year)$ component versus years; b) fit of $m_w(week)$ component versus weeks; c) fit of $m_r(rain)$ component versus rain values; d) fit of $m_t(temperature)$ component versus temperature values; e) fit of $m_h(humidity)$ component versus humidity values; f) fit of $m_{w.d.s.}(wind.direction.speed)$ component versus wind values; of *Model b* for $\ln(SO_2)$ monitored at Deuselbach (DE04).

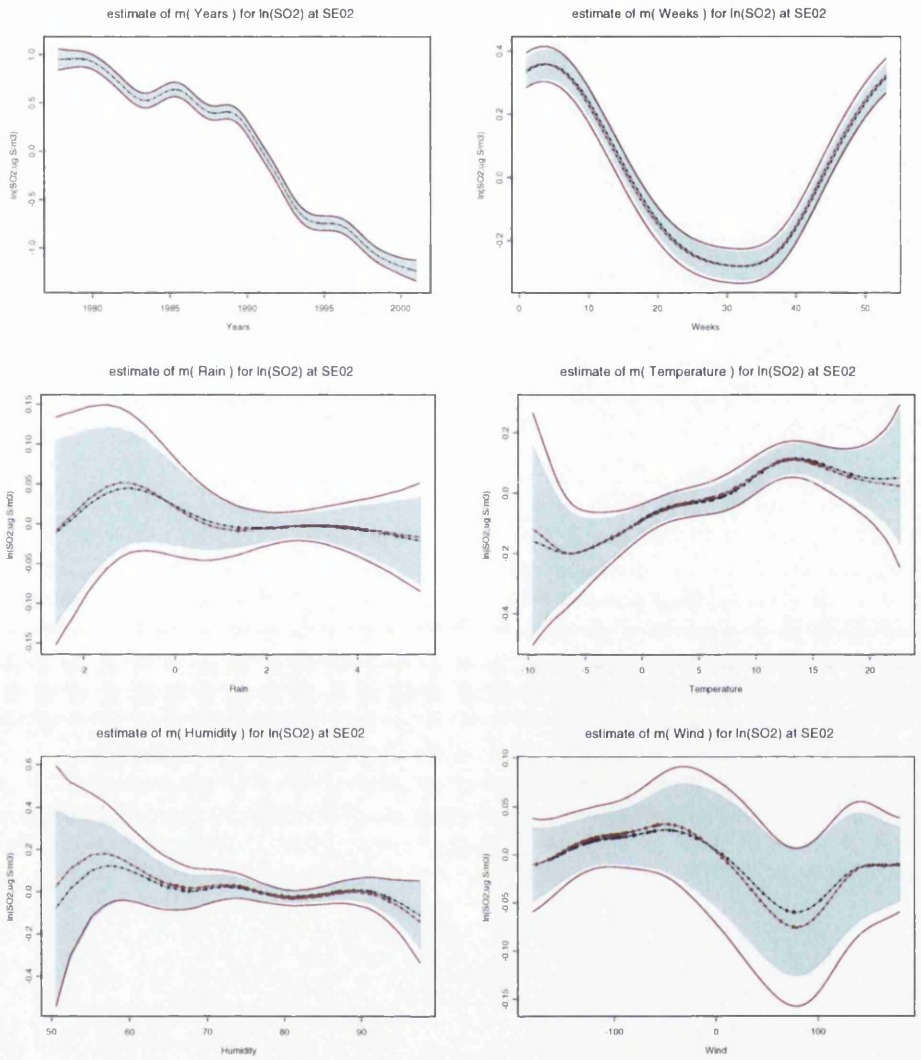

Figure 5.2: a) fit of $m_y(year)$ component versus years; b) fit of $m_w(week)$ component versus weeks; c) fit of $m_r(rain)$ component versus rain values; d) fit of $m_t(temperature)$ component versus temperature values; e) fit of $m_h(humidity)$ component versus humidity values; f) fit of $m_{w.d.s.}(wind.direction.speed)$ component versus wind's values; of *Model b* for $\ln(SO_2)$ monitored at Rörvik (SE02).

all the sites, are displayed in Figure 5.3 and Figure 5.4, where the thin dashed red line is the fit of *Model b*, while the thick continuous blue line is the fit of *Model d*. At Rörvik (SE02) and Bredkälen (SE05), the two lines are almost indistinguishable, giving no indication of a significant effect of the meteorological variables. By contrast at the other nine sites, *Model b* clearly tracks the data more closely than *Model d*, especially where there are rapid fluctuations. This indicates that useful explanatory information is contained in the meteorological data.

These impressions are confirmed by comparing *Model b* and *Model d* more formally through the approximate F test and the Pseudo Likelihood Ratio test. The results of both tests give p values close to 0 across all the sites, except for Rörvik (SE02) and Bredkälen (SE05) whose p values are bigger than 0.05. In other words, this means that the meteorology is statistically significant in explaining the variability of SO_2, except for two Swedish sites.

These results are confirmed by the R^2 values in Table 5.1. The R^2 value of a

Table 5.1: R^2 values from the Additive models

R^2	*Model a*	*Model b*	*Model c*	*Model d*
GB02	0.786	0.774	0.648	0.623
DE01	0.526	0.470	0.241	0.184
DE02	0.801	0.794	0.644	0.626
DE03	0.686	0.666	0.425	0.382
DE04	0.721	0.699	0.530	0.487
DE05	0.665	0.646	0.421	0.381
CZ03	0.688	0.670	0.486	0.449
SE02	0.604	0.582	0.599	0.570
SE05	0.617	0.596	0.609	0.587
SE08	0.557	0.540	0.391	0.359
CH02	0.471	0.417	0.352	0.283

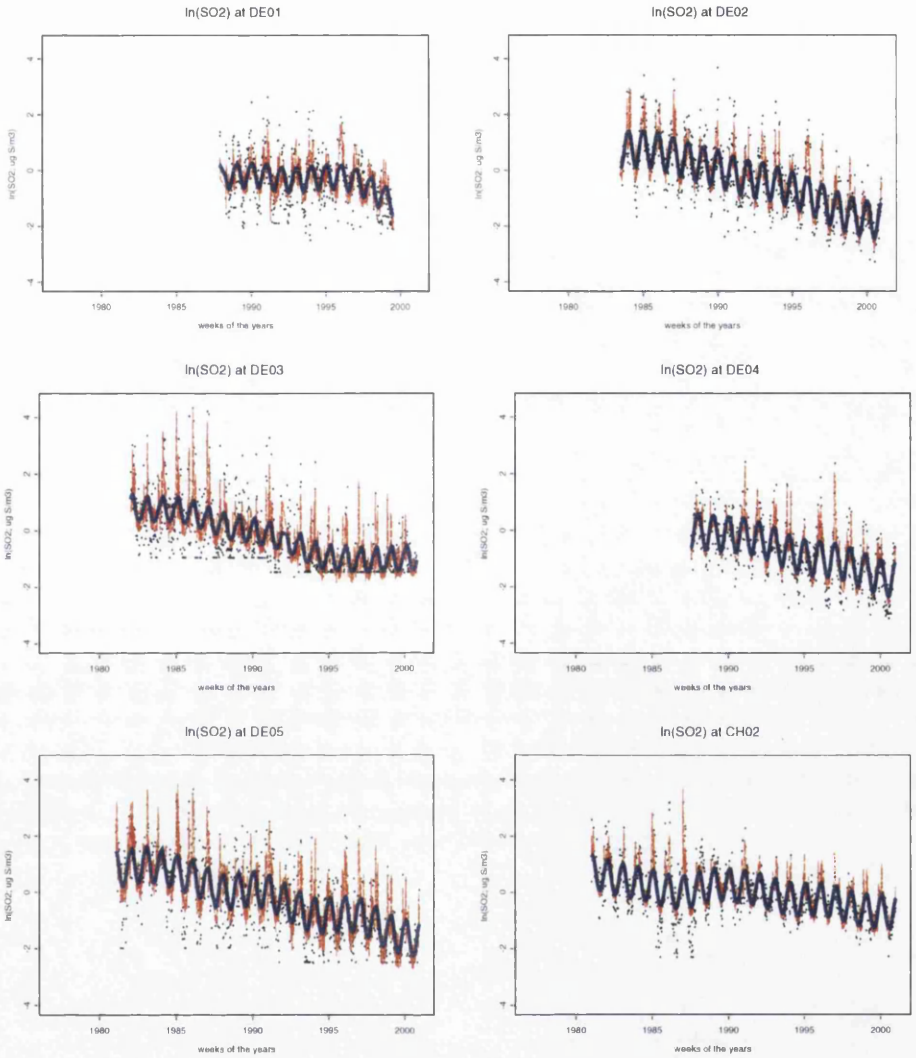

Figure 5.3: fits of *Model b* and *Model d* across all sites.

model such as $M_1: \ y = \alpha + m(x) + \varepsilon$ is obtained by

$$R^2 = \frac{RSS_0 - RSS_1}{RSS_0} \tag{5.1}$$

where RSS_1 is the residual sum of squares of model M_1, and RSS_0 is the residual

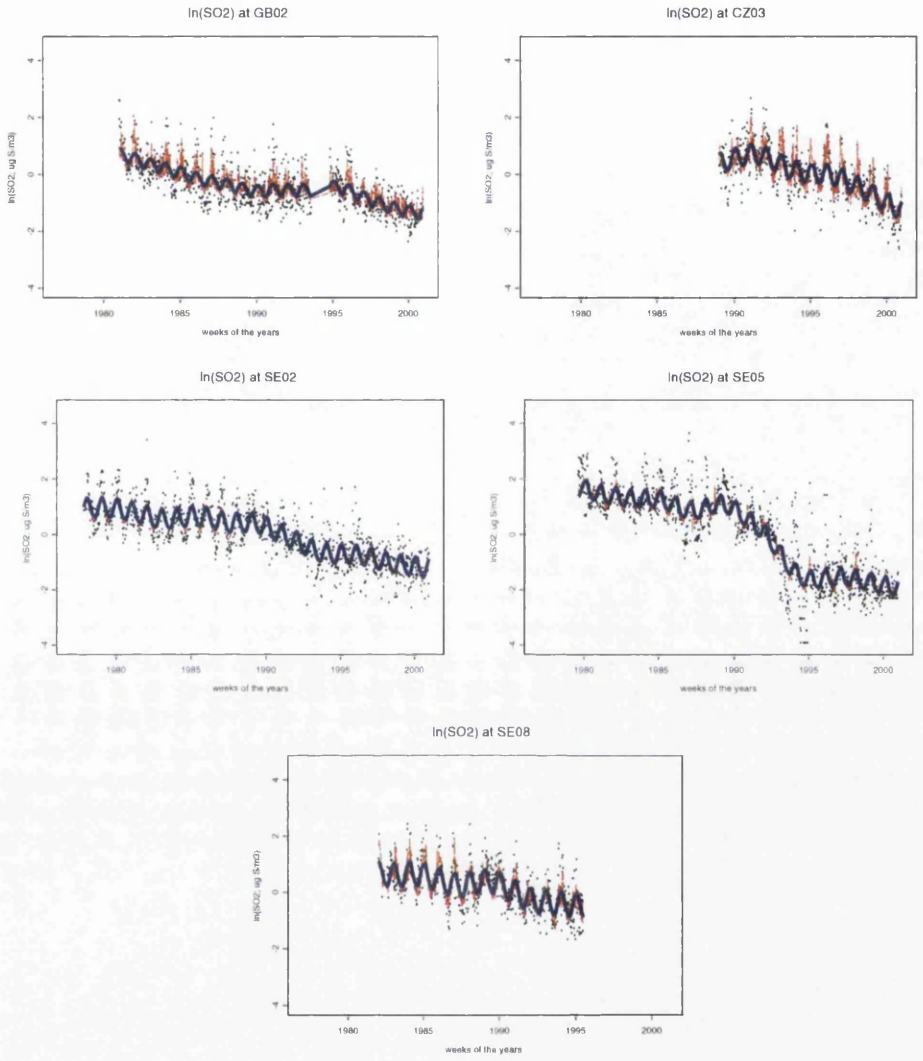

Figure 5.4: fits of *Model b* and *Model d* across all sites.

sum of squares of a model M_0 : $y = \alpha + \varepsilon$ with no covariates (i.e. $RSS_0 = (y - \bar{y})^T V^{-1}(y - \bar{y})$). For Rörvik (SE02) and Bredkälen (SE05), the R^2 values of *Model b*, and the ones of *Model d* are much closer than those for the other sites.

Figure 5.5 shows a map of part of Europe with the estimates of the wind

component from *Model a*. In red are plotted those directions whose estimates were negative, and in black are plotted the positive ones. This means that at each site, the wind that is blowing from the "red" direction reduces the concentration of $\ln(SO_2)$, while those winds that blow from the "black" direction increase the concentration of $\ln(SO_2)$. In the southern sites, concentration of $\ln(SO_2)$ decreases when the wind is blowing from the south and south-east, and increase when the wind blows from the north and north-west. For the northern German sites, the British and the southern Swedish (SE08) sites the concentrations of $\ln(SO_2)$ seem to increases when the wind blows from the west.

5.1.2 Comparing trend and seasonality estimates for the effect of meteorology.

The second question of interest was to analyze if trend and seasonal cycle estimates change when meteorology is accounted for. In other words this means that the $m_y(years)$ and the $m_w(weeks)$ components of *Model b* have to be tested versus the $m_y(years)$ and the $m_w(weeks)$ components of *Model d*. In order to give an answer, test (4.17), presented in Section 4.4.3, has been applied across all the eleven sites, and results are presented in Table 5.2. For the trend component $m_y(years)$, across most of the sites the p values are significant, which means that the trend estimates differ significantly if meteorology is included in the model. The only exceptions are two Swedish sites $SE02$ and $SE05$, whose trends estimates do not change significantly if meteorology is accounted for. However these two sites ($SE02$ and $SE05$) already showed nonsignificant meteorological

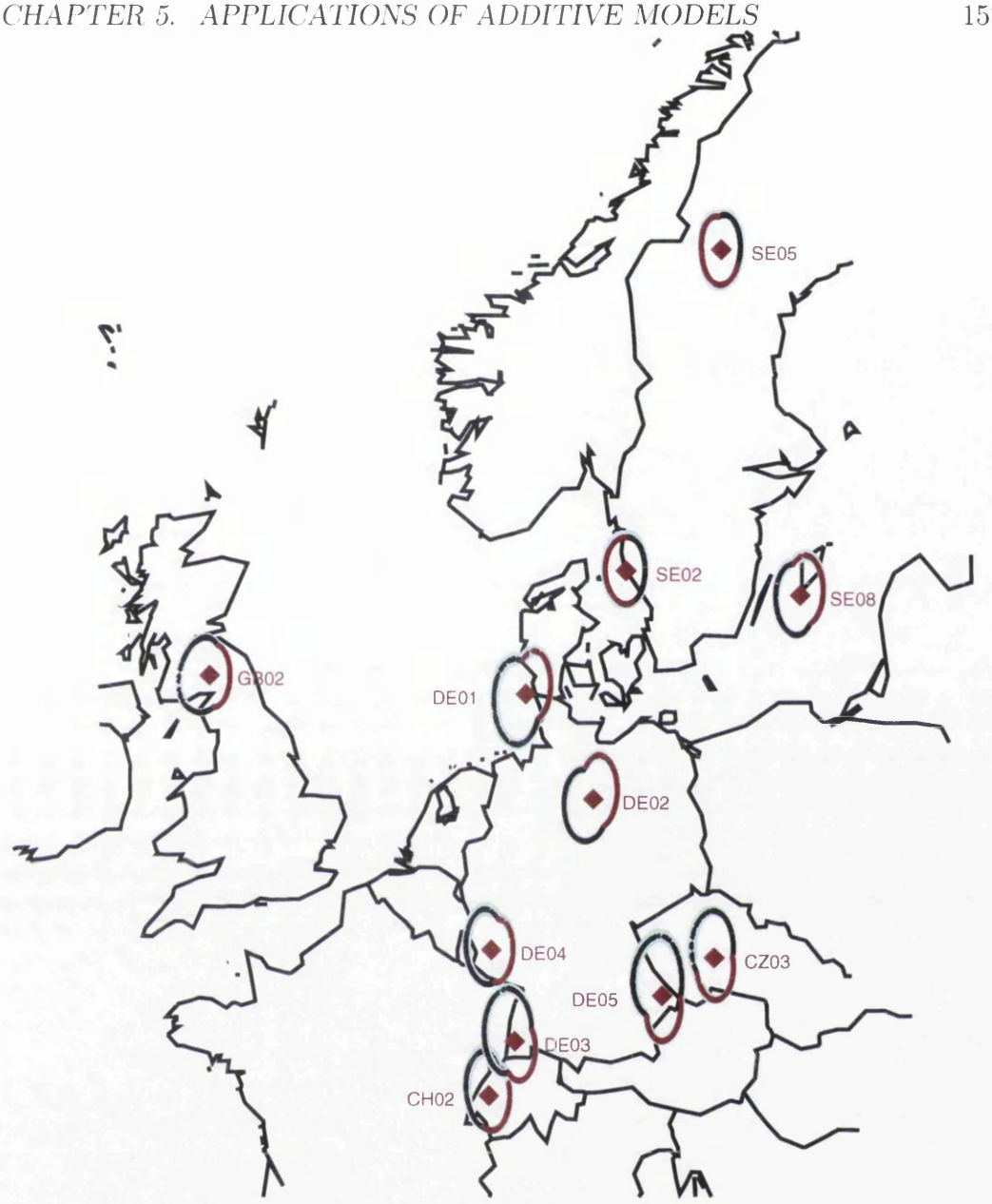

Figure 5.5: Map of the wind estimates from *Model a*.

effects. For the seasonal component, $m_w(weeks)$, the results show that the estimates change significantly if the model accounts for meteorology, except at SE05 where the seasonal signal remains the same, with or without the meteorological

information.

Figures 5.6, 5.7, 5.8 and 5.9 compare the trend, $m_y(years)$, and the seasonal components, $m_w(weeks)$, of *Model b* (continuous line) and *Model d* (dashed line) across all the sites. A standard error reference band, defined in equation (4.21), is also shown and this indicates where the estimates are more than two standard errors apart. For the trend components, in most of the sites the two estimates match very well. The trends diverge temporarily at a few points but the general shape is preserved. These deviations are possible explanations for the significant p values. For the seasonal component, $m_w(weeks)$, the change is much smaller at SE02 and at SE05, where the meteorology was not statistically significant, than the other 9 sites. Apart from SE02 and SE05, the seasonality changes dramatically from *Model b* to *Model d* across all the other sites. In particular the seasonal pattern is stronger when the meteorology is not accounted for. This arises from the fact that some meteorological variables present strong seasonal signals and when these variables are not included in the model, the $m_w(weeks)$ component will be estimating part of the seasonality as well.

5.1.3 Testing for significant changes in seasonality across years.

The question that we tackle now concerns the analysis of possible changes of the seasonal cycles across time. In order to answer this question, *Model a* has been fitted at DE01, DE02, DE03, DE04, DE05, SE08, GB02, CH02, CZ03 and *Model c* has been fitted at SE02 and SE05. Both *Model a* and *Model c* allow the seasonal component to change over time, however *Model a* still accounts for

Table 5.2: Testing for equal trends and seasonalities with and without accounting for meteorology.

	Trends		Seasonalities	
	approximate F test	Pseudo Likelihood Ratio test	approximate F test	Pseudo Likelihood Ratio test
GB02	0.005	0	$4.2e^{-5}$	$3.1e^{-8}$
DE01	0.003	0	$3.6e^{-8}$	$2.2e^{-16}$
DE02	0.004	$8.2e^{-15}$	0	$8.2e^{-15}$
DE03	$1.5e^{-4}$	0	0	0
DE04	0.001	0	0	0
DE05	$8.2e^{-4}$	$2.6e^{-9}$	0	$2.6e^{-9}$
CZ03	$5.0e^{-4}$	0	$5.2e^{-15}$	0
SE02	0.117	0.090	0.024	$1.4e^{-5}$
SE05	0.173	0.287	0.252	0.327
SE08	$2.5e^{-4}$	0	$8.5e^{-12}$	0
CH02	0.017	$9.9e^{-16}$	0	0

meteorological effects, while *Model c* does not. Examples of the fits of *Model a* at DE04 and *Model c* at SE02 are shown in Figures 5.10 and 5.11 respectively.

In order to test for significant changes of the seasonal component across time, at DE01, DE02, DE03, DE04, DE05, SE08, GB02, CH02, CZ03, *Model a* has been tested versus *Model b*, and at SE02 and SE05, *Model c* has been tested versus *Model d*, using approximate F tests and Pseudo Likelihood Ratio tests. The p values are listed in Tables 5.3 and 5.4. From these tables, it can be seen that at Waldhof (DE02), Brotjacklriegel (DE05), Kosetice (CZ03), and Hoburg (SE08) the best model to be fitted is *Model b*, where the trend and seasonal component are fitted as univariate. This means that seasonality does not change significantly across years. Instead, at Eskdalemuir (GB02), Westerland (DE01), Schauinsland (DE03), Deuselbach (DE04) and Payerne (CH02) the trend and seasonal component give significant extra information in explaining the variability of SO_2 if

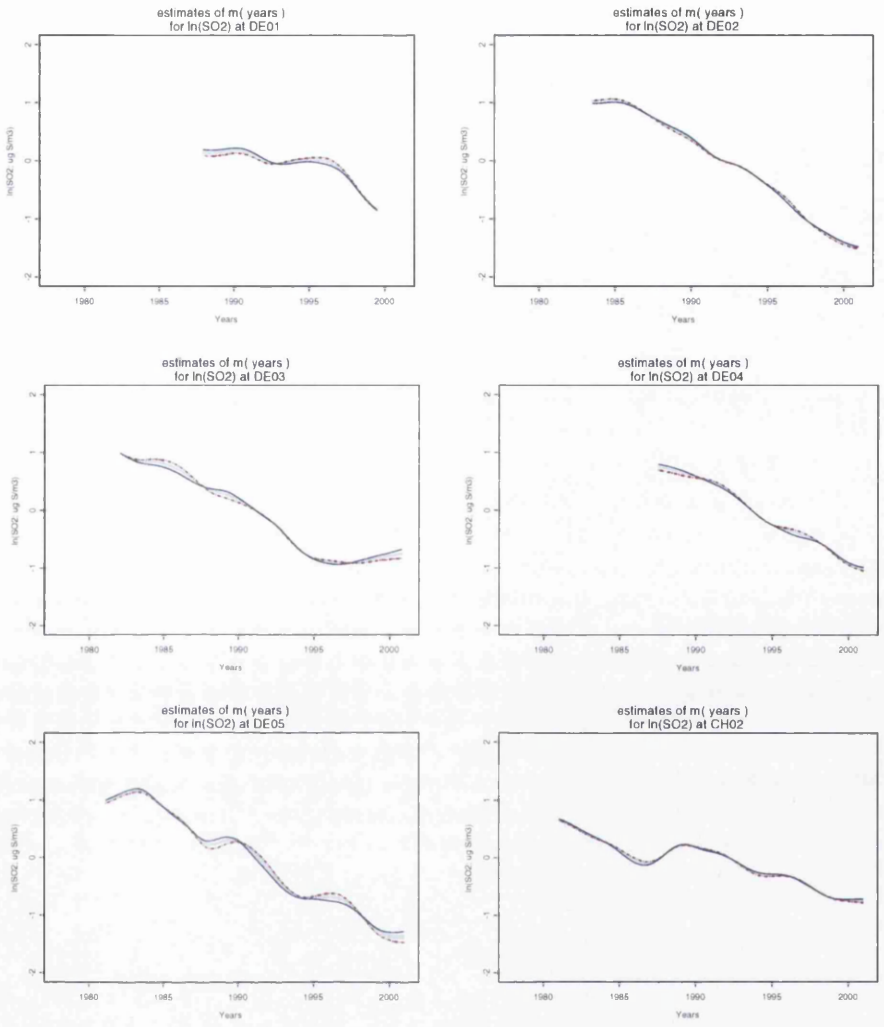

Figure 5.6: Fits of the trends ($m_y(years)$) for *Model b* (continuous line) and *Model d* (dashed line).

they are modeled with a bivariate smoother by *Model a*. At Rörvik (SE02) and at Bredkälen (SE05), it is possible to say that there was a statistically significant change in seasonality across years, and therefore at both sites the best model is *Model c*. These results are also confirmed by Figures 5.12 and 5.13, which show

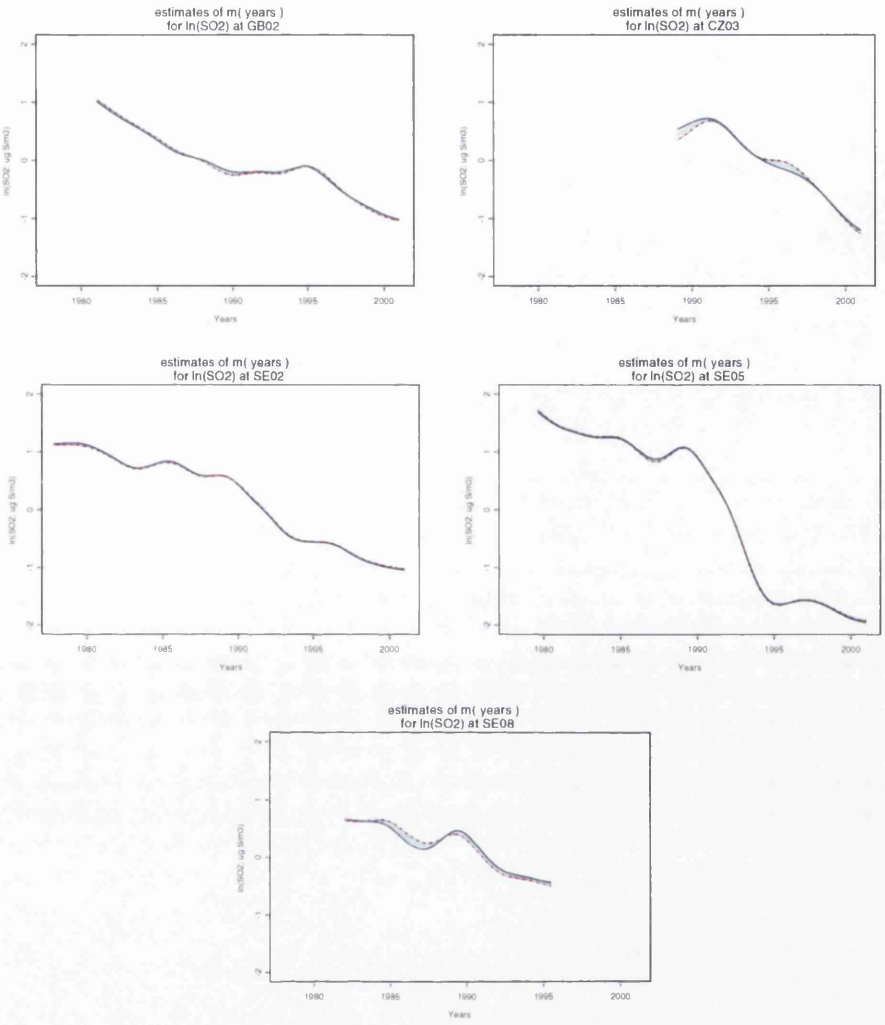

Figure 5.7: Fits of the trends ($m_y(years)$) for *Model b* (continuous line) and *Model d* (dashed line).

the different seasonal cycles, allowing the seasonality to change or to be fixed across years. The dashed black line shows the $m_y(years)+m_w(weeks)$ component of *Model b* at DE01, DE02, DE03, DE04, DE05, SE08, GB02, CH02, CZ03 and *Model d* at SE02 and SE05. The continuous red line shows the $m_{yw}(years, weeks)$

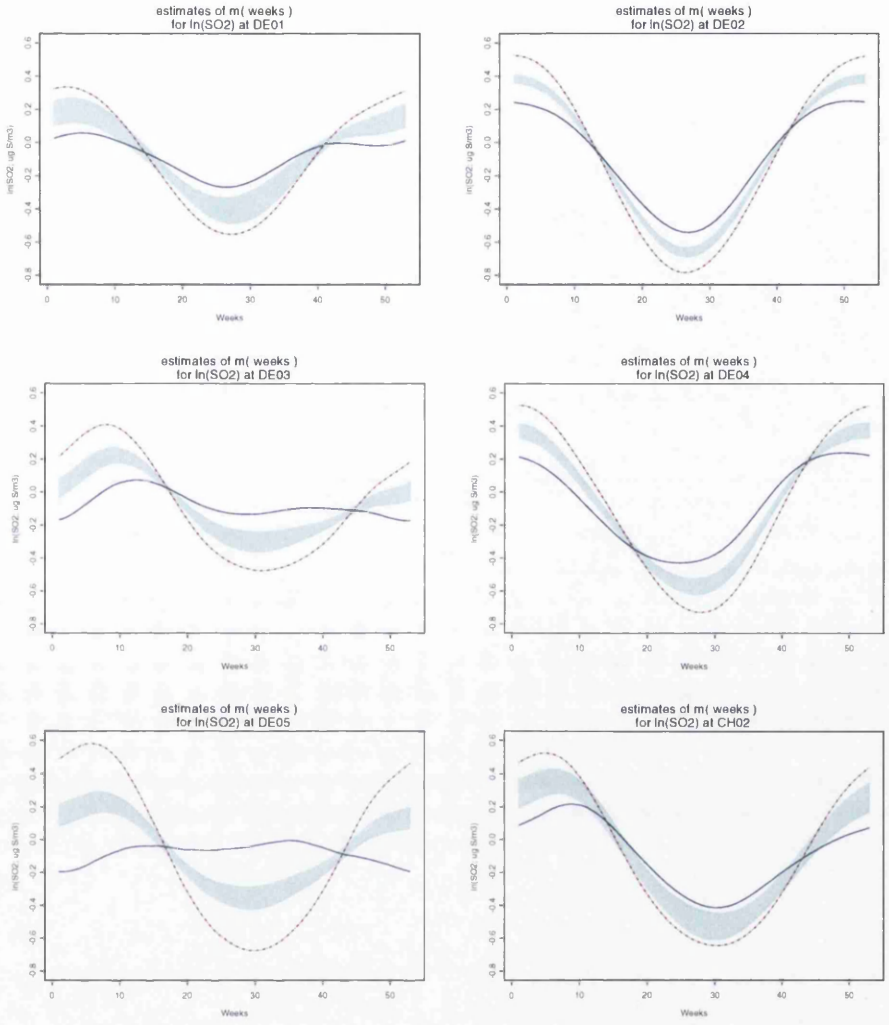

Figure 5.8: Fits of the seasonalities ($m_w(weeks)$) for *Model b* (continuous line) and *Model d* (dashed line).

component of *Model a* at DE01, DE02, DE03, DE04, DE05, SE08, GB02, CH02, CZ03 and of *Model c* at SE02, SE05.

At those sites where the seasonality changes significantly over the years, graphs such as Figures 5.14 and 5.15 have been also produced to analyze whether,

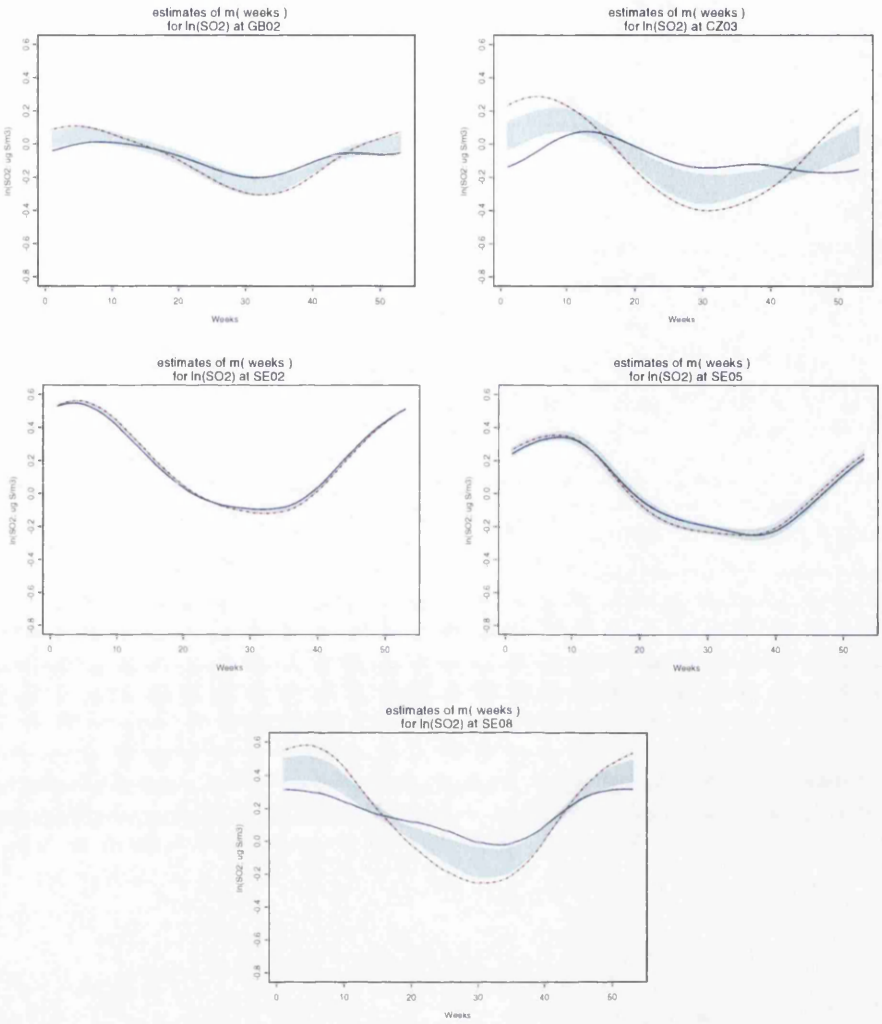

Figure 5.9: Fits of the seasonalities ($m_w(weeks)$) for *Model b* (continuous line) and *Model d* (dashed line).

and if so how, the peaks and troughs change over the years. Figure 5.14 shows the weeks where the yearly troughs (red) and peaks (black) of $m_{yw}(years, weeks)$ component, have been registered for *Model a* at DE04. Figure 5.15 shows the weeks where the yearly troughs (red) and peaks (black) of $m_{yw}(years, weeks)$

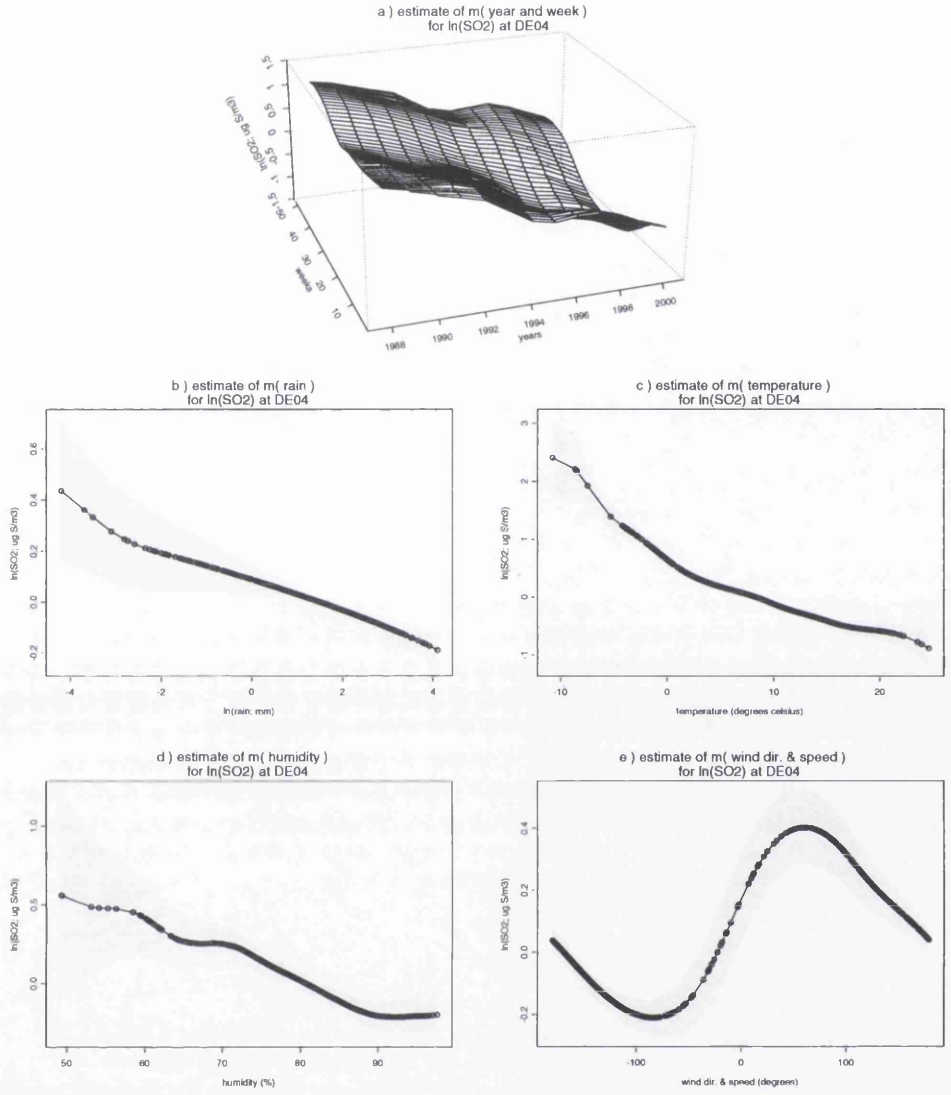

Figure 5.10: a) fit of $m_{yw}(years, weeks)$ component versus years and weeks component; b) fit of $m_r(rain)$ component versus rain values; c) fit of $m_t(temperature)$ component versus temperature values; d) fit of $m_h(humidity)$ component versus humidity values; e) fit of $m_{w.d.s.}(wind.direction.speed)$ component versus wind values, of *Model a* for $\ln(SO_2)$ monitored at Deuselbach (DE04).

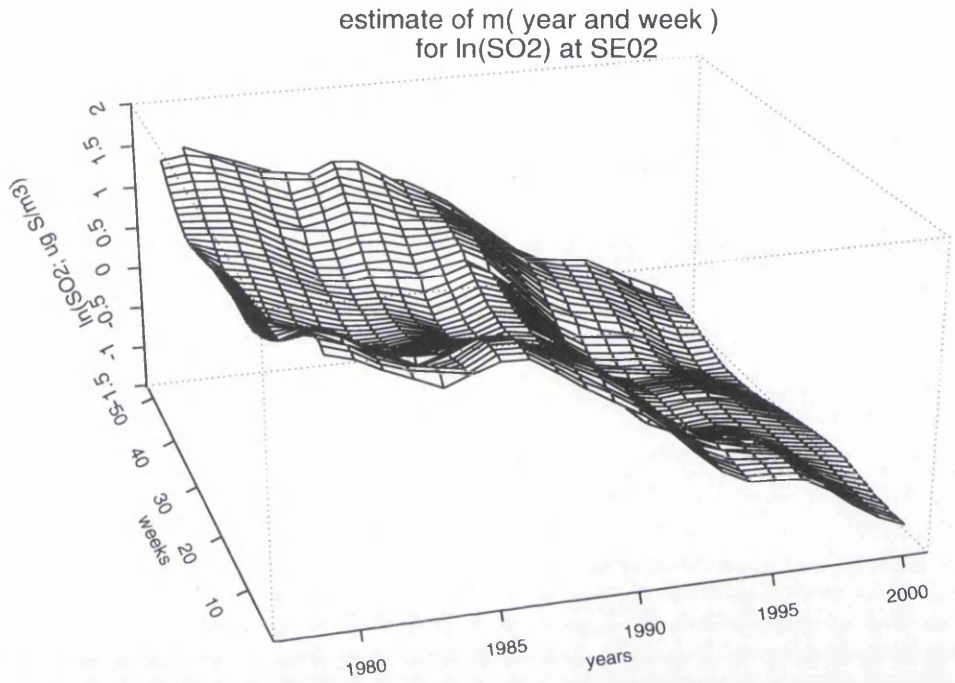

Figure 5.11: a) fit of $m_{yw}(years, weeks)$ component versus years and weeks of *Model c* for $\ln(SO_2)$ monitored at Rörvik (SE02).

component, have been registered for *Model c* at SE02. At DE04, the troughs are in the period May-July, while the peaks are around December-February. At SE02, the troughs are in the period July-September, while the peaks are around December-February.

Figures 5.16 and 5.17 show the pattern followed by the peaks (continuous lines) and troughs (dashed lines) of $m_{yw}(years, weeks)$ component for *Model a* at DE01, DE03, DE04, CH02, GB02, and for *Model c* at SE02 and SE05 respectively. At DE01, the peaks move from winter months to spring ones, while troughs move from summer to winter. At DE03, peaks move from spring to summer, while troughs, after a variable period, seem to converge toward winter

Table 5.3: p values from testing *Model a* versus *Model b*

p-values	approximate F test	Pseudo Likelihood Ratio test
GB02	0.016	0.002
DE01	$8.4e^{-5}$	$8.8e^{-8}$
DE02	0.664	0.739
DE03	0.050	0.014
DE04	0.009	$8.5e^{-4}$
DE05	0.172	0.104
CZ03	0.217	0.149
SE08	0.649	0.722
CH02	$4.3e^{-5}$	$8.8e^{-8}$

Table 5.4: p values from testing *Model c* versus *Model d*

p-values	approximate F test	Pseudo Likelihood Ratio test
SE02	0.006	$4.1e^{-4}$
SE05	0.073	0.025

months. At DE04, the peaks remain steady in winter months, while troughs move from the late summer to early summer months. For CH02, both troughs and peaks remain constant in summer and in winter, apart from a shift around 1995 where the trough moved to winter and the peak to spring. At GB02 the peaks and the troughs seem to stay in winter and summer months respectively until 1992 where they seem to swap, with peaks in summer time and troughs in winter. At SE02 and SE05, the peaks and the troughs seem to move from early winter and summer weeks to late winter and summer weeks respectively.

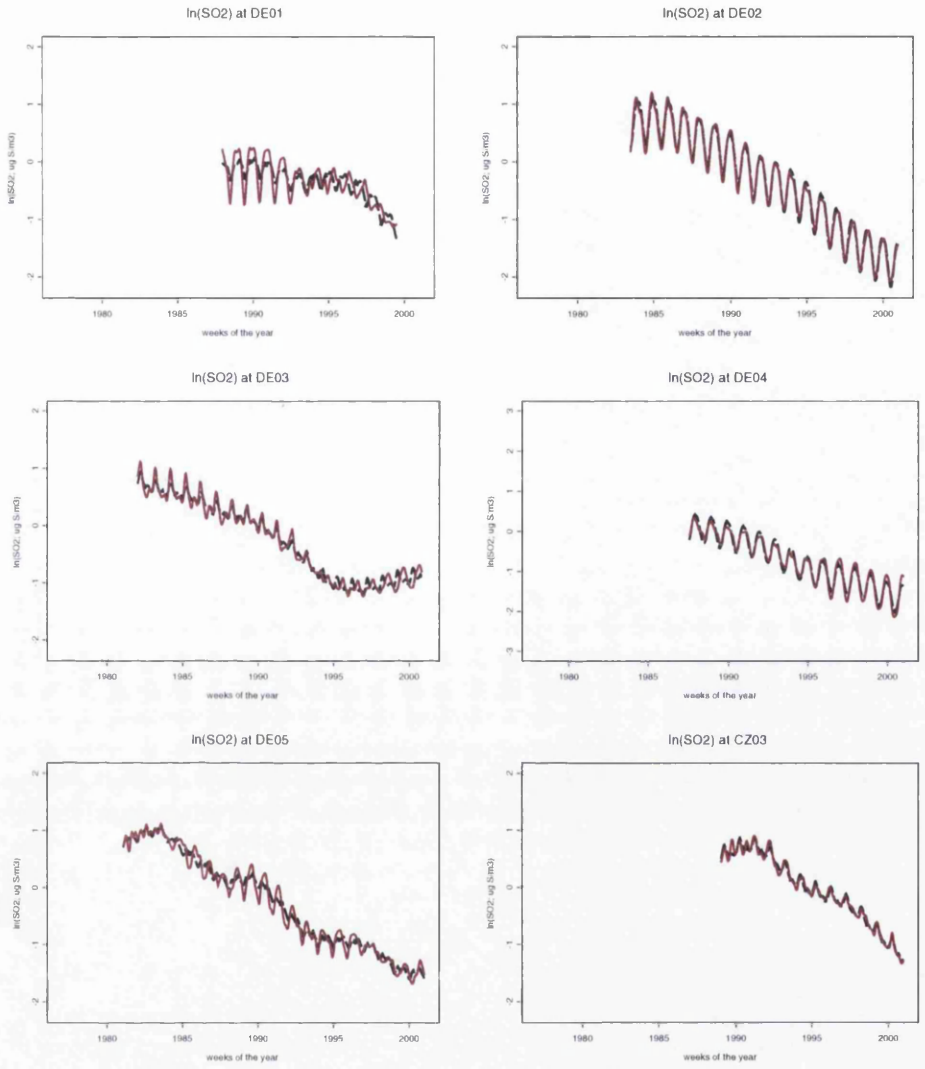

Figure 5.12: The dashed black line shows the $m_y(years) + m_w(weeks)$ component of *Model b*, and the continuous red line shows the $m_{yw}(years, weeks)$ component of *Model a* at DE01, DE02, DE03, DE04, DE05, CZ03.

5.2 Sensitivity analysis of the test for changes in correlation

All the results that have been shown in the previous sections are based on the assumption that the correlation estimate is correct. In contrast with the simulation study, with real data it is not possible to have knowledge of the correlation,

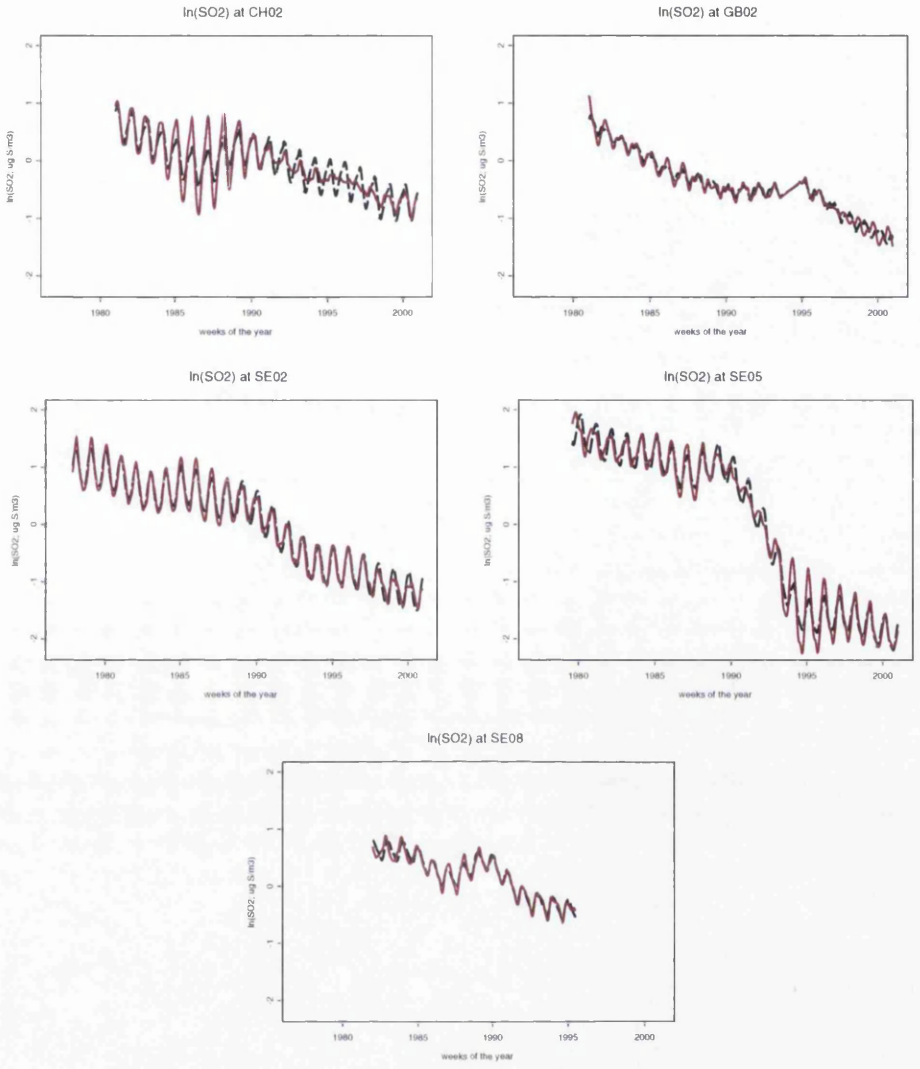

Figure 5.13: The dashed black line shows the $m_y(years) + m_w(weeks)$ component of *Model b* at SE08, GB02, CH02 and *Model d* at SE02 and SE05. The continuous red line shows the $m_{yw}(years, weeks)$ component of *Model a* at SE08, GB02, CH02 and of *Model c* at SE02, SE05.

therefore there could be the risk that the smoothing parameter chosen would produce a biased estimate of the correlation, leading to poorer fits of the models,

and consequently misleading results of the tests.

The models presented in Section 5.1 have been fitted, with estimation of the correlation parameters as 0.20 for Deuselbach (DE04), and 0.27 for Rörvik (SE02). In order to analyze the sensitivity of the results to the correlation estimates, this section presents the results using five different plug in correlation values (0, 0.2, 0.3, 0.4, 0.6), and these results will be compared to those of Section 5.1. Indeed according to the analysis in Section 5.1, the best model at Rörvik (SE02) was *Model c*, while at Deuselbach (DE04) it was *Model a*. This means that at Rörvik (SE02), the meteorological variables are not significant, while at Deuselbach (DE04) the meteorological variables give significant extra information. However, at both sites, the changes in seasonality across years have been identified as statistically significant. Table 5.5 presents the R^2 values for the models fitted at Rörvik (SE02) and Deuselbach (DE04). From Table 5.5 it can be seen that the R^2 decreases as the correlation increases, but the general picture is similar to the one obtained in Section 5.1. With a high correlation (0.6) the R^2 can not be calculated for *Model a* and *Model b* since their residual sums of squares are higher than the ones for the model with no covariates.

Table 5.6 presents the p values that have been obtained from testing the models using different plug-in correlation values, using both the approximate F test (upper value in the cells) and Pseudo Likelihood Ratio test (lower value in the cells). From Table 5.6, it can be seen that at SE02, whatever correlation parameter is used, the meteorology remains not statistically significant when testing *Model a* against *Model c*. This result agrees with the one obtained in Section 5.1. From Table 5.6, it is also clear that testing *Model c* against *Model d*, when the plug-in correlation is below 0.4, the p values agree with the result of

significant change in seasonality across years of Section 5.1. For correlation values greater than or equal to 0.4, significant change in seasonality across years is not identified.

The results from Table 5.6 seem to indicate that at DE04, for correlation values of equal or higher than 0.4, the tests for assessing the change in seasonality give non-significant results, in contrast with results described in Section 5.1. The tests for assessing the statistical significance of the meteorological variables give the same results as Section 5.1, apart from a correlation value of 0.6.

Finally from Table 5.6, for high correlation (0.6), both the approximate F test and the Pseudo Likelihood Ratio test do not seem to work, because the residual sum of squares of the reduced models becomes smaller than that for the full model, implying that the models do not seem to perform well with high correlation.

Table 5.5: R^2 values from the additive models, using ρ=0, 0.2, 0.3, 0.4, 0.6 at SE02 and at DE04

SE02	ρ=0	ρ=0.2	ρ=0.3	ρ=0.4	ρ=0.6
Model a	0.736	0.652	0.588	0.507	0.281
Model b	0.709	0.628	0.566	0.487	0.271
Model c	0.732	0.647	0.583	0.501	0.289
Model d	0.697	0.616	0.555	0.476	0.273

DE04	ρ=0	ρ=0.2	ρ=0.3	ρ=0.4	ρ=0.6
Model a	0.783	0.722	0.675	0.605	-
Model b	0.758	0.701	0.656	0.589	-
Model c	0.628	0.532	0.463	0.382	0.199
Model d	0.596	0.488	0.448	0.430	0.493

Table 5.6: p values from testing the Additive models, using ρ=0, 0.2, 0.3, 0.4, 0.6 at SE02 and at DE04. The values above are obtained from the approximate F test, the ones below from the Pseudo Likelihood Ratio test.

SE02	ρ=0	ρ=0.2	ρ=0.3	ρ=0.4	ρ=0.6
Model a Vs *Model c*	0.677	0.860	0.893	0.890	-
	0.733	0.909	0.925	0.896	0.263
Model c Vs *Model d*	$1.1e^{-9}$	$1.5e^{-4}$	0.016	0.294	0.998
	$1.1e^{-16}$	$6.6e^{-7}$	0.002	0.236	0.999

DE04	ρ=0	ρ=0.2	ρ=0.3	ρ=0.4	ρ=0.6
Model a Vs *Model b*	$1.5e^{-5}$	0.009	0.087	0.602	-
	$3.1e^{-9}$	$7.7e^{-4}$	0.048	0.656	0.562
Model a Vs *Model c*	0	0	0	0	-
	0	0	0	0	1

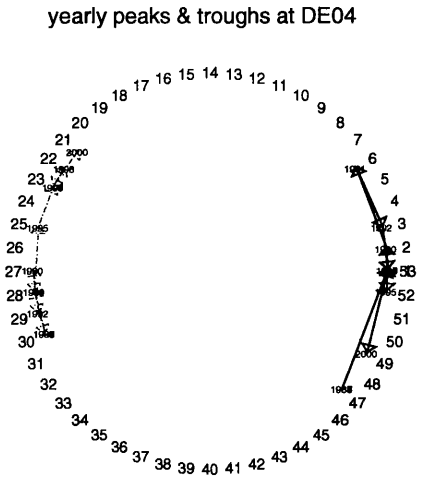

Figure 5.14: yearly troughs (dashed line) and peaks (continuous line) for the estimates of $m_{yw}(years, weeks)$ of *Model a* at Deuselbach (DE04).

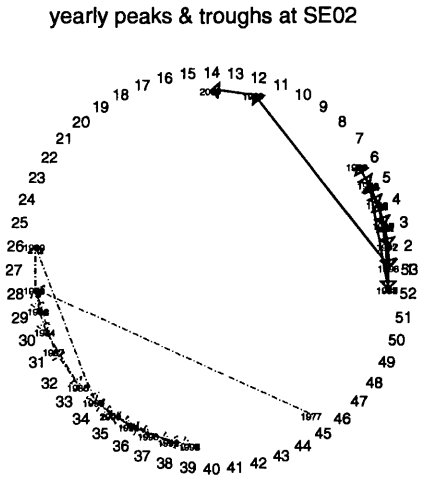

Figure 5.15: yearly troughs (dashed line) and peaks (continuous line) for the estimates of $m_{yw}(years, weeks)$ of *Model c* at Rörvik (SE02).

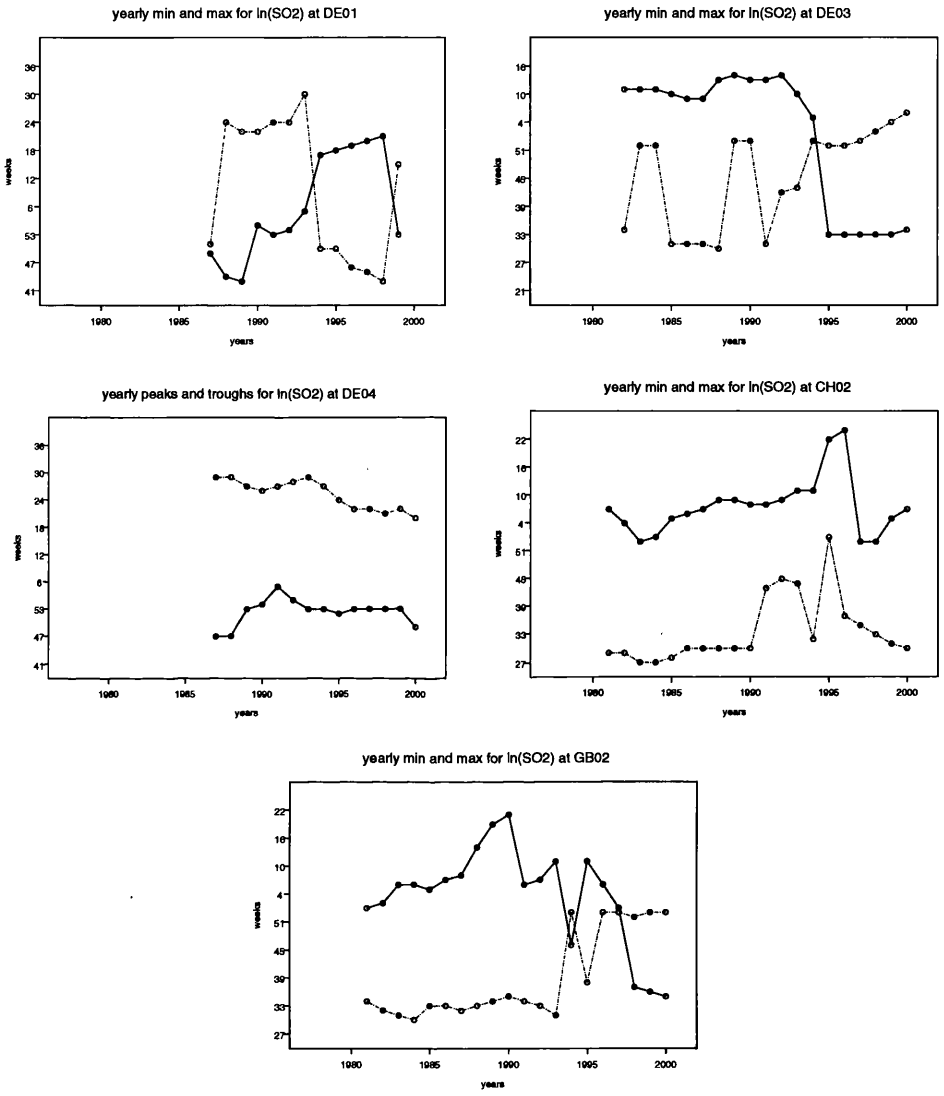

Figure 5.16: Yearly troughs (dashed lines) and peaks (continuous lines) of $m_{yw}(years, weeks)$ component of *Model a* at DE01, DE03, DE04, CH02, GB02.

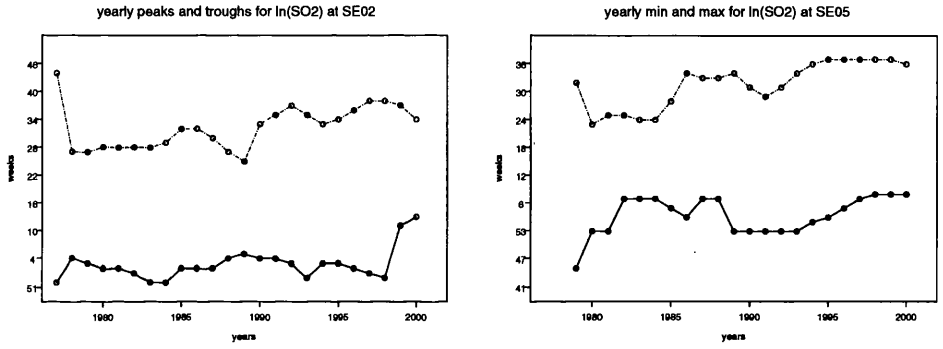

Figure 5.17: Yearly troughs (dashed lines) and peaks (continuous lines) of $m_{yw}(years, weeks)$ component of *Model c* at SE02 and SE05).

5.3 Detecting discontinuities after the removal of smooth seasonality and meteorological effects through additive models

Chapter 3 showed an example of how discontinuities can be detected in a variable of interest. In chapter 3, seasonality was removed by fitting a factored model, with weeks of the year as factors. In the present context, the fit of an additive model suggests another way of deseasonalizing the data. Indeed, from fitting a simple additive model $\ln(SO_2) = \mu + m_y(years) + m_w(weeks) + \varepsilon$ it is possible to obtain a smooth estimate of the component $m_w(weeks)$ which, when subtracted from the observed pollutant concentrations, gives a different deseasonalization procedure. The two different ways of deseasonalizing the data, by a factor model and by a smooth term, could also result in different outcomes for the test results. In fact, the factor model is able to fit abrupt changes over years, while the fit of an additive model would produce a smooth estimate that would not catch these abrupt changes. Therefore some discontinuities could be removed by deaseasonalizing the data through the factor model, and hence not be detected by the test. In other words, by applying the discontinuity test to the deseasonalized data obtained from the additive model fit, it could be possible to detect some discontinuities, that may not be found from fitting a factor model for the seasonal pattern.

Figures 5.18, 5.19, and 5.20 present the discontinuities detected across the 11 sites when seasonalities are removed by factor models (left hand plots), and when seasonalities are removed by smooth terms (middle plots). Table 5.7 shows the p

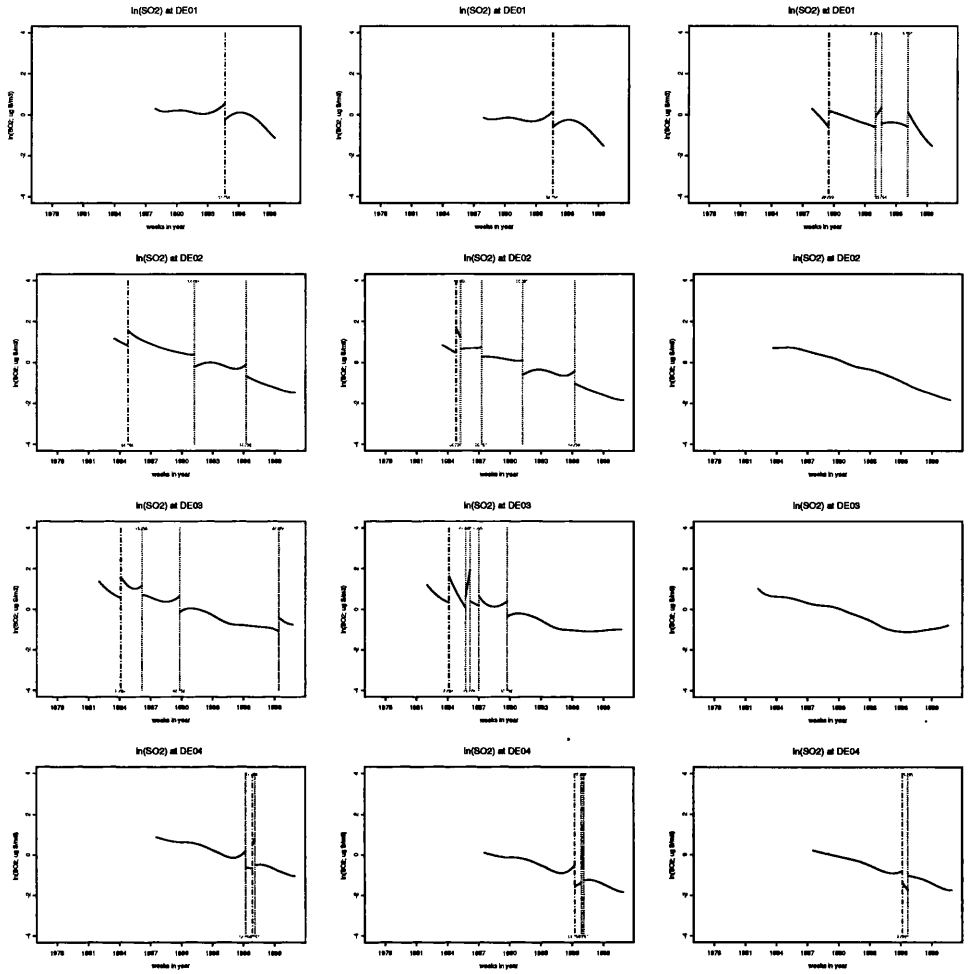

Figure 5.18: Discontinuities detected removing the seasonal pattern by a factor term (left hand plots), by a smooth term (middle plots), and removing meteorology and the seasonal pattern by smooth terms (right hand plots).

values of discontinuity tests computed by removing the seasonal component as a factor term (1st column), and removing the seasonal component as a smooth term (2nd column). From Figures 5.18, 5.19, and 5.20, and from the first two columns of Table 5.7, removing the seasonality with a factor model does not change the conclusions of the test, except for Hoburg (SE08). As expected removing the

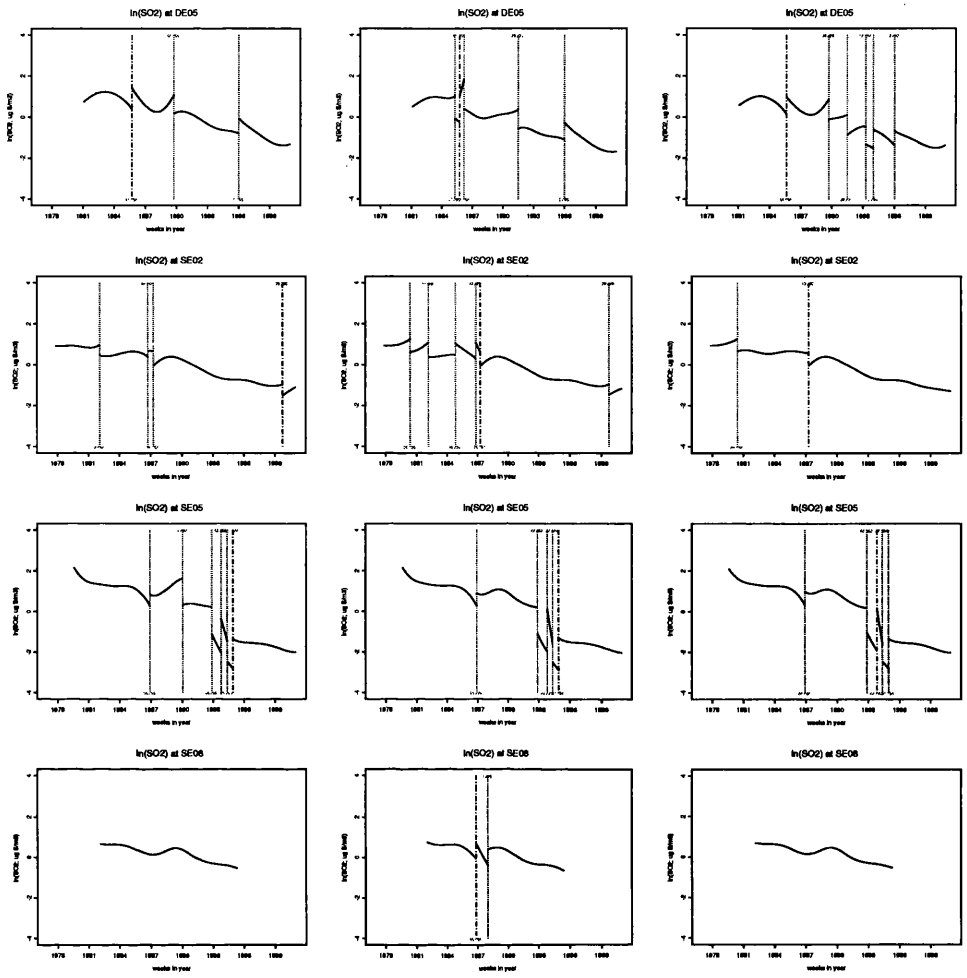

Figure 5.19: Discontinuities detected removing the seasonal pattern by a factor term (left hand plots), by a smooth term (middle plots), and removing meteorology and the seasonal pattern by smooth terms (right hand plots).

seasonality with a smooth term rather than with a factor model, results in a slight increase in the number of discontinuities detected and lower p values obtained by removing a smooth seasonal term. In particular at Hoburg (SE08), the p value is not significant when the factor model is used to remove the trend, while it is lower than 5% when the seasonal term is modeled by a smooth term. In this context

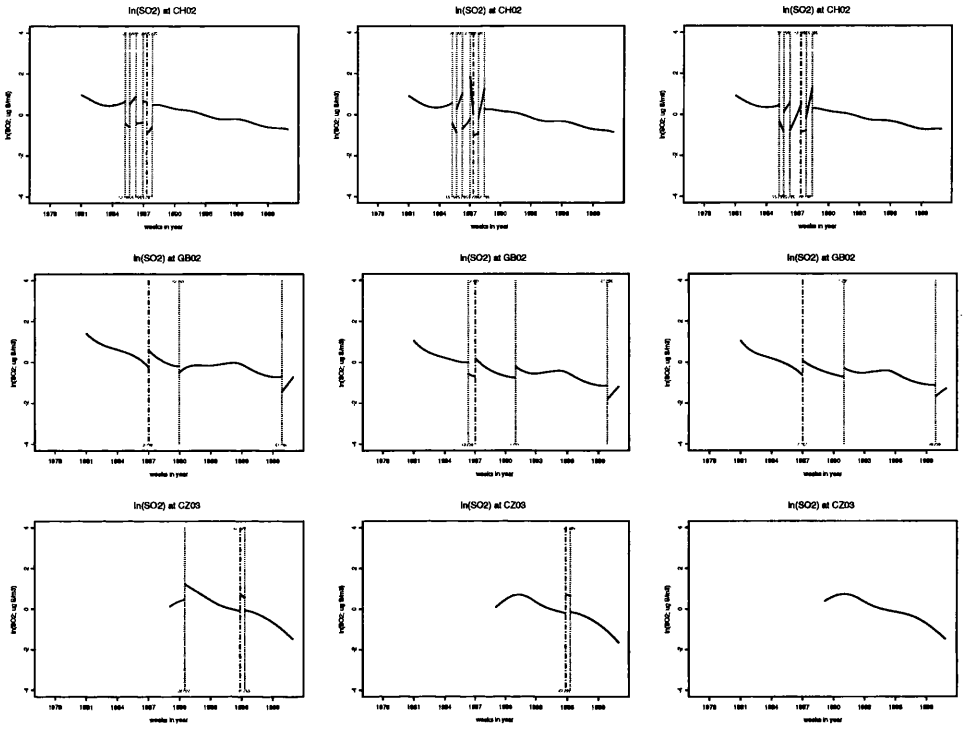

Figure 5.20: Discontinuities detected removing the seasonal pattern by a factor term (left hand plots), by a smooth term (middle plots), and removing meteorology and the seasonal pattern by smooth terms (right hand plots).

some further analysis can be conducted on the cause of discontinuities. In fact chemistry experts suggested that some of these discontinuities could be due to meteorology effects. Therefore it will be interesting to see what would happen in detecting discontinuities after the meteorological effects are removed from the pollutant concentrations. Smooth estimates of the meteorological variables can be obtained from *Model b* that was fitted in section 5, here recalled as:

$$\begin{aligned} \ln(SO_2) = \mu &+ m_y(years) + m_w(weeks) + m_r(rain) + m_t(temperature) \\ &+ m_h(humidity) + m_{w.d.s.}(wind.direction.speed) + \varepsilon \quad Model\ b \end{aligned}$$

Table 5.7: p values of discontinuities tests computed removing the seasonal component as a factor term (1st column), removing the seasonal component as a smooth term (2nd column), removing the seasonal component and the meteorology as smooth terms (3rd column), at Eskdalemuir (GB02), Westerland (DE01), Waldhof (DE02), Schauinsland (DE03), Deuselbach (DE04), Brotjacklriegel (DE05), Kosetice (CZ03), Rörvik(SE02), Bredkälen (SE05), Hoburg (SE08), and Payerne (CH02).

p values	Factor Seasonality	Smooth Seasonality	Smooth Seasonality & Meteorology
GB02	$2.2e^{-4}$	$3.5e^{-4}$	0.025
DE01	0.001	0.001	$1.3e^{-5}$
DE02	0.003	$5.7e^{-4}$	0.211
DE03	$1.0e^{-4}$	$4e^{-4}$	0.067
DE04	$3.8e^{-5}$	$1.8e^{-5}$	0.006
DE05	0.002	$8.2e^{-4}$	0.002
CZ03	0.027	0.015	0.166
SE02	$2.6e^{-4}$	$1.3e^{-4}$	0.011
SE05	$7.1e^{-6}$	$4.9e^{-6}$	$2.9e^{-5}$
SE08	0.072	0.029	0.282
CH02	$1.8e^{-8}$	$1.2e^{-9}$	$1.8e^{-6}$

By fitting *Model b*, it is possible to obtain an estimate of the effect of each covariate and to remove it in order to test for the presence of discontinuities. In the present work, the effect of all the meteorological variables and the seasonal term have been removed from the observations, and the residuals tested for discontinuities.

The right hand plots of Figures 5.18, 5.19, and 5.20 and the right hand column of Table 5.7 present the discontinuities detected and the p values across the 11 sites when seasonalities and meteorological variables are removed by smooth terms. The conclusions of the discontinuity test are different if meteorological effects are removed from the SO_2 concentrations. Indeed at Waldhof (DE02),

Schauinsland (DE03), Kosetice (CZ03) and Hoburg (SE08), the discontinuities detected are no longer significant after removing the effect of the meteorology. As expected as well in the other sites, the number of discontinuities detected is reduced and the p values increase.

5.4 Conclusions

This chapter presents an application of the techniques presented in Chapter 4 to air pollution data. Analysis of sulfur dioxide at Eskdalemuir (GB02, Scotland), Westerland (DE01, Germany), Waldhof (DE02, Germany), Schauinsland (DE03, Germany), Deuselbach (DE04, Germany), Brotjacklriegel (DE05, Germany), Kosetice (CZ03, Czech Republic), Rörvik (SE02, Sweden), Bredkälen (SE05, Sweden), Hoburg (SE08, Sweden), Payerne (CH02, Switzerland), from 1973 up to 2001 shows that the meteorology is statistically significant in explaining the variability of SO_2 across all sites except for two Swedish sites (Rörvik (SE02), Bredkälen (SE05)). The inclusion of meteorological variables changes significantly the trend estimates at a few points. However generally the shapes of trend estimates do not seem to change when meteorology is accounted for. Accounting for meteorology does seem to change the estimates of the seasonal component. Indeed when meteorology is excluded, the seasonal component is estimating part of the seasonal cycle of some meteorological variables. Concerning the seasonal cycle, the analysis also showed that at Payerne (CH02), Brotjacklriegel (DE05), Kosetice (CZ03) and Hoburg (SE08) the seasonality of SO_2 did not change significantly from 1973 to 2000, while it did at Eskdalemuir (GB02), Westerland (DE01), Schauinsland (DE03), Deuselbach (DE04), Payerne (CH02),

Rörvik (SE02) and Bredkälen (SE05). Meteorological variables seem also to be the cause of significant discontinuities detected at Waldhof (DE02), Schauinsland (DE03), Kosetice (CZ03) and Hoburg (SE08).

Therefore it can be finally concluded that for further applications, if the interest lies in analyzing the shapes of trends, there is no need to look at the meteorological variables. On the other hand, if the aim of the analysis is to have fewer frequency components, such as seasonal cycles, peaks and troughs, meteorological variables represent significant components to be accounted for in the model.

Chapter 6

Spatiotemporal Analysis

6.1 Introduction

In the previous chapters, the main problems associated with modeling air pollution have been presented. Fitting and testing additive models were reviewed in chapter 2 and their generalizations to deal with correlated data were presented in chapter 4. These models have been used to analyze the effects of meteorological variables on SO_2 (Chapter 5) showing that in most cases meteorological variables have a statistically significant effect in explaining the variability of SO_2. However the estimated trend does not seem to change dramatically when meteorological information is not accounted for in the model.

However the analysis performed up to now has focused entirely on the time series analyses of SO_2, without considering the spatial pattern. The aim of this chapter is to describe the spatiotemporal SO_2 trend through additive models accounting for spatiotemporal correlation, using time trend, seasonality and spatial locations as covariates. In order to determine the spatiotemporal correlation,

two marginal analyses have been performed: a spatial analysis across time, and a time series analysis across space. These spatiotemporal additive models will also be compared through an approximate F test that accounts for spatiotemporal correlation.

The data analyzed in this chapter are the natural logarithms of the monthly means of sulphur dioxide concentrations $\ln(SO_2)$, monitored from 1990 to 2001 at 130 sites across Europe. The decision to focus on these 12 years of data is a consequence of the low percentages of missing values present after 1990. The data are therefore quite dense in time, but sparse in space, which is a common feature in many spatial problems.

There are a number of approaches to modeling space-time data in the statistical literature and the following section briefly reviews some approaches relevant to the analysis presented here.

6.2 Literature review of spatiotemporal trend analysis

Environmental processes, such as atmospheric pollutant concentration, are characterized by spatial and temporal variability. In order to analyze similarities and differences between the approaches that are present in the literature, the following notation will be used consistently across the different techniques described below. Whether the aim of the analysis is description, inference or prediction, space-time data are often modeled as a realization of a spatiotemporal random function (RF) $Z(s,t)$, indexed in space by $s \in D \subseteq R^d$ and in time by $t \in T \subseteq R$.

The space-time covariance function is defined as:

$$c(s_1, s_2; t_1, t_2) = cov[Z(s_1; t_1), Z(s_2; t_2)], \quad s_1, s_2 \in R^d, t_1, t_2 \in R \tag{6.1}$$

A useful introduction to the main approaches available in the literature is provided by Gneiting and Schlather (2002). They propose the classification of space-time modeling, and space-time covariance functions, into two general approaches:

1. *Geostatistical methods.* These methods give highest priority to the fitting of the space-time covariance function, that is usually expressed in simple or closed form.

2. *Model based approach.* The choice of the space-time covariance function is subordinated to the choice of the stochastic model that has higher priority. Therefore the covariance function can range from simple to very complex forms.

Therefore geostatistical approaches are conceptually simple, but may not offer the flexible covariance structures of the model based approaches. Subsequent sections will review the main techniques presented in the literature for each of the two approaches. The geostatistical approach will be described in more detail than the model based approach, since in the following sections, a technique that can be characterized as a geostatistical approach will be proposed.

In the following sections more emphasis will be given to the different approaches to the analysis of the spatiotemporal correlation structure. This is an active area of research and there are many studies concerning spatiotemporal non-stationarity.

6.2.1 Geostatistical space-time models

Since the principal feature of geostatistical space-time models is the covariance function, there are two main groups that can be distinguished: separable models and non-separable models. The difference between separable and non-separable models concerns the structure of the covariance function (6.1). In particular separable models make the assumption that expression (6.1) can be simplified as follows:

$$\begin{aligned} c(s_1, s_2; t_1, t_2) &= cov[Z(s_1; t_1), Z(s_2; t_2)], \quad s_1, s_2 \in R^d, t_1, t_2 \in R \\ &= cov[Z(s_1; s_2)] + cov[Z(t_1; t_2)] \qquad (6.2) \\ &or \\ &= cov[Z(s_1; s_2)]cov[Z(t_1; t_2)] \qquad (6.3) \end{aligned}$$

In other words, expressions (6.2) or (6.3) assume that the spatial covariance structure is the same across time, and the temporal covariance matrix is the same across space. Non-separable models do not make this assumption and give the space-time covariance function a more general form. This section is divided into two parts: a review of separable models, and a review of non-separable ones.

6.2.1.1 Geostatistical space-time models: the separable approach

Kyriakidis and Journel (1999) presented a useful review of geostatistical space-time separable models. They distinguished two approaches of modeling spatiotemporal data.

1. The first approach consists in modeling a single spatiotemporal function

$Z(s,t)$, typically decomposed into a trend component, describing smooth patterns, and a stationary residual component, describing higher frequency fluctuations.

2. The second approach analyzes multiple vectors of spatial functions or vectors of time series. Within this approach, it is then possible to distinguish: models that treat the spatiotemporal process $Z(s,t)$ as a collection of a finite number T of temporally correlated space functions $Z(s)$, and models that view the $Z(s,t)$ as a collection of a finite number N of spatially correlated time series $Z(t)$.

1) Single spatiotemporal model.

Kyriakidis and Journel (2001a) propose a methodology for stochastic spatiotemporal modeling applied to atmospheric pollution. They summarized their approach in five steps:

1. *Station specific temporal trend models.*
2. *Regionalization of temporal trend coefficients.*
3. *Simulation of spatiotemporal trend.*
4. *Location specific temporal residual models.*
5. *Simulation of spatiotemporal residuals.*

Kyriakidis and Journel (2001b) applied this spatial time series methodology to monthly averaged daily values of particulate sulphate SO_4 dry deposition over Europe. The data were provided by the Norwegian Institute for Air Research (NILU), and were collected through the Cooperative Program for Monitoring

and Evaluation of the Long-Range Transmission of Air Pollutants in Europe (EMEP).

There are several other works which follow the single spatiotemporal model. Bogaert and Christakos (1997) analyzed the spatiotemporal pattern of calcium, chloride and nitrate which provide important indicators of water contamination. The regression space-time model they proposed consists of a purely spatial component, a purely temporal component, a space homogeneous-time stationary component, and a space-time mean function. In order to make the model operational and describe its main features, Bogaert and Christakos (1997) assumed that the spatiotemporal covariance function was separable, the random components of the regression model were statistically independent, a parametric expression for the mean was available and the spatial locations remained the same for all time points.

Luo et al. (1998) used "smoothing splines ANOVA" to produce spatiotemporal estimates of air temperatures data monitored by a network of stations. They show how the "smoothing spline ANOVA" can correct for the biases that result from the usual smoothing spline methods due to the incompleteness of sampling over time.

Kammann and Wand (2003) presented *geoadditive models* obtained from the fusion of geostatistical and additive models. There are several ways to combine the ideas of geostatistics and additive modeling. They propose to incorporate a geographical component expressing kriging as a linear mixed model and merging it with an additive model to obtain a single mixed model.

Christakos (1992) developed a spatiotemporal Random Field (RF) for analyzing complicated spatiotemporal deposition trends. A mathematical operator

Q can be defined in space-time that transforms the RF $Z(s,t)$ to a zero mean homogeneous-stationary process $Y(s,t) = Q[Z(s,t)]$. The Q operator filters out any existing space-time trends. The ordinary covariance $c(s_1,t_1;s_2,t_2)$, which is generally space inhomogeneous and time nonstationary, can be decomposed into a spatially homogeneous and temporally stationary part $g(h,\tau)$ ($h = s_1 - s_2$ and $\tau = t_1 - t_2$), which is called the generalized spatiotemporal covariance, and a polynomial in s_1, s_2, t_1, t_2.

Vyas and Christakos (1997) applied this spatiotemporal random field model to sulphate deposition over Eastern Europe. Christakos and Vyas (1998) applied the same spatiotemporal random field model to ozone concentration over Eastern Europe. Their analysis showed that temporal and spatial variations cannot be separated in simple ways, as they interact and influence each other. Theoretical arguments as well as numerical results show that composite spatio-temporal deposition maps lead to improved estimates of concentrations compared to purely spatial or purely temporal analysis.

In later work, Christakos and Serre (2000) proposed a Bayesian version of the spatiotemporal random field model, called Bayesian Maximum Entropy (BME). They discussed an application to particulate matter concentration (PM_{10}) in the state of North Carolina. The PM_{10} maps show significant variability both spatially and temporally, a finding that may be associated with geographical characteristics, climatic change, seasonal patterns and random fluctuations.

2) Multiple vectors of spatial functions or vectors of time series.

The second approach views the spatiotemporal process $Z(s,t)$ as a set of temporally correlated spatial functions, or a set of spatially correlated time series.

Sampson and Guttorp (1992) presented a nonparametric method for estimating the spatial covariance assuming temporal stationarity, but neither spatial isotropy nor spatial stationarity. The model is constructed in two steps. Firstly, multidimensional scaling (MDS) is used to generate, from the original geographical coordinates (also called the G plane), a two-dimensional coordinate representation of the sampling stations (also called the D plane) whose spatial dispersion is stationary and isotropic. Secondly, a thin plate spline interpolation is used to relate the two coordinate systems. Estimates of the covariance between observations at any two locations are smoothed functions of the geographical coordinates.

Guttorp et al. (1994) examined hourly ozone data at 17 sites around the Sacramento area. Their analysis showed a different covariance structure between night-time and day-time and therefore the spatial and temporal correlation structures of the residuals could not be assumed separable.

In later work, Meiring et al. (1998) applied a space deformation approach to ozone data and found a diurnally varying covariance structure. The covariance structure is clearly nonstationary and cannot be separated into purely spatial and purely temporal components. The focus of their analysis was mainly the spatial structure of the residuals from site-specific time series models. The procedure proposed consists firstly in a temporal pre-whitening of the time series at multiple monitoring stations and then the computation of spatial and space-time covariances between the pre-whitened series at different sites.

Mardia and Goodall (1993) apply kriging after detrending the data and covariance transformations in order to view the data as repeated measurements in space. When the data showed non stationarity and anisotropy in space, they apply a space deformation in order to obtain new coordinates for the monitoring

stations so that the spatial covariance became stationary and isotropic.

6.2.1.2 Geostatistical space-time models: the nonseparable approach

Among geostatistical approaches, Gneiting and Schlather (2002) present a review of nonseparable covariance functions. Nonseparable covariance functions can be modeled using mainly three different approaches: models based on space-time metrics, physically based models, and models based on Fourier analysis.

Among the physically based models, Cox and Isham (1988) present a physical model for rainfall. Jones and Zhang (1997) discuss the space-time covariance functions associated with the solutions of certain stochastic partial differential equations. Brown et al. (2000) build on physical dispersion models which could correspond to phenomena such as the spread of an air pollutant. These models are again generated by stochastic differential equations. The physical background of these models is appealing but the approach does not readily lead to closed form expressions for the space-time covariance functions.

Among the models based on Fourier analysis, the approach of Cressie and Huang (1999) focuses on the analytical derivation of covariance functions through Fourier inversion. Gneiting (2002) provides a Fourier-free implementation of the Cressie and Huang (1999) approach and enlarges the class of valid spatiotemporal covariance functions.

6.2.2 Model Based Approaches

A good overview of the model based approach is given by Diggle et al. (2002). They set out the basic methodologies for dealing with geostatistical problems.

Diggle et al. (2002) distinguish two main cases of study: the Gaussian and the non-Gaussian model. The first is based on the assumption that the "signal process" is Gaussian, while the second relaxes this assumption. For the Gaussian model two main methods of inference can be distinguished: parametric and Bayesian methods. In the parametric approach estimation is based on variogram analysis. Bayesian methods of inference for Gaussian models treat parameters in the model as random variables, allowing for parameter uncertainty in predictive inference. Among the non-Gaussian models, Generalized Linear Spatial Models (GLSM) represent a widely used class of models.

Huerta et al. (2004) propose a model within the Bayesian framework by using dynamic linear models to analyze hourly ozone levels in Mexico City accounting for temporal non-stationarities in the data. Markov Chain Monte Carlo methods are used to produce predictions in time and interpolation in space.

Wikle et al. (1998) illustrate the Bayesian hierarchical view in space-time settings. A flexible five-stage hierarchical model is presented by the authors: "The first stage of the model specifies a measurement error model for the observational data. The second stage of the model allows for site specific time series models and the incorporation of space-time dynamics. In the third stage, the parameters of the site specific time series models are described with priors through a Markov random field that generates spatial dependence structures. The final two stages complete the Bayesian formulation by specifying priors on the parameters". The aspect that is central to this article and distinguishes this approach from the other hierarchical space-time formulations is the third stage, where the dynamic terms are modeled. The Bayesian hierarchical strategy that they propose allows complicated structures to be modeled in terms of means at various stages, rather

than a model for a massive joint covariance matrix.

Wikle and Cressie (1999) try to combine the geostatistical approach with the model based approach. In fact the former approach is limited in that difficulties often arise in the specification and implementation of realistic covariance functions. They therefore combine both approaches through the spatiotemporal Kalman filter. The Kalman filter, commonly used by control engineers and other physical scientists, has been successfully used in such diverse areas as the processing of signals in aerospace tracking and underwater sonar and the statistical control of quality. Meinhold and Singpurwalla (1983) present an interesting article on how the Kalman filter can be easily understood and useful to statisticians if a Bayesian formulation and some well known results in multivariate statistics are used. The essential difference between the Kalman filter and the conventional linear model representation is that, in the former, the state of nature analogous to the regression coefficients of the latter is not assumed to be a constant but may change with time.

Brown et al. (2000) proposed a non-separable spatiotemporal model based on a physically sensible dispersion model which can model phenomena such as the spread of an air pollutant. The model consists in producing spatial maps that evolve in time by "blurring" the values at the previous time point and adding a spatial random field.

Oehlert (1993) presented a spatiotemporal model to estimate the ability to detect trends in wet-deposition sulphate, using data monitored in North America. This model includes spatial and temporal correlation among the monitoring stations and provides a way to estimate regional values from a scattering of stations by using a discrete smoothing prior. The first problem addressed by this article

is to determine the variance of regional trend estimates for response variables of interest, while the second considers the addition and the deletion of stations to the current network. The first involved modeling the spatial covariance (between stations) using an exponential model, while the second one was based on two criteria: minimizing the average regional variance, and keeping the largest of the local variances from becoming too large.

Studies conducted by Shannon (1999) regarding data collected during the last 16 years in US and in Canada, and by Lynch et al. (1995) on 13 years of data from the National Atmospheric Deposition Program (NADP), used regression analysis to reveal a considerable inter- and intra-regional variability, even in small regions with multiple sites. These studies showed that if, on the one hand, regional averaging is recommended to remove small scale variability, on the other hand, it is possible to obtain non-significant estimates at regional scale even if significant estimates of trends are obtained at the individual station level.

The approach taken in this work differs from the ones presented in this literature review because it tackles the spatiotemporal problem using a nonparametric approach that does not make any distributional assumptions about the data. The methodology proposed here describes the spatial pattern of the data through a smooth surface that can change its mean level smoothly over time, taking into account the spatiotemporal correlation. The flexibility of this model comes at the price that a separable model is assumed for the error term and that the spatial surface does not change in shape over time.

6.3 Spatial analysis across time

This section will present an analysis of the spatial trend at 130 sites in Europe across each month from 1990 to 2001. Time series analysis of the spatial correlation structure will be also performed.

6.3.1 Spatial analysis across time: Notation

The model assumed for the data is:

$$z_{ij} = Z_j(s_i) = \mu_j(s_i) + \eta_j(s_i); \quad i = 1, \ldots, n; \; j = 1, \ldots, t,$$

where $n = 130$ is the number of sites, $t = 144$ is the number of time points and $s_1, \ldots, s_n$ denote the positions of the monitoring sites. μ_j represents the spatial trend at time j, and η_j is the error term at time j whose expected value is zero. (It should be noted that not all sites monitored SO_2 over the period of analysis and some values of z_{ij} are therefore missing). The process is said to be (weakly) stationary if $E\{Z_j(s_i)\} = \mu_j$, and $Cov\{Z_j(s_i), Z_j(s_{i'})\} = C(s_i - s_{i'})$ (or equivalently $Cov\{\eta_j(s_i), \eta_j(s_{i'})\} = C(s_i - s_{i'})$), and it is said to be intrinsically stationary if $E\{Z_j(s_i)\} = \mu_j$, and $Var\{Z_j(s_i) - Z_j(s_{i'})\} = 2\gamma(s_i - s_{i'})$ (or equivalently $Var\{\eta_j(s_i) - \eta_j(s_{i'})\} = 2\gamma(s_i - s_{i'})$). The spatial processes η_j are modeled here through semivariogram functions γ_j given by:

$$2\gamma_j(s_i - s_{i'}) = Var\{\eta_j(s_i) - \eta_j(s_{i'})\} \tag{6.4}$$

where s_i and $s_{i'}$ are the locations of two stations. If the argument of γ_j depends only on $d = ||s_i - s_{i'}||$, where the norm $||.||$ usually represents the standard

Euclidean metric, the process is said to be isotropic.

This model raises the questions of what appropriate structures should be used for the trend functions μ_j and the variograms γ_j. In particular, evidence for differences in the γ_j functions over j has important implications for the construction of a suitable spatiotemporal model. These questions will be addressed below.

Since spatial analysis is based on distances between sites, the measure of distance used must be considered. The locations of the sites are expressed in latitude and longitude. In order to account for the curvature of a globe, the location coordinates will be translated to the coordinates of a tangent plane to the north pole. The translations have been done using the Lambert (or Schmidt) projections that translate equal areas on the surface of the spherical map to equal areas on the plane projection (Fisher et al., 1993). The distances between sites can then be computed using Euclidean distances. Indicating latitude with θ and longitude with ϕ, the Lambert projections are given by:

$$x = 2\sin\left(\frac{1}{2}(90-\theta)\right)\cos\phi \tag{6.5}$$

$$y = 2\sin\left(\frac{1}{2}(90-\theta)\right)\sin\phi \tag{6.6}$$

All the results in this paper are obtained on the projected plane, and are translated back to the original latitude and longitude for visualization.

6.3.2 Spatial analysis across time: model fitting

Exploratory graphs (Figure 6.1a) of the observed values of $\ln(SO_2)$ clearly indicate the presence of spatial trends.

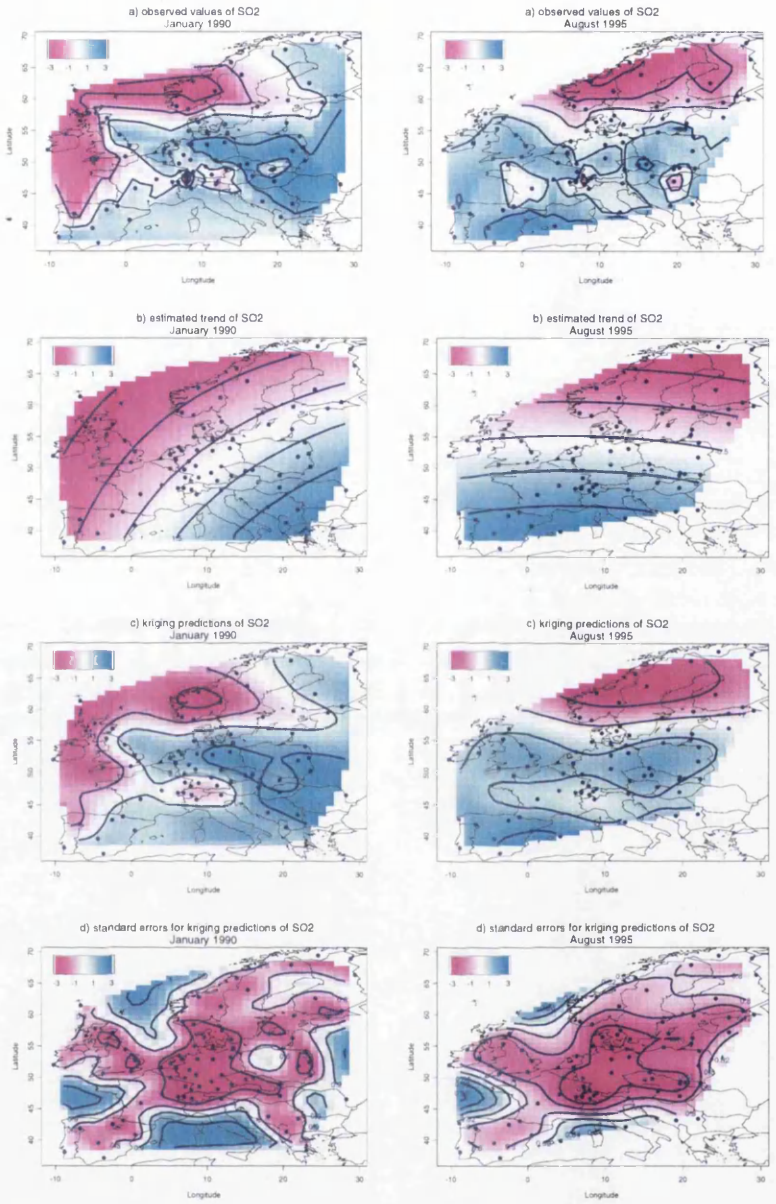

Figure 6.1: Contour plots of: a) observed values, b) estimated trend, c) kriging predictions, d) standard errors for kriging predictions for SO_2 in January 1990 (left hand side) and in August 1995 (right hand side).

In order to construct variograms to model the spatial covariance with an assumption of intrinsic stationarity, the trends must be removed. However, a unique methodology for trend detection and identification does not exist. In this analysis four different trend models have been fitted:

- Model 1: $\ln(SO_2)_{ij} = \beta_{0j}+\beta_{1j}x_i+\beta_{2j}y_i+\varepsilon_{ij}$, for each time point $j = 1,\ldots,t$, across n sites $i = 1,\ldots,n$.

- Model 2: $\ln(SO_2)_{ij} = f_j(x_i, y_i)+\varepsilon_{ij}$, for each time point $j = 1,\ldots,t$, across n sites $i = 1,\ldots,n$.

- Model 3: $\ln(SO_2)_{ij} = \alpha + \pi_{k_j} + \varphi_{l_j} + \nu_i + \varepsilon_{ij}$, where π_k, φ_l, ν_i denote factor levels for month, year and site, and the notations k_j and l_j indicate that the month index k and the year index l are identified from the index j of the time point.

- Model 4: $\ln(SO_2)_{ij} = \alpha + \pi_{k_j} + \varphi_{l_j} + \nu_i + (\pi\varphi)_{k_j l_j} + (\pi\nu)_{k_j i} + \varepsilon_{ij}$, where years, months and site codes are again treated as factors.

Model 1 describes the spatial structure with a different plane at each time point ($j = 1,\ldots,t$). Model 2 allows the spatial structure to be a smooth surface at each time point. Models 3 and 4 describe the spatial structure by including year, month and site as factors. Both these models assume that the time trend is the same across sites, but Model 3 assumes in addition that the seasonality is the same across sites and across years, while Model 4 allows the seasonality to change across sites and across years. Model 4 was the most general factor model fitted, since it was not possible to fit the interaction term between year and site, due to the fact that not all the sites have monitored SO_2 across the same years.

The fitting of Models 1, 3 and 4 was carried out using least squares, while Model 2 has been fitted by a bivariate local linear regression smoothing procedure (Bowman and Azzalini, 1997). For n sites with coordinates $(x_i, y_i), i = 1, 2, \ldots, n$, the bivariate local linear regression smoother involves solving the weighted least squares problem:

$$\min_{\alpha_j, \delta_j, \tau_j} \sum_{i=1}^{n} \{\ln(SO_2)_{ij} - \alpha_j - \delta_j(y_i - y) - \tau_j(x_i - x)\}^2 w_1(y_i - y; h_1) w_2(x_i - x; h_2) \tag{6.7}$$

for each time point j. The estimate of the surface at position (x, y) is given by the value of $\hat{\alpha}_j$. The smoothing parameters h_1 and h_2 control the width of the kernel functions $w_1(.)$ and $w_2(.)$ respectively. Here the smoothing parameters have been chosen subjectively as $h = (h_1, h_2) = (0.06, 0.06)$.

For the residuals of each of the four models fitted, the variogram has been computed using a "robust" alternative version of the usual variogram proposed by Cressie (1991):

$$2\tilde{\gamma}(d) = \frac{1}{0.457 + \frac{0.494}{n}} \left\{ \frac{1}{|N(d)|} \sum_{N(d)} |Z(s_i) - Z(s_k)|^{1/2} \right\}^4 \tag{6.8}$$

where, given a finite number of observations $s = s_1, \ldots, s_n$, $N(d)$ denotes a collection of (s_i, s_k) pairs of sites whose Euclidean distance lies within a given neighborhood of d, and $|.|$ denotes cardinality. The semivariogram estimate (6.8) is an approximately unbiased estimator when the data are normally distributed, but is less affected by outliers than the common estimator of the semivariogram. Figure 6.2 gives examples of the empirical variograms of the residuals from the

four models, at two different time points (January 1990, August 1995). In these figures, as in most of the cases that have been observed, the fitting of Model 1, $\ln(SO_2)_{ij} = \beta_{0j} + \beta_{1j}x_i + \beta_{2j}y_i + \varepsilon_{ij}$, seems to be the only trend surface that gives a bounded variogram which is monotonically increasing. Models 3 and 4 give unbounded variograms at most of the time points, which apparently suggests that spatial correlation is always present between any two points across Europe. Since the variograms increase indefinitely with increasing lag distance, the process is not second order stationary and the covariance does not exist. One possible explanation for this is that the residuals from Model 3 and Model 4 still include some trend, since these two models assume that the time trends are the same across sites. The variograms from Model 2 show no evidence of positive correlation, which suggests that this model is overfitting the data, capturing not only the trend but also part of the correlation structure as well. The most plausible trend estimate is therefore the plane (Model 1). The bounded pattern of the variograms obtained from Model 1 provides a suitable model of the underlying spatial covariance.

The results for Model 1 could be obtained from Model 2 by using a very large smoothing parameter. Indeed, the local linear regression smoother can be viewed as a relaxation of the usual linear regression model. Indicating with h the two smoothing parameters of equation (6.7), $h = (h_1, h_2)$, and changing their values by the same amount, it is clear that as the smoothing parameter h becomes very large, the curve estimate approaches the fitted least squares regression surface. It is appealing to have this standard model within the nonparametric formulation. Figure 6.3 shows the empirical variograms of the residuals obtained by fitting a bivariate local linear regression smoother to the data for January 1990 with

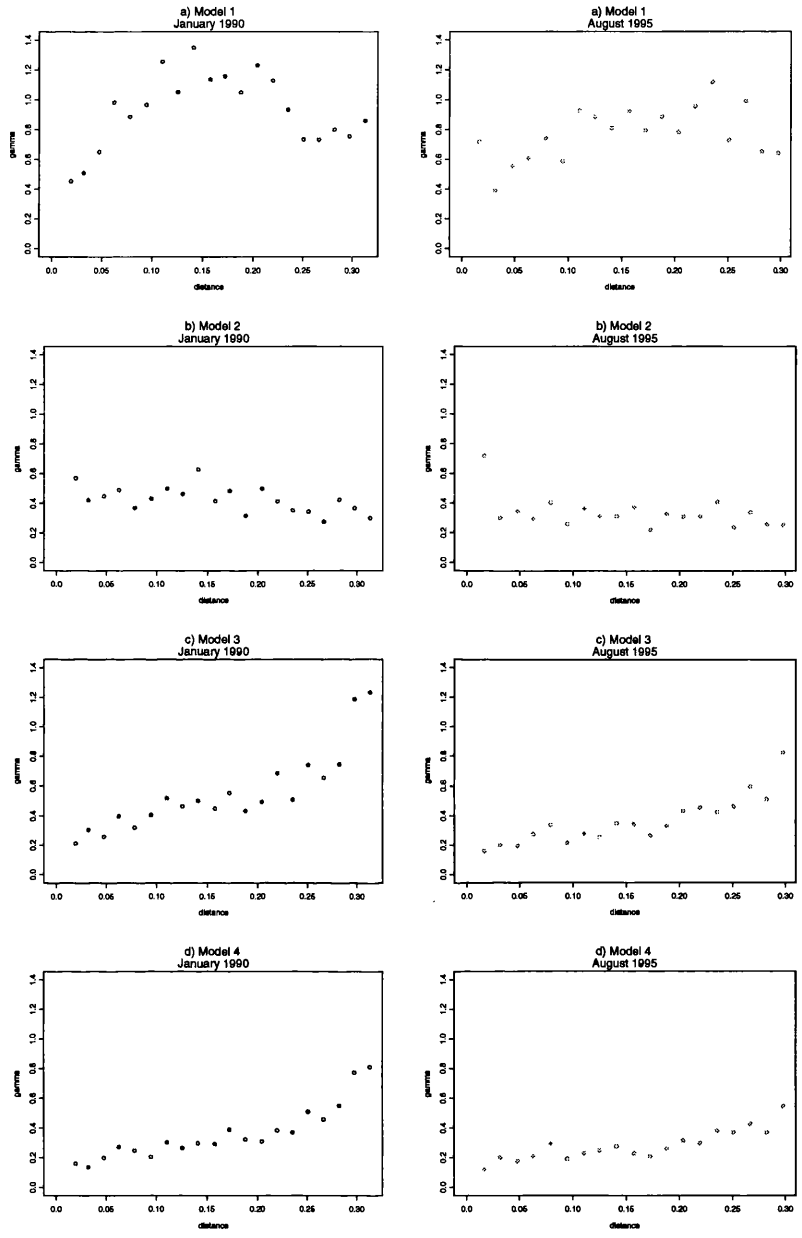

Figure 6.2: Observed variograms of the residuals from: a) Model 1, b) Model 2, c) Model 3, d) Model 4 for SO_2 in January 1990 (left hand side) and in August 1995 (right hand side).

different smoothing parameters. As the smoothing parameter increases, the shape of the variogram approaches the one obtained from the plane. A visual impression

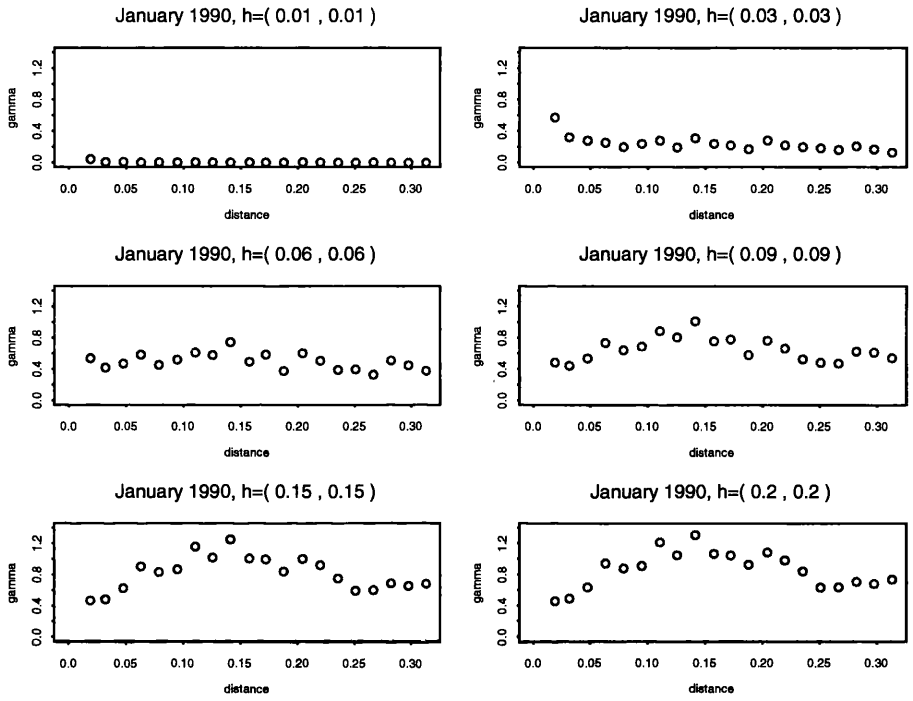

Figure 6.3: Observed variograms of residuals obtained from a bivariate local linear regression smoother with different smoothing parameters of January 1990.

of the fit of the plane of Model 1 is given by Figure 6.1 b. Generally, the fits seem to agree quite well with the observed values shown in Figure 6.1 a.

Having established that Model 1 describes the trend, three different theoretical variograms were fitted to the residuals:

- Spherical model

$$\gamma_0(d) = \left\{ \begin{array}{cc} 0, & d = 0, \\ c_0 + c_1 \left\{ \frac{3}{2}\frac{d}{R} - \frac{1}{2}\left(\frac{d}{R}\right)^3 \right\}, & 0 < d \leq R, \\ c_0 + c_1 & d \geq R \end{array} \right\} \tag{6.9}$$

- Exponential model

$$\gamma_0(d) = \left\{ \begin{array}{cc} 0, & d = 0, \\ c_0 + c_1(1 - e^{-d/R}), & d > 0, \end{array} \right\} \tag{6.10}$$

- Gaussian model

$$\gamma_0(d) = \left\{ \begin{array}{cc} 0, & d = 0, \\ c_0 + c_1(1 - e^{(-d/R)^2}), & d > 0, \end{array} \right\} \tag{6.11}$$

In each of the expressions (6.9), (6.10), (6.11), $R > 0$, a scale parameter, is the *range*, $c_0 \neq 0$ is the *nugget* effect, and c_1 is the *sill*. All the resulting functions are negative definite.

Each of them has been fitted by Cressie's weighted least squares procedure (Cressie, 1991): given a sample variogram $\hat{\gamma}(d)$ evaluated at a finite number of values of d, say $d_1, d_2, \ldots$, and a model $\gamma(d; \lambda)$ depending on unknown parameters λ, where λ is chosen to minimize:

$$\sum_i |N(d_i)| \left\{ \frac{\hat{\gamma}(d_i)}{\gamma(d_i; \lambda)} - 1 \right\}^2 \tag{6.12}$$

This method is not dependent on a particular sample estimator. It is relatively

straightforward to calculate through nonlinear optimization, and no complicated likelihood evaluation is required. A practical disadvantage is that there is no easy way to obtain standard errors for the estimators, or to test hypotheses about the parameters.

To minimize expression (6.12), the Gauss-Newton algorithm has been used. Results from fitting the three theoretical variograms at each time point seem to indicate that the Gaussian model can be easily fitted in most cases, while with the Spherical and the Exponential models the algorithm does not converge at some time points. Figure 6.4 shows the fit of the estimated Gaussian variograms, obtained from the residuals of Model 1, for January 1990 and August 1995.

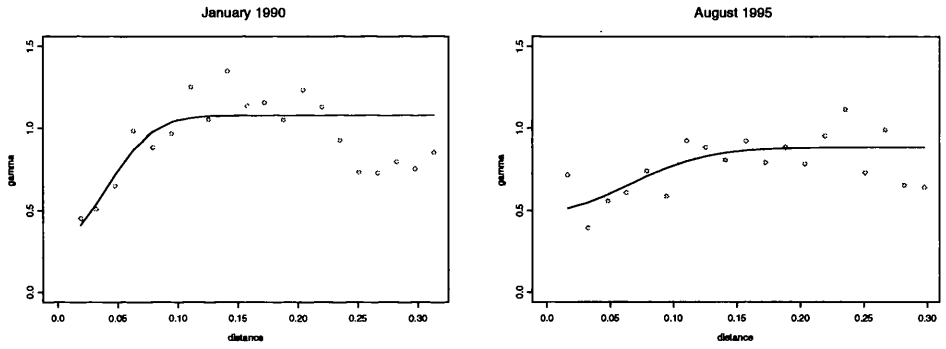

Figure 6.4: Fitting Gaussian variograms to the residuals of SO_2 from fitting Model 1 to monthly means of January 1990 and August 1995.

Once the variograms for each time point have been estimated, an ordinary kriging procedure has been used. Figure 6.1 c and d show respectively the predictions and the standard errors for kriging predictions using contour plots. The predicted values of Figure 6.1 c have been obtained by adding the ordinary kriging predictions to the plane surface shown in Figure 6.1 b. The predictions seem

to agree quite well with the main features of the data.

6.3.3 Spatial analysis across time: parameter estimates

It is now of interest to consider the temporal dimension of the data by first inspecting the pattern of the coefficients of the plane for projected latitude and longitude and to assess if they vary over time.

Figure 6.5 shows time series plots of the estimates of the intercept (β_0), the coefficients for x (β_1) and for y (β_2) of Model 1. Smooth estimates of trend (local linear regression smoothers) for each time series are plotted (continuous lines), together with global averages (dashed lines) and reference bands for no effect. A test for no effect of the parameters (introduced in Section 4.4.4) has been implemented and p values for the test applied to the Model 1 parameters are respectively $0.342, 0.358, 0.819$ (Pseudo Likelihood Ratio test). From the test results and from Figure 6.5 it is clear that no significant changes across years are affecting Model 1 parameters. Figure 6.6 shows the plots of the estimated monthly parameters of the variograms (range, sill, and nugget effect) across time. Figure 6.6 also shows smooth trends, global average and a reference band for no effect. A test for no effect of the parameters (introduced in section 4.4.4) has been implemented and p values for the test applied to the range, the sill and the nugget are respectively $< 0.0001, 0.255, 0.0003$ (Pseudo Likelihood Ratio test). For the sill it is therefore possible to say that there has not been a significant change across years. The results for the range and the nugget, have shown significant p values. However it is possible to see from Figure 6.6, that the trends are not monotonic and the effects are small (the values for the range and the nugget at

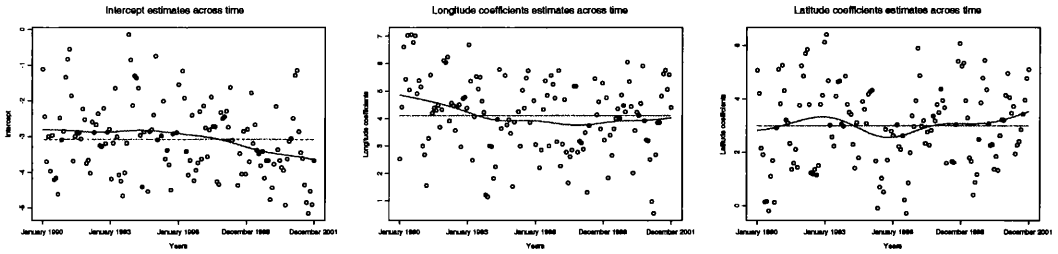

Figure 6.5: Temporal plots of β_{0j} β_{2j}, β_{1j} across t time points $j = 1, \ldots, t$. Displayed are smooth trend curves (continuous lines) and global averages (dashed lines) superimposed. Reference bands for no effect are also displayed.

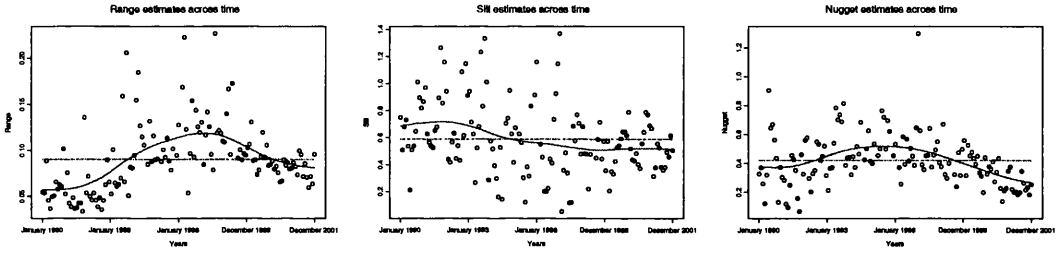

Figure 6.6: Time series of monthly estimates of Range (R_j), Sill (c_{0j}) and Nugget (c_{1j}) parameters across t time points $j = 1, \ldots, t$. Displayed are smooth trend curves (continuous lines) and global averages (dashed lines) superimposed. Reference bands for no effect are also displayed.

1990 are very close to the ones in 2001). Therefore we have proceeded by defining a spatial correlation structure for the entire period 1990-2001 by averaging the monthly estimates of the range, sill, and nugget effect parameters. Monthly averages of the variograms' parameters across all the time points are displayed by dashed lines in Figure 6.6. Using these values in the theoretical variogram defined in section 6.3.2 would result, assuming a separable model, in a matrix which could be used together with a time correlation matrix across space, to

undertake spatiotemporal analysis.

6.3.4 Spatial analysis across time: Conclusions

This section presents an approach to the spatial analysis of sulphur dioxide from 1990 to 2001 at 130 sites across Europe. At each time point, a plane seems to be the most appropriate model for the SO_2 trend surface. Analysis of the resulting residuals showed that the Gaussian variogram model fits better than the Exponential and the Spherical models. Time series analysis of the range, the nugget, and the sill showed no evidence that the spatial correlation changes markedly over time. Therefore the average of the estimates for each parameter could be used to define a spatial correlation matrix for the period of analysis. Thus a separable space-time covariance structure for SO_2 on this spatial scale and over this time period is appropriate.

However in order to build a separable model, there is a need to define a time correlation matrix that could be used in combination with the space correlation matrix to model the spatiotemporal trend. Therefore, the following sections will present some time series analyses across space, in order to define a time correlation matrix of the monthly SO_2 data from 1990 to 2001, for the 130 sites.

6.4 Time series analysis across space

Using the same data set analyzed in Section 6.3, in this section time series analysis across space is performed. The techniques used to perform these analyses are

identical to those presented in Chapter 4.

$$\ln(SO_2)_u = \alpha_u + m_u(years, months) + \varepsilon_u \quad u = 1, \ldots, 130 \tag{6.13}$$

At each of the 130 sites, (Model 6.13) has first been fitted assuming independent errors. An AR(1) model has then been fitted to the residuals. Visual inspection across the sites showed that the AR(1) assumption is not unrealistic. The top graph of Figure 6.7 shows the fit of Model (6.13) at an Italian site (IT04). The middle left and middle right graphs of Figure 6.7 show the *acf* and *pacf* of the residuals from the fitting of Model (6.13) at IT04. The bottom left and right graph of Figure 6.7 show respectively a contour plot and a boxplot of the estimated correlation coefficients across all sites. Both contour plots and boxplot indicate some variability in the correlation coefficient, but the average of the 130 estimated correlation coefficients ($\frac{\sum_{u=1}^{130} \hat{\rho}_u}{130} = 0.23$) is proposed to provide a single estimate of the temporal correlation across all sites. Consequently the element $[i, j]$ of the time correlation matrix across all sites will be $\hat{\rho}^{|i-j|}$.

6.5 Spatiotemporal additive model

In this section spatiotemporal additive models will be fitted and tested, accounting for the spatiotemporal correlation, obtained by combining the spatial correlation matrix across time (Section 6.3) with the time correlation matrix across space (Section 6.4).

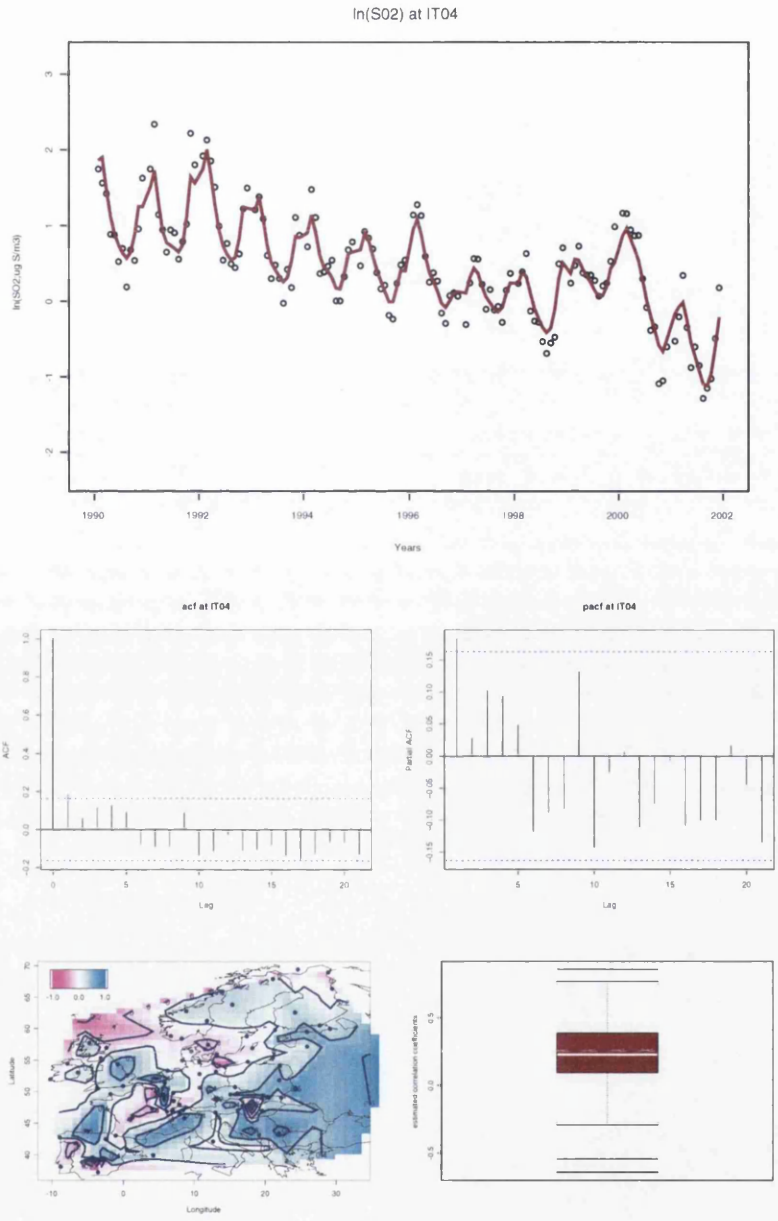

Figure 6.7: Fitted Model 6.13 to an Italian site (IT04); *acf* and *pacf* of the residuals. Contour plot and boxplot showing the distribution of the time correlation estimates across all sites.

6.5.1 Spatiotemporal additive model: introduction

Sections 6.3 and 6.4 showed "marginal spatial analysis" and "marginal time series analysis" respectively. This section proposes a spatiotemporal model that models the spatial and the temporal trends using at the same time all the data that are available, across space and time simultaneously.

Additive Models 6.14 and 6.15 are the additive models that we want to fit here, accounting for circular smoothers for the variable "months", and for the spatiotemporal correlation.

$$y = \alpha + m_{1,2}(years, months) + m_{3,4}(latitude, longitude) + \varepsilon \quad (6.14)$$

$$y = \alpha + m_1(years) + m_2(months) + m_{3,4}(latitude, longitude) + \varepsilon \quad (6.15)$$

For an additive model of the form 6.15 or 6.14, it is necessary to think about its computational applicability. For fitting a model with such characteristics it is necessary to build a smoothing matrix of dimension $n \times n$, where n is the sample size. If we think about monthly means of 10 years data across 100 sites, it would be necessary to build and to do computations with smoothing matrices of dimension 12000×12000. Computationally, it is understandable that this is extremely expensive, and because of that it is necessary to use an algorithm to avoid this dimensional problem. The next section will develop an approach using binning in the fitting of additive models.

6.5.2 Spatiotemporal additive model: binning large data sets

In nonparametric statistics, binning represents one of the most common tools to apply to computationally expensive procedures (Bowman and Azzalini, 2003). The concept of binning consists of reducing the raw data to frequencies over a fine grid. If the number of grid points is set to be large then the accuracy of the simplest form of binning can be maintained at a high level without introducing complications into the computational formulae. When a fine grid is placed over the sample space, the original data can be recoded as frequencies at grid locations and, for regression data, sample means and standard deviations of the response variable. Bins containing no observations should be omitted from the recorded list. Using the notation:

- b the number of bins,
- y_{ij} refers to observation j which lies in bin i,
- $\bar{x}_i$ the location of the bin i,
- $\bar{y}_i$ and s_i denote the mean and the standard deviation of the responses for the data in bin i,
- n_i denotes the bin frequencies.

it is possible to rewrite the least squares problem for fitting a local linear regression, minimizing, at each evaluation point z, the following expression:

$$\sum_{ij}\{y_{ij} - \alpha - \beta(\bar{x}_i - z)\}^2 w(\bar{x}_i - z; h) \tag{6.16}$$

Expression 6.16 can be expanded also in the following formula:

$$\sum_i n_i s_i^2 w(\bar{x}_i - z; h) + \sum_i n_i \{\bar{y}_i - \alpha - \beta(\bar{x}_i - z)\}^2 w(\bar{x}_i - z; h) \qquad (6.17)$$

which shows that the estimation of α, and hence the construction of the non-parametric regression curve, can be based solely on the bin means, locations and frequencies.

Denote the local mean estimator as $\hat{y} = Hy$, where H denotes the smoothing matrix $(n \times n)$. With binning the dimensionality of the smoothing matrices is determined by b rather than by n, which is a very considerable reduction. In fact the local mean estimator can be written as $\hat{y} = BSDy$, where:

- S is the smoothing matrix of dimension $b \times b$ defined by $\bar{x}_i$;
- D is the matrix that reduces the response variable y to the binned data $\bar{y}$ ($\bar{y} = Dy$), reducing the dimensionality from n to b. D is of dimension $b \times n$, whose elements of row j are all zeros, except those (j, i) elements at position i of y that belong to bin j that are defined by $1/n_j$, where n_j is the frequency of bin j.
- B is the matrix that expands the values of $\bar{y}$ back to a vector corresponding to the original data y, assigning the value $\bar{y}_i$ to the position j, in the vector y_{ij}. B is of dimension $n \times b$ whose elements of column j are all zeros, except those (i, j) elements at position i of y that belong to bin j, that are defined by 1.

6.5.3 Spatiotemporal additive model: binning additive models

The fitting of additive models has been carried out in Section 4.3 by defining and computing a smoothing matrix $H^{(i)}$ at iteration (i), such that $\hat{y} = H^{(i)}y$. The backfitting algorithm described in Section 4.3 starts by computing the smoothing matrices of each component. For the binned data, the smoothing matrix of the additive model for component j will be defined as $H_j = B_j S_j D_j$, where these matrices have been defined in Section 6.5.2.

In this section, a computationally less expensive version of the backfitting algorithm is proposed. This consists in updating the matrices B_j, S_j, D_j separately at each step rather than H_j, so dealing with matrices of lower dimensions. Indicate by N, a matrix of dimension $n \times n$ whose elements are $\frac{1}{n}$.

Using a similar approach to the one described in Section 4.3, for the simplest case of two variables, it is possible to write the first two steps of the backfitting algorithm in matrix form as follows:

1. first step:

$$\begin{aligned} \hat{m}_1^{(1)} &= (I - N)B_1 S_1 D_1 y \\ \hat{m}_2^{(1)} &= (I - N)B_2 S_2 [D_2 - D_2(I - N)B_1 S_1 D_1] y \\ &= (I - N)B_2 S_2 [D_2 - Q_2^{(1)}] y \end{aligned}$$

2. second step:

$$\hat{m}_1^{(2)} = (I - N)B_1 S_1 [D_1 - D_1(I - N)B_2 S_2 (D_2 - Q_2^{(1)})] y$$

$$
\begin{aligned}
&= (I - N)B_1S_1[D_1 - Q_1^{(2)}]y \\
\hat{m}_2^{(2)} &= (I - N)B_2S_2[D_2 - D_2(I - N)B_1S_1(D_1 - Q_1^{(2)})]y \\
&= (I - N)B_2S_2[D_2 - Q_2^{(2)}]y
\end{aligned}
$$

Using induction, it is possible to derive the projection matrix for variable j, updated at iteration (i), from the following formula:

$$H_j^{(i)} = (I - N)B_jS_j(D_j - Q_j^{(i)})$$

where,initializing $Q_j^{(1)} = I$, at iteration i it is possible to write:

$$Q_j^{(i)} = \sum_{k>j} P_{jk}(D_k - Q_k^{(i-1)}) + \sum_{k<j} P_{jk}(D_k - Q_k^{(i)})$$

where

$$P_{jk} = D_j(I - N)B_kS_k$$

It is worth noting that the previous expression can still be reduced in dimensionality. For example by expressing

$$P_{jk} = D_jB_kS_k - D_jNB_kS_k$$

it is possible to see that the first term of the difference is no longer an $n \times n$ matrix, and the second term can be simplified further as

$$D_jNB_kS_k = C_{jk}S_k$$

where C_{jk} is a matrix of dimension $b_j \times b_k$, where b_j and b_k are the numbers of bins present respectively in variables x_j and x_k. The C_{jk} matrix for each of the j rows has b_k elements: $n_1/n, n_2/n, \ldots, n_{b_k}/n$, where n_i denotes the frequencies of bin i.

In a similar way

$$H_j^{(i)} = (I - N)B_j S_j (D_j - Q_j^{(i)})$$

can be expressed as

$$H_j^{(i)} = (B_j S_j - N B_j S_j)(D_j - Q_j^{(i)})$$

where it is possible to note that the first term of the first difference is no longer an $n \times n$ matrix, and the second term can be further simplified by expressing it as:

$$N B_j S_j = C_j S_j \tag{6.18}$$

where C_j is a matrix of dimension $n \times b_j$, where for each of the n rows, there are b_j elements: $n_1/n, n_2/n, \ldots, n_{b_j}/n$.

Once the algorithm has converged at iteration i, it is simply necessary to sum together the estimates of each component $\hat{m}_j^{(i)} = H_j^{(i)} y, j = 1, \ldots, p$, to obtain an estimate of y. It is worth remembering that for the sum of the projection matrices of each component it is necessary to include the mean of y, in order to get a final H matrix that, when multiplied by y, gives an estimate $\hat{y}$. In fact the

projection matrices were computed assuming that:

$$\hat{y} = \bar{y} + (H_1 + H_2 + \ldots + H_p)y = (N + H_1 + H_2 + \ldots + H_p)y = Hy \quad (6.19)$$

Having obtained a hat matrix (H) for the additive model, it is now possible to compute residual sums of squares and degrees of freedom, in order to compute the approximate F test.

6.5.4 Spatiotemporal additive model: binning additive models with correlated errors

For the correlated errors case, the binned version of the backfitting algorithm needs to be extended. It is firstly necessary to define the correlation matrix Σ for the response variable y that accounts for both spatial and temporal correlation. Assuming a separable model, it is possible to obtain the full correlation matrix Σ as the direct product (known also as the Kroneker product) of the temporal correlation matrix Θ with the spatial correlation matrix Γ, namely $\Sigma = \Theta \otimes \Gamma$.

As expressed in Chapter 4, the computation of the smoothing matrices needs to be amended for the correlation of the errors, as does the definition of degrees of freedom and residual sum of squares of the model. The correlation matrices needed for estimating the local linear smoothers can be easily obtained using the matrix D defined in Section 6.5.2. Therefore the correlation matrix of Dy is $D\Sigma D^T$.

The computation of degrees of freedom and residual sum of squares are also different in the correlated case. Using the definition of degrees of freedom of the error ($df_{err.c.}$) given in section 4.4.2, using the present notation, it is possible to

write the following expression:

$$df = tr(\Sigma^{-1}H\Sigma + H^T - H^T\Sigma^{-1}H\Sigma) \tag{6.20}$$

In order to reformulate the previous expression to a less expensive one, the final smoothing matrix H can be decomposed as the sum of smoothing matrices $H = N + H_1 + H_2 + \ldots + H_p$.

Since the matrices involved in (6.20) are of dimension $n \times n$, making their computation extremely expensive, it is necessary to decompose each of the elements of the sum in (6.20) as follows:

$$\begin{aligned} tr(\Sigma^{-1}H\Sigma) &= \sum_{j=1}^{p} tr(\Sigma^{-1}H_j\Sigma) = \sum_{j=0}^{p} tr(\Sigma\Sigma^{-1}H_j) = \\ &= \sum_{j=0}^{p} tr[\Sigma\Sigma^{-1}(I-N)B_jS_j(D_j - Q_j^{(i)})] \end{aligned} \tag{6.21}$$

$$tr(H^T) = \sum_{j=0}^{p} tr(H_j^T) = \sum_{j=0}^{p} tr[(D_j - Q_j^{(i)})^T S_j^T B_j^T (I-N)^T] \tag{6.22}$$

$$\begin{aligned} tr(H^T\Sigma^{-1}H\Sigma) &= tr[(\sum_{j=0}^{p} H_j^T\Sigma^{-1})(\sum_{j=0}^{p} H_j\Sigma)] = tr[\sum_{k=0}^{p}\sum_{j=0}^{p} H_k^T\Sigma^{-1}H_j\Sigma] = \\ &= tr[\sum_{k=0}^{p}\sum_{j=0}^{p}(D_k - Q_k^{(i)})^T S_k^T B_k^T (I-N)^T\Sigma^{-1} \\ &\quad (I-N)B_jS_j(D_j - Q_j^{(i)})\Sigma] \end{aligned} \tag{6.23}$$

where H_0 corresponds to the N matrix. From expressions 6.21, 6.22 and 6.23, it is possible to obtain more efficient computations, calculating the matrix product

in blocks to avoid the computation of full $n \times n$ matrices.

For the computation of the Residual Sums of Squares (RSS), the calculation of $(y - \hat{y})\Sigma^{-1}(y - \hat{y})$, can be written in a simpler computational formula as $(y - \hat{y})\Theta^{-1} \otimes \Gamma^{-1}(y - \hat{y})$.

6.5.5 Spatiotemporal additive model: application

On the basis of the methodology presented above, the additive models that have been fitted here are:

$$\ln(SO_2) = \alpha + m_1(year) + m_2(month) + m_3(lon., lat.) + \varepsilon \quad (6.24)$$

$$\ln(SO_2) = \alpha + m_{12}(year, month) + m_3(lon., lat.) + \varepsilon \quad (6.25)$$

Both models have been fitted to the monthly means of SO_2 from 1990 to 2001 at 130 sites across Europe. The temporal correlation matrix (Θ as expressed in Section 6.5.4) has been computed by the procedure shown in Section 6.4. The spatial correlation matrix (Γ as expressed in Section 6.5.4) has been computed by the procedure shown in Section 6.3. The fits of both models are shown in Figures 6.8 and 6.9.

The bottom panel of Figure 6.8 and the right panel of Figure 6.9 describe the spatial pattern of $\ln(SO_2)$ concentrations of Model 6.24 and Model 6.25 respectively. Both show higher values in Eastern and Central Europe.

It is possible to note that Model 6.24 fits the trend and the seasonality as two univariate components, assuming then that the seasonal pattern (top right hand panel in Figure 6.8) is constant over the years, or that the trend component (top left hand panel in Figure 6.8) is constant in each month. Model 6.25 instead

presents trend and seasonality as a bivariate term and therefore the seasonal cycle is allowed to change over time. Indeed from the left panel of Figure 6.9 it is possible to analyze the three-dimensional nature of the time component, where the x-axis is the trend, the y-axis is the seasonal component and the z-axis is the $\ln(SO_2)$ concentration. It can be seen that the 12 lines along the x-axis show a decreasing trend across all the months, while the 12 lines along the the y-axis describe the seasonal pattern per each year.

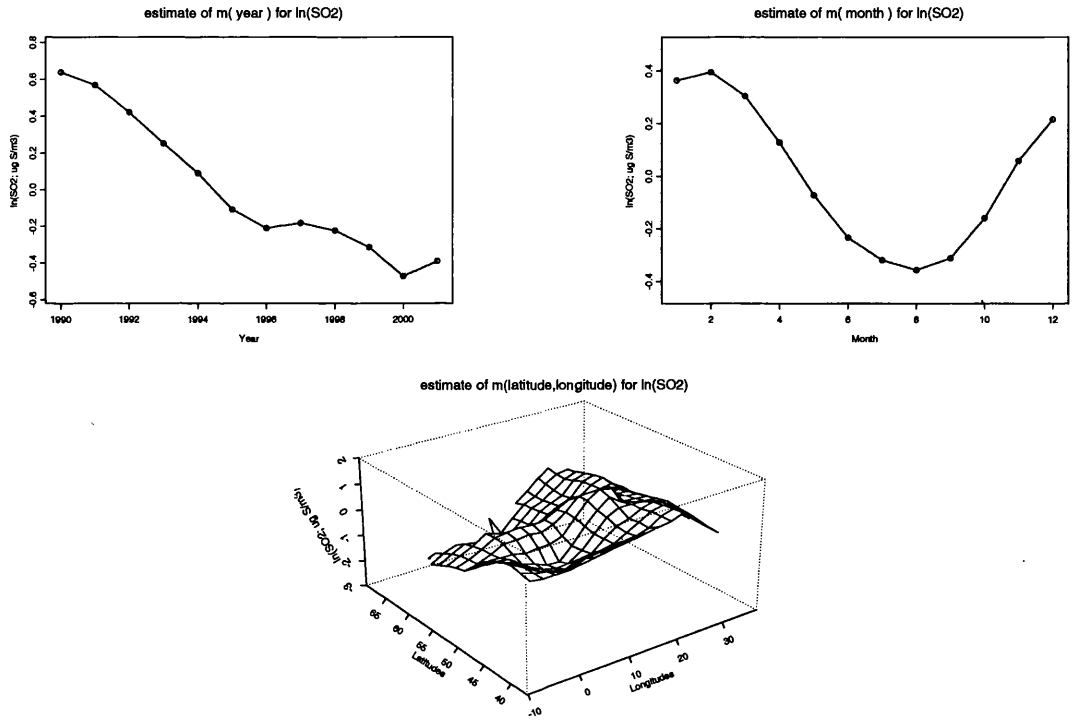

Figure 6.8: Fits of the components $m(year)$, $m(month)$, and $m(latitude, longitude)$ for model 6.24.

These models have been tested, applying the approximate F test, in order to check for changes in seasonality. The resulting p value is 0.011. This significant p value means that Model 6.25 is a better fit compared to Model 6.24. In other words it means that there is evidence of a change in seasonality across years over

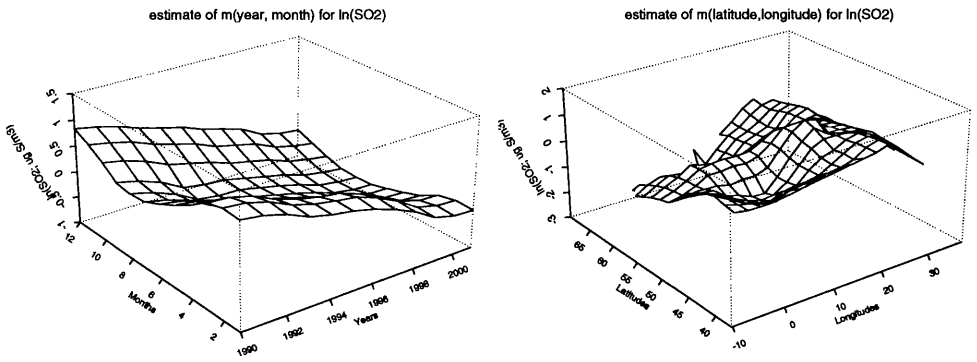

Figure 6.9: Fits of the components $m(year, month)$, and $m(latitude, longitude)$ for model 6.25.

Europe. Following the analysis of section 6.3, two semiparametric models have been fitted:

$$\ln(SO_2) = \alpha + m_1(year) + m_2(month) + \beta_1 lat. + \beta_2 lon. + \varepsilon \quad (6.26)$$

$$\ln(SO_2) = \alpha + m_{12}(year, month) + \beta_1 lat. + \beta_2 lon. + \varepsilon \quad (6.27)$$

In fact, from the analysis of section 6.3, the plane seemed to be the most reasonable model for the spatial trend. However, those conclusions were made just from a visual inspection of the spatial trends at each time point separately. Therefore, from fitting model 6.26 and model 6.27, and testing them versus model 6.24 and model 6.25, it will be possible to analyze if the spatial trend can be adequately described by a plane. From the tests that compare the four models, it appears that model 6.25 is the best. In other words, both nonparametric components, temporal and spatial, are needed to describe their pattern. For the temporal component, years and months need to be modeled by a bivariate term, rather than two univariate ones.

It is necessary to clarify, that the choice of a linear plane in Section 6.3 in order to estimate correlation, does not contradict the choice of the smooth surface that is selected here to describe the trend. Indeed, the linear plane is supposed to remove just the main part of the trend, leaving the residuals that will be used to quantify the correlation. Once the correlation is computed and accounted in the model, then the trend surface to be selected will be the one that better fits the data.

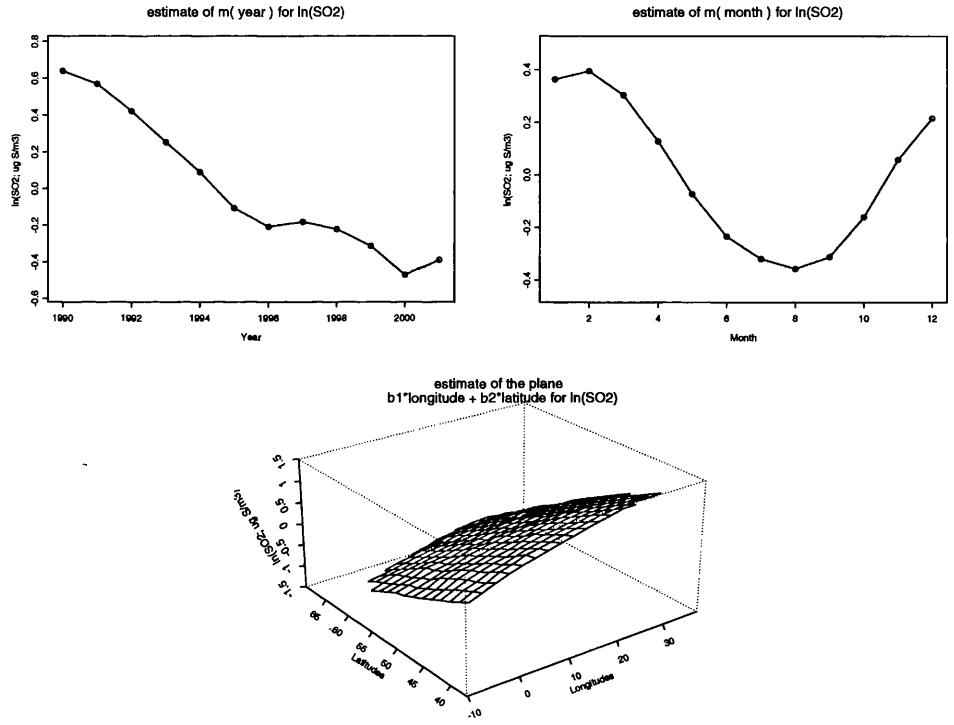

Figure 6.10: Fits of the components $m(year)$, $m(month)$, and $\beta_1 * Latitude + \beta_2 * Longitude$ for model 6.26.

In order to have a better understanding of how the seasonality changed, the 12 seasonal cycles from 1990 to 2001 are displayed in Figure 6.12. It is possible to note that the seasonal cycle keeps the same shape in terms of location of

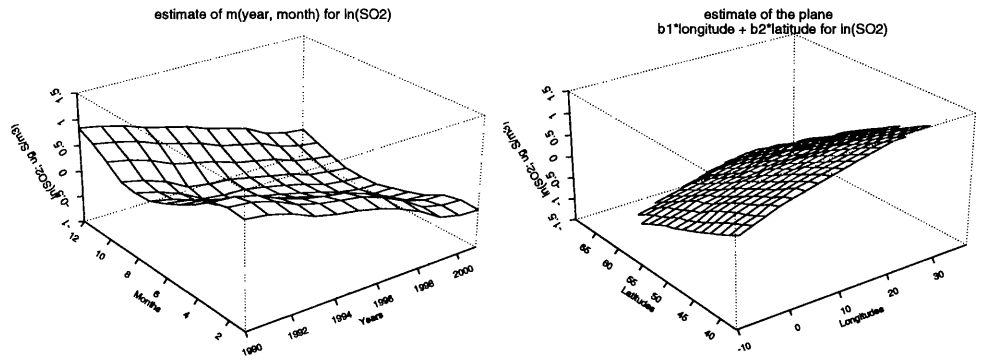

Figure 6.11: Fits of the components $m(year, month)$, and $\beta_1 * Latitude + \beta_2 * Longitude$ for model 6.27.

Table 6.1: p-values from testing Models: 6.24, 6.25, 6.26 and 6.27

Models	6.24	6.27
6.25	$4.8e^{-10}$	$1e^{-20}$
6.26	$1e^{-20}$	$1.3e^{-9}$

peaks (winter months) and troughs (summer months). However the winter values increase and the summer ones reduce from 1990 up to 1993, stressing more the difference between peaks and troughs. From 1994 to 2001, the winter values reduce and the summer ones increase, giving a smaller difference between peaks and troughs, and indicating that further movement in this direction, might cause the seasonal signal to disappear.

6.5.6 Spatiotemporal additive model: conclusion

This section presents a spatiotemporal analysis of $\ln(SO_2)$ across Europe from 1990 to 2001. Fitting and testing techniques for additive models that can deal with large data sets and that account for correlation have been illustrated. The

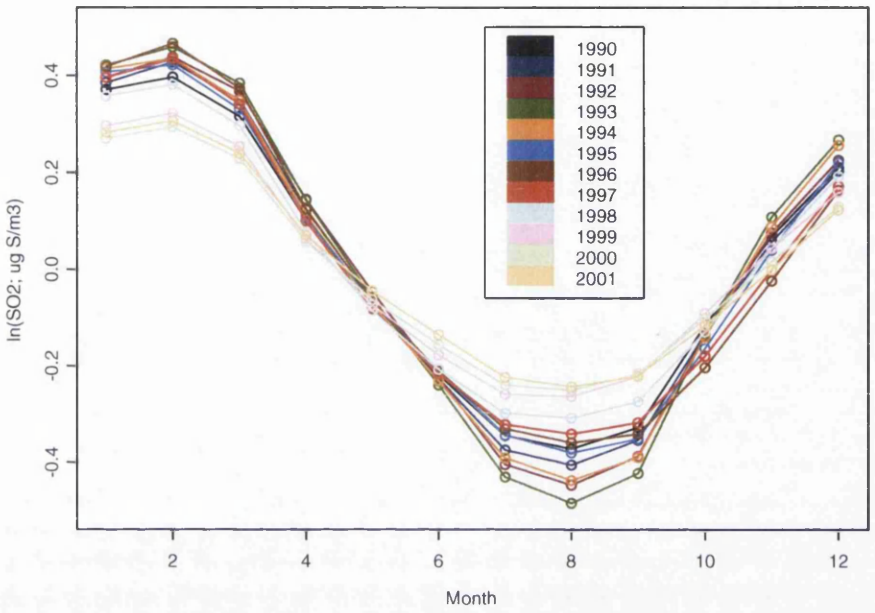

Figure 6.12: Seasonal cycles from 1990 to 2001.

results show a decreasing time trend across Europe and a seasonal cycle that changes significantly across time over Europe. The seasonal signal shows higher values in winter and lower ones in summer. However in the later years this seasonal pattern became flatter. A possible explanation for this could be due to the increase in the use of air conditioning, leading to higher emissions during summer months. The spatial pattern seems to indicate higher concentrations in the centre and in the east of Europe and lower values in the north west.

6.6 Spatiotemporal Analysis: Conclusions

Two marginal analyses, the spatial trend across time and time trend across space, have been performed, from which spatial correlation across time and time correlation across space have been obtained respectively. The spatial correlation does not seem to change significantly over time, and the time correlations seem quite homogeneous over Europe.

Fitting and testing techniques for additive models that can deal with large data sets and that account for correlation have been shown. Analysis of SO_2 data monitored at 130 sites from 1990 to 2001 showed higher concentrations in the center and in the east of Europe and lower values in the north west. A decreasing time trend across Europe and a tendency of the seasonal cycle to disappear over time have also been noted.

Chapter 7

Analysis Of Emissions' Effects

7.1 Introduction

In the previous chapters, we analyzed some of the main characteristics of the spatiotemporal pattern of sulphur dioxide concentrations monitored across Europe in the last 30 years. As stated at the very beginning of this work (chapter 1) the main purpose was to see if the efforts that have been made during the last quarter of the twentieth century by European countries to reduce emissions have resulted in a real improvement in environmental quality and in a real change in the acidifying environment. In this chapter we present an analysis of the relationship between sulphur dioxide concentrations monitored at 112 sites across Europe from 1990 to 2001, and the emissions data that European countries publish every year.

Pollutant concentrations have been the subject of a great deal of study. However few of them have looked at the relationship with emissions. Hůnová et al.

(2004) presented the observed trends of some pollutants in air and in precipitation at rural sites in the Czech Republic from 1993 to 2001. A statistically significant decreasing trend in SO_2 was explained by the political and economic change in the Czech Republic and neighbouring countries in the 1990s, and by the adoption of new technologies. They also pointed out the non-linearity in the response of sulphur to emission reductions.

Barbieri et al. (2004) analyzed a series of wet deposition samples from 1982 to 1998 in the southern region of the Central Alps. Relationships between depositions and distance from the emission sources were explored and quantified using principal component analysis and linear models. Results showed the existence of an ionic concentration gradient along a south-north axis and with altitude.

Vuorenmaa (2004) analyzed the long-term changes of acidifying deposition in Finland from 1973 to 2000. In order to determine deposition trends with respect to implementation of international emission reduction agreements (CLRTAP), annual means were divided into two time periods 1973-1985 and 1986-2000. These time periods represent periods prior to and after sulphur emission reduction abatements. The Kendall-τ test was applied to examine the significance of the trends, and a simple linear regression model was used for slope estimates. For the period 1973-1985, no significant changes were observed for sulphate deposition, while increasing trends were observed for deposition of nitrogen compounds. For the second period 1986-2000, substantial decreases (30% in northern and 60% in southern Finland) were observed for sulphate deposition. Nitrogen deposition also decreased but less than sulphate deposition.

Sirois (1997) presented a study of the temporal variation of the oxides of

sulphur and nitrogen at 8 sites in eastern Canada. Using kernel smoothing regression and spectral analysis, Sirois noted a difference between the long-term trends of SO_2 and SO_4, and his explanation was based on the fact that SO_2 is influenced mainly by local sources (Canadian sources) and SO_4 by more distant sources (USA sources).

Berge et al. (1999) analyzed the temporal trend of the atmospheric emissions and depositions of sulphur and nitrogen compounds in Europe using EMEP data. Source-receptor matrices, which quantify the transboundary transport between the European countries, were presented. They proposed the use of source-receptor matrices, as a convenient way of presenting the budgets of the transboundary fluxes between European countries. In such a matrix the amount of airborne transport of acidifying sulphur and nitrogen from one country to any other country is established by the use of a deterministic model. The model used by Berge et al. (1999) is the one developed by EMEP, that is a two dimensional Lagrangian trajectory model utilized to calculate the transboundary fluxes and depositions of acidifying compounds in Europe from 1985 to 1995. Berge et al. (1999) found that the total deposition to all grid cells in the EMEP domain, from 1985 to 1995, reduced by 34%, 9% and 12% for sulphur, oxidized and reduced nitrogen respectively. For the same period, the deposition reductions were 5%, 1% and 6% smaller than the emission reductions for sulphur, oxidized and reduced nitrogen respectively. Berge et al. (1999) pointed out that, apart from reduced nitrogen which has a shorter lifetime, for sulphur and oxidized nitrogen, many countries in Europe receive most of the acidifying compounds from emissions in other countries. Berge et al. (1999) also commented that the seasonal variability in the meteorological conditions affects the annual deposition.

In the USA, studies conducted both by Holland et al. (1999) and by Lynch et al. (2000) confirmed a strong correspondence between the 35% reduction in reported emission and the 32% reduction in sulphur dioxide concentrations calculated with a regression technique. The difference in reductions become null if the data from only the last three years are considered, despite the fact that they are obviously affected by short time scale deviations in meteorology. Holland et al. (1999) examined both trends at individual sites and those for aggregated regions, finding it more problematic to assign regional descriptors to complex terrain. Problems of both meteorological variability influencing short records and of difficulty in defining regions was taken up by Blanchard et al. (1996). They noted that in areas subject to higher levels of deposition ($> 20kg$ S 1/ha 1/yr) the power to detect trends could be expected to reach 90% within around two years, while monitoring regions with levels of deposition below this may take twice as long to identify a trend. However the time taken to quantify that trend is likely to take an additional 6-7 years even in the higher deposition zone. Their study also revealed that even if identification of a trend may be possible with a limited monitoring network, defining isopleths for given deposition criteria then becomes quite uncertain. An important study in the United Kingdom was conducted by Downing et al. (1995) who compared the dry and the wet deposition of sulphur between 1978 and 1993. They constructed maps for wet deposition for 1978-1980 to compare with equivalent maps for 1989-1993 for the whole of mainland Britain. Wet deposition of sulphur for the UK as a whole declined by 43% while UK emissions fell by 32%. During the same period decline in SO_2 concentrations and dry deposition in remote areas reached as much as 70%. This indicates that UK emissions alone could not account for the changes in British

air quality.

In contrast with the approaches described in this brief literature review, this chapter will propose a statistical analysis of the emission effects on the observed concentrations using nonparametric inferential tools. On the basis of the EMEP data sets, sulphur dioxide (SO_2) concentrations, monitored in Europe from 1990 to 2001, are related to "own" and neighbouring countries emissions, through additive models accounting for the data correlations. The nature and the statistical significance of the relationships are examined.

The data analyzed are the monthly means of the natural logarithm of SO_2 concentrations monitored daily from 1990 up to 2001 at 112 sites across Europe by the EMEP network and the natural logarithm of the annual total emissions per each European country, obtained from the UNECE/EMEP emissions database, available at `http://webdab.emep.int`. Annual emissions have also been expressed monthly, dividing the annual values by 12.

For those 11 sites that monitored meteorological variables (Eskdalemuir GB02, Westerland DE01, Waldhof DE02, Schauinsland DE03, Deuselbach DE04, Brotjacklriegel DE05, Kosetice CZ03, Rörvik SE02, Bredkälen SE05, Hoburg SE08, Payerne CH02) wind data will be used to build a neighbouring countries emissions covariate. The models fitted in this chapter are mainly additive models whose fitting and testing techniques used will be those presented in Chapter 4.

7.2 Analyzing relationship between observed trends and emissions

This section looks at the relationship between observed SO_2 concentrations and reported emissions. The additive models that have been fitted in this analysis are the following ones:

$$\ln(SO_2) = \alpha + m_m(month) + m_y(year) + m_o(\ln(oe)) + \varepsilon \quad (7.1)$$

$$\ln(SO_2) = \alpha + m_m(month) + m_o(\ln(oe)) + \varepsilon \quad (7.2)$$

$$\ln(SO_2) = \alpha + m_m(month) + m_y(year) + \varepsilon \quad (7.3)$$

$$\ln(SO_2) = \alpha + m_m(month) + \varepsilon \quad (7.4)$$

$$\ln(SO_2) = \alpha + m_m(month) + \beta_1 \ln(oe) + \varepsilon \quad (7.5)$$

where *oe* stands for own country emissions.

For six countries, estimates $\hat{m}_y(year)$ in Model 7.3 have been plotted in Figure 7.1 for each site with the annual emissions of the country to which the site belongs. The estimated trends are plotted as thin dotted lines with confidence bands in light gray and the trend scale is plotted on the left axis. The emissions are plotted as triangles linked by thick continuous lines and the scale is plotted on the right axis. For the six countries displayed, the observed trends and the emissions show a similar decreasing pattern. However, since emissions and estimated trends are plotted on different scales, no comments on proportionality can be made, and further analysis of their relationship will be investigated later.

In order to analyze the significance of the "own emissions" component, consideration needs to be given to the relationship of the $m_y(year)$ and the $m_o(\ln(oe))$

components. Both terms describe the $\ln(SO_2)$ trend, as $\ln(oe)$ is itself a function of *year*. Therefore a model that includes both terms could be affected by concurvity (the nonparametric analogue of collinearity), causing problems for its fitting and testing. In order to compare the amount of variability of $\ln(SO_2)$ that each of the two terms explains, the R^2s of models 7.1 - 7.4 have been computed. A histogram of the R^2s of Model (7.2) has been plotted in Figure 7.2. The histogram shows a wide range of R^2 values across these sites.

For each site, differences in R^2s of Model 7.1 (x), Model 7.3 (o) and Model 7.4 (+) from the R^2s of Model 7.2 have been plotted in Figure 7.3. The continuous lines are the distances from R^2 of Model 7.3 (o) to R^2 of Model 7.2, while the dotted lines are the distances from R^2 of Model 7.4 (+) to R^2 of Model 7.2. From the length of the dotted lines, it is clear that the model with the seasonal term alone has a large increase in R^2 if either the emissions or the years component is included. It is also possible to note from the continuous lines, that the R^2s of Model 7.3 don't differ too much from the R^2s of Model 7.2 apart from a few sites where the differences in R^2 is about 0.2. Also a model that includes both *year* and $\ln(oe)$ leads to only a very small increase in R^2 compared to the models with just one of these two terms, *year* or $\ln(oe)$. It is therefore possible to conclude that a reasonable model that describes the SO_2 variability need contain only the emission term and the seasonal component.

To explore the form of the relationship between SO_2 emissions and SO_2 observations, Model 7.5 was fitted across all sites to assess if a linear relationship between log emissions and observed $\ln(SO_2)$ concentrations was adequate, rather than the nonparametric term for $\ln(SO_2)$ concentrations fitted in Model 7.2. The difference between those two models was investigated using the Pseudo Likelihood

Ratio test. The p values are shown in Figure 7.4, where triangles (at 66 sites) show nonsignificant p values, while circles (at 46 sites) show significant p values. This shows that at 46 sites it is not possible to model the $\ln(SO_2)$ concentrations by a linear function of the log emissions, while at 66 sites, log own country emissions are linearly related to $\ln(SO_2)$ concentrations. It can also be seen that by rewriting equation (7.5) as equation (7.6), the proportionality assumption between observed concentrations and emissions is valid only for $\beta_1 = 1$.

$$SO_2 = e^{\alpha + m_m(months)} o e^{\beta_1} e^{\varepsilon} \tag{7.6}$$

Figure 7.5 shows the estimated β_1 coefficients, and the respective confidence intervals for those 66 sites where it is possible to model the $\ln(SO_2)$ concentrations by a linear function of the log emissions. It is clear that the confidence intervals for the β_1 estimates generally do not include the value of 1 (where the dashed line is displayed). This means that strict proportionality between concentrations and own country emissions does not hold at the majority of sites. Most of the β_1 estimates are greater than one indicating that a decrease in emissions corresponds to a greater decrease in concentrations. Figure 7.6 shows a contour plot of the β_1 estimates at those 66 sites where Model 7.5 has not been rejected. It can be noted that the central-east and UK areas are characterized by low positive β_1 estimates, while the Alps and the Netherlands show higher β_1 estimates.

7.3 Analyzing neighbouring countries emissions

For each of the 11 sites that reported wind information, an analysis of the effects of the neighbouring countries emissions has been performed.

In order to define the neighbouring countries emissions covariates (ne), at each of the 11 stations where meteorological variables were reported ($k = 1, \ldots, 11$), for each of the 12 years ($t = 1990, \ldots, 2001$), weighted sums of the 33 log neighbouring countries emissions have been calculated using the following expression.

$$ne_{kt} = \sum_{j=1}^{33} w_d(d_{jk}) w_t(\theta_{jk}) \ln(E_{jt}), \quad k = 1, \ldots, 11, \quad t = 1990, \ldots, 2001 \qquad (7.7)$$

Here E_{jt} are the emissions of country j in year t, $w_d(d_{jk})$ are weights based on the distance between site k and country j, $w_t(\theta_{jk})$ are the weights based on wind direction and speed, from direction country j to site k in year t.

An appropriate weight function for the distances between each of the 11 monitoring sites with each of the 33 neighbouring countries is

$$w_d(d_{jk}) = e^{-\frac{1}{2}\left\{\frac{d_{jk}}{h_d}\right\}^2}, \quad j = 1, \ldots, 33, \quad k = 1, \ldots, 11, \qquad (7.8)$$

where d_{jk} is the distance between site k and the center of country j and h_d is a smoothing parameter which regulates the weights given to each distance. This means that the smaller the value of h_d, the smaller the weight given to the long distances, the bigger the value of h_d, the higher the weights given to long distances. To define the distance between each of the 11 sites and each of 33 neighbouring countries, "country-centers" have been defined subjectively by inspection of a map. In order to compute distances that account for the

curvature of a globe, the latitude (ω) and longitude (λ) location coordinates have been translated to the x and y coordinates of a tangent plane to the north pole using the Lambert (or Schmidt) projections (Fisher et al., 1993) introduced in Section 6.3.1.

The weight function for wind direction and wind speed was created using weighted annual averages of the weekly wind directions with the weekly wind speed as weights. For each year t, site k and country j, the weight function for wind is given by

$$w_t(\theta_{jk}) = \sum_{i=1}^{53} \nu_{ikt} e^{h_w \cos(\theta_{jk} - \phi_{ikt})}, \quad \begin{aligned} j &= 1, \ldots, 33, \quad k = 1, \ldots, 11 \\ t &= 1990, \ldots, 2001 \end{aligned} \tag{7.9}$$

where θ_{jk} is the angle between site k and country j, and ϕ_{ikt} and ν_{ikt} are the weekly values for wind direction and speed for each of the 11 sites. h_w is a smoothing parameter that regulates the weights given to each wind direction and speed. Larger values of h_w lead to larger weights associated with winds from directions close to the direction of site k from country j.

The weighted neighbouring countries emissions computed by expression (7.7) and averaged from 1990 to 2001 are plotted in Figure 7.7. This plot gives an indication of the effects of the neighbouring countries emissions on each of the 11 sites. At each site, the longer the spikes, the higher the effect of the emissions coming from that direction on the SO_2 concentrations.

Having defined a neighbouring countries emissions covariate, Model (7.10) has

been fitted at each of the 11 sites;

$$\ln(SO_2) = \alpha + m_m(month) + m_o(\ln(oe)) + m_n(ne) + \varepsilon \qquad (7.10)$$

where *ne* stands for neighbouring countries emissions. Model (7.10) has been compared with Model (7.2) using the quadratic form tests in order to analyze the effects of the emissions of neighbouring countries. Results are listed in Table 7.1. Neighbouring countries' emissions were significant at one German site (DE03), at the Czech, at the British, and at the Swedish sites. The rest of the German and the Swiss sites do not show any significant effect of the neighbouring countries' emissions.

One possible explanation of these results is that the Swedish and the British sites are mainly affected by the emissions coming from the European mainland, and the Czech site by the eastern European countries. The German site could be affected by neighbouring emissions because it is situated near the border and at the top of a mountain, at an altitude of over 1200 meters above sea level and therefore more exposed to air mass of SO_2 coming from neighbouring countries.

The R^2s of Model 7.10 (o), Model 7.2 ($\triangle$) and Model 7.4 (+) for each of the 11 sites have been plotted in Figure 7.8. The dashed lines show the differences in R^2s between Model 7.2 ($\triangle$) and Model 7.4 (+), and the dotted lines show the difference in R^2s between Model 7.2 ($\triangle$) and Model 7.10 (o). It is possible to note again large differences in R^2 when the own country emission term is included.

It is interesting to note that there is a big difference in R^2 of Model (7.10) (o) and Model (7.2) ($\triangle$) only at SE02 and SE05, while at SE08, CZ03, DE03, and GB02, where the neighbouring emissions were significant, the differences in R^2

Table 7.1: p values from testing neighbouring countries emissions.

p-values	Model (7.10) Versus Model (7.2)
GB02	0.020
DE01	0.431
DE02	0.324
DE03	0.023
DE04	0.315
DE05	0.228
CZ03	0.007
SE02	< 0.0001
SE05	< 0.0001
SE08	< 0.0001
CH02	0.081

are not large.

Table 7.2 presents a list of the final models selected at each site.

Table 7.2: Models selected at each site.

site	Models selected
DE01	$ln(SO_2) = \alpha + m_m(month) + m_o(\ln(oe)) + \varepsilon$
DE02	$ln(SO_2) = \alpha + m_m(month) + m_o(\ln(oe)) + \varepsilon$
DE03	$ln(SO_2) = \alpha + m_m(month) + m_o(\ln(oe)) + m_n(ne) + \varepsilon$
DE04	$ln(SO_2) = \alpha + m_m(month) + m_o(\ln(oe)) + \varepsilon$
DE05	$ln(SO_2) = \alpha + m_m(month) + m_o(\ln(oe)) + \varepsilon$
SE02	$ln(SO_2) = \alpha + m_m(month) + m_o(\ln(oe)) + m_n(ne) + \varepsilon$
SE05	$ln(SO_2) = \alpha + m_m(month) + m_o(\ln(oe)) + m_n(ne) + \varepsilon$
SE08	$ln(SO_2) = \alpha + m_m(month) + m_o(\ln(oe)) + m_n(ne) + \varepsilon$
CZ03	$ln(SO_2) = \alpha + m_m(month) + m_o(\ln(oe)) + m_n(ne) + \varepsilon$
GB02	$ln(SO_2) = \alpha + m_m(month) + m_o(\ln(oe)) + m_n(ne) + \varepsilon$
CH02	$ln(SO_2) = \alpha + m_m(month) + m_o(\ln(oe)) + \varepsilon$

7.4 Conclusions and discussions

This chapter proposes a statistical analysis of observed $\ln(SO_2)$ concentrations as a function of $\ln(SO_2)$ emissions in Europe from 1990 to 2001. Across most of the European sites, the observed SO_2 concentrations show a clear relation to the own country emissions and reported R^2 values show that own country emissions explain a substantial part of $\ln(SO_2)$ variability. However, proportionality between observed SO_2 concentrations and SO_2 emissions does not generally apply. Some sites show a nonlinear relationship between own country emissions and observed concentrations. For most of those sites where the relationship appears to be linear, the rate of decrease of the observed SO_2 concentrations is higher than the decrease of the own country SO_2 emissions, and in particular a much faster decrease of the observed SO_2 concentrations than own country emissions has been noted in the Alps and in the Netherlands areas.

The new neighbouring countries emissions covariate that has been defined has resulted in a statistically significant improvement in the model at more than half of the sites analyzed. In particular, one German site (DE03), the British, the Czech and the Swedish sites show statistically significant effects of the neighbouring emissions over the observed SO_2, while the Swiss and the other four German sites do not show any significant effects of the neighbouring emissions over the monitored SO_2 concentrations.

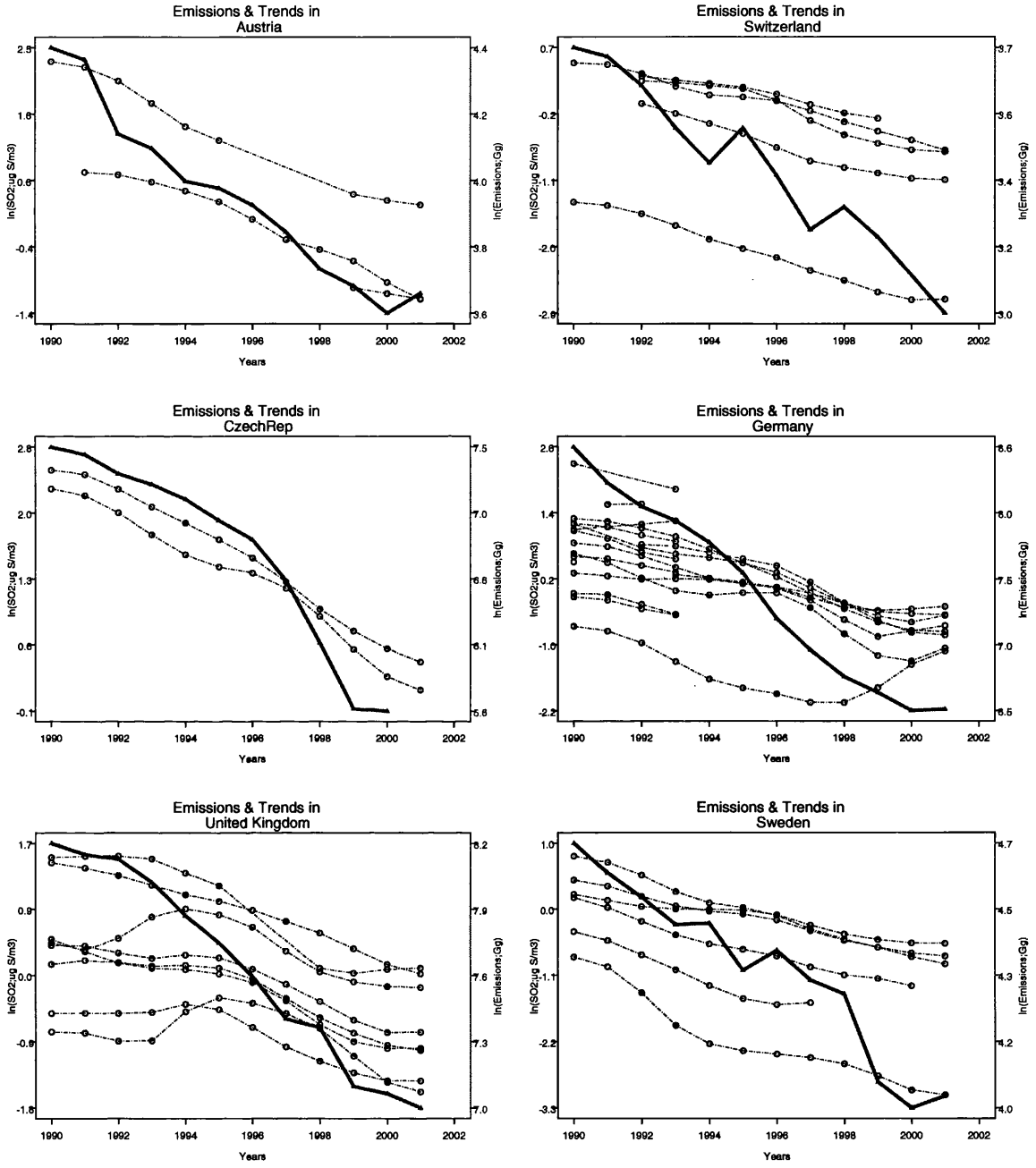

Figure 7.1: Annual emissions data (thick continuous line), and estimates $\hat{m}_y(year)$ of model (7.3) (thin dotted line) with standard errors bands (shaded regions).

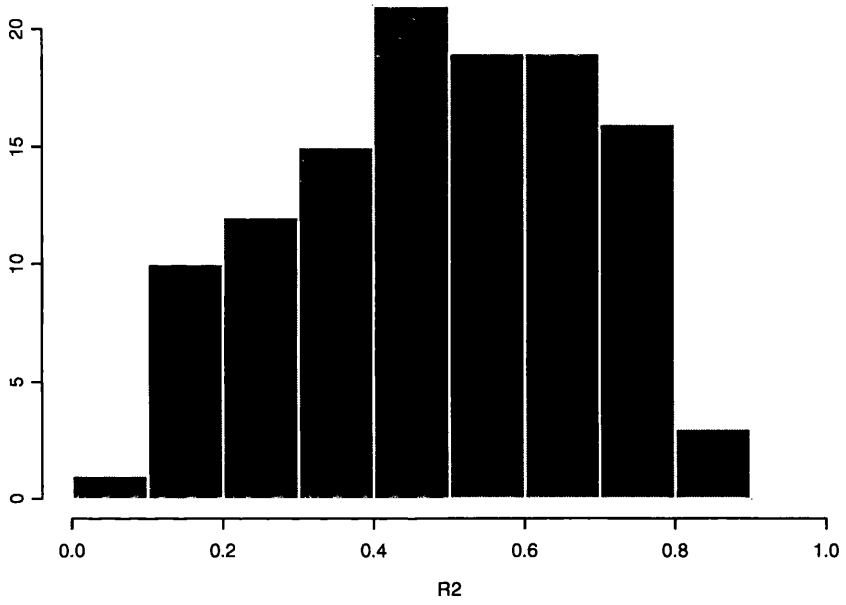

Figure 7.2: Histogram of R^2s from fitting Model (7.2) to each site.

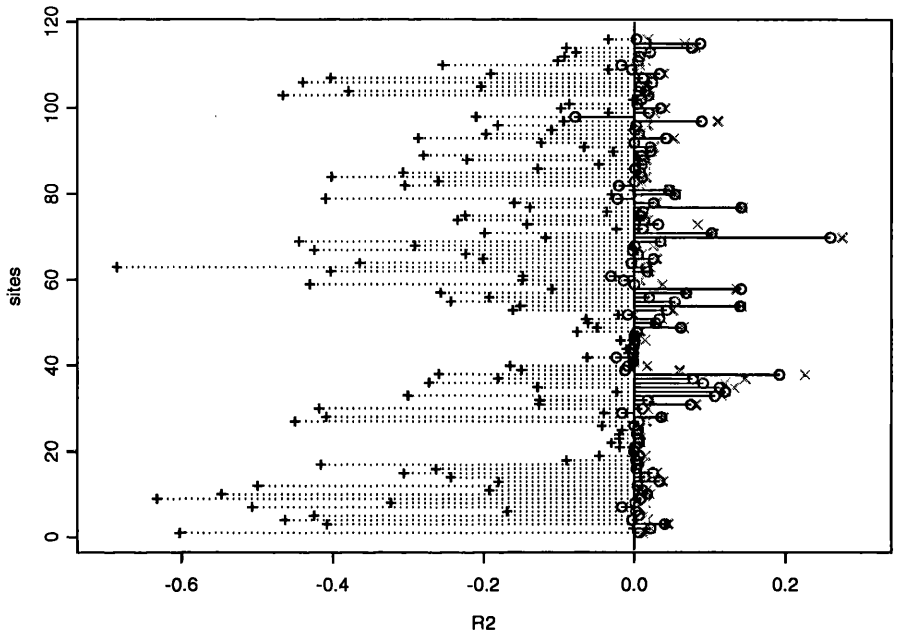

Figure 7.3: Differences of R^2s of Model (7.1) (x), Model (7.3) (o) and Model (7.4) (+) from R^2s of Model (7.2). The continuous lines are the distances from R^2 of Model (7.3) (o) to R^2 of Model (7.2), while the dotted lines are the distances from R^2 of Model (7.4) (+) to R^2 of Model (7.2).

Figure 7.4: Testing Model (7.5) versus Model (7.2) using the quadratic form test. Circles mean statistically significant non linear effect of log own emissions over the $\ln(SO_2)$ concentrations ($p < 0.05$). Triangles mean not statistically significant non linear effect of log own emissions over the $\ln(SO_2)$ concentrations ($p > 0.05$).

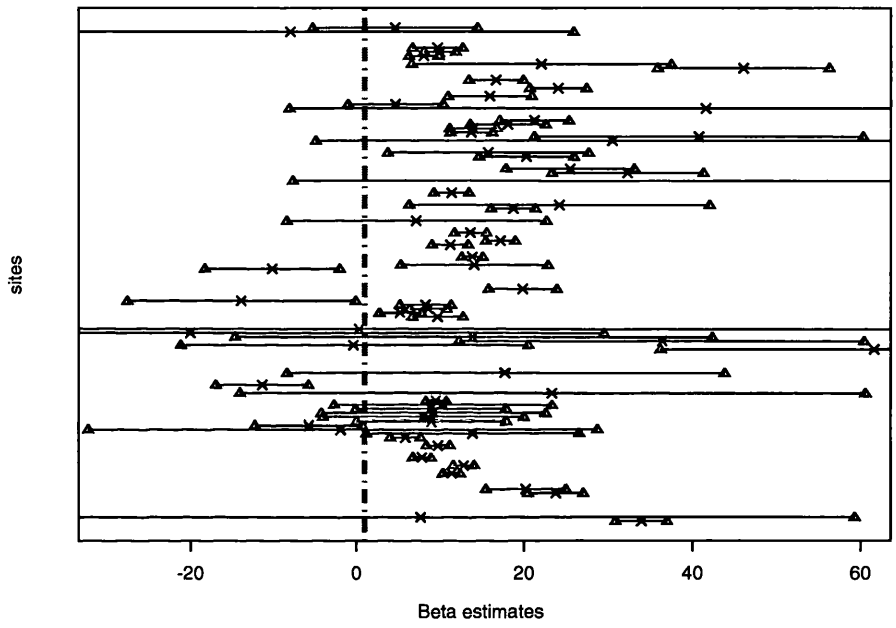

Figure 7.5: Plot of the β_1 estimates (x), and confidence intervals (continuous lines limited by $\triangle$) of Model (7.5), obtained at those 66 sites where Model (7.5) has been accepted. A dotted & dashed line is drawn at $\beta_1 = 1$.

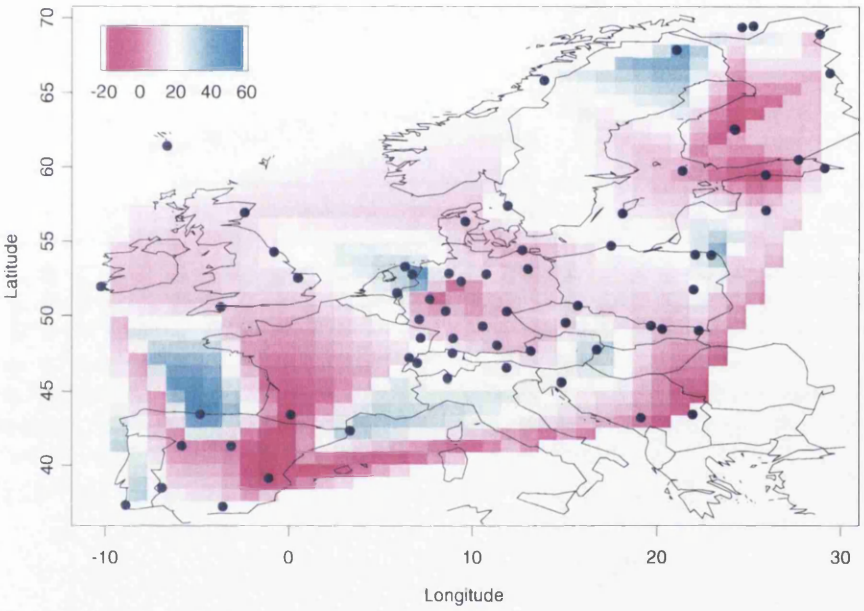

Figure 7.6: Contour plot of the β_1 estimates at those 66 sites where Model (7.5) has been accepted.

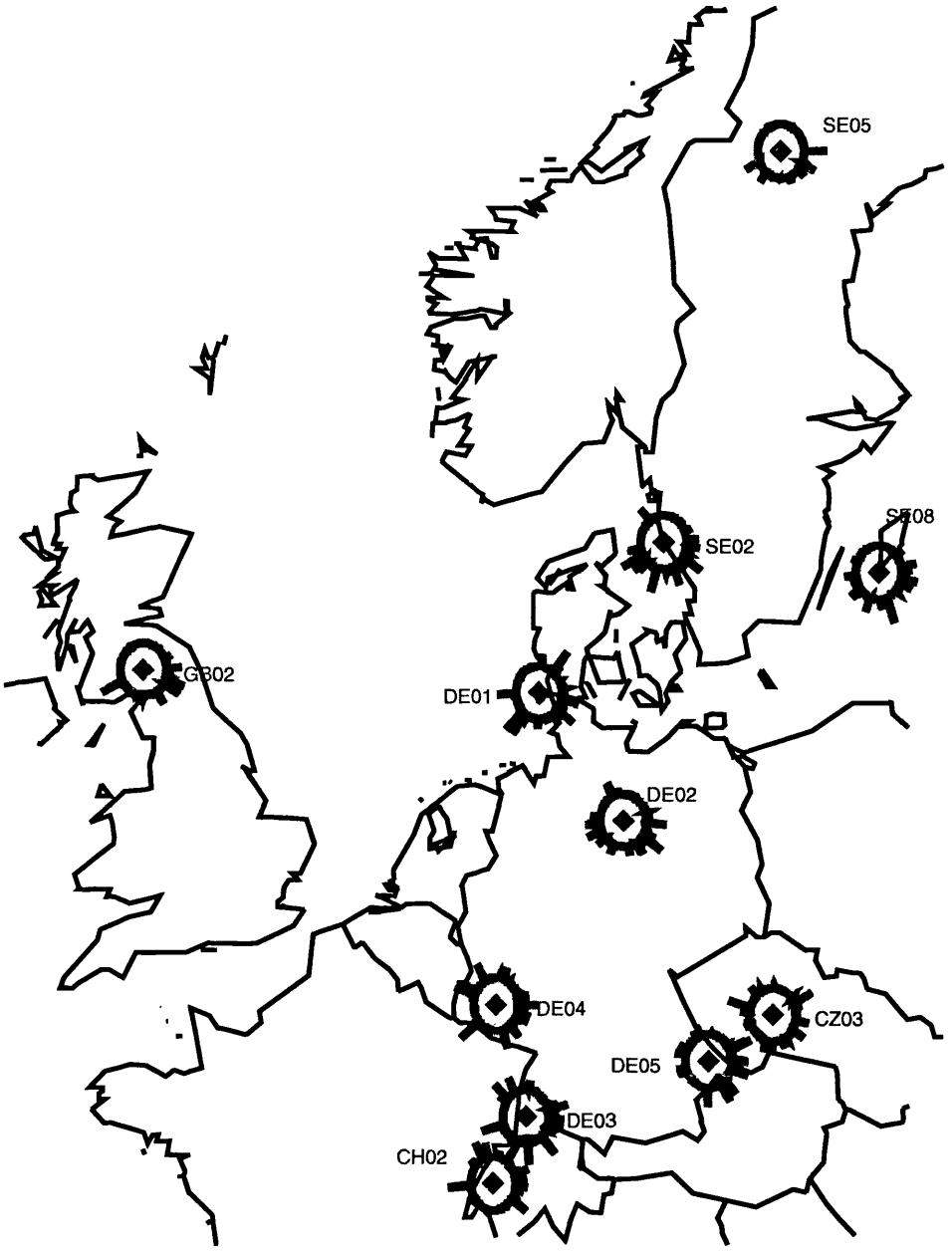

Figure 7.7: Weighted neighbouring emissions for each of the 11 sites computed from equation (7.7), averaged across $t = 1990, \ldots, 2001$. The longer the spikes, the higher the effect of the emissions coming from that direction on the SO_2 concentrations.

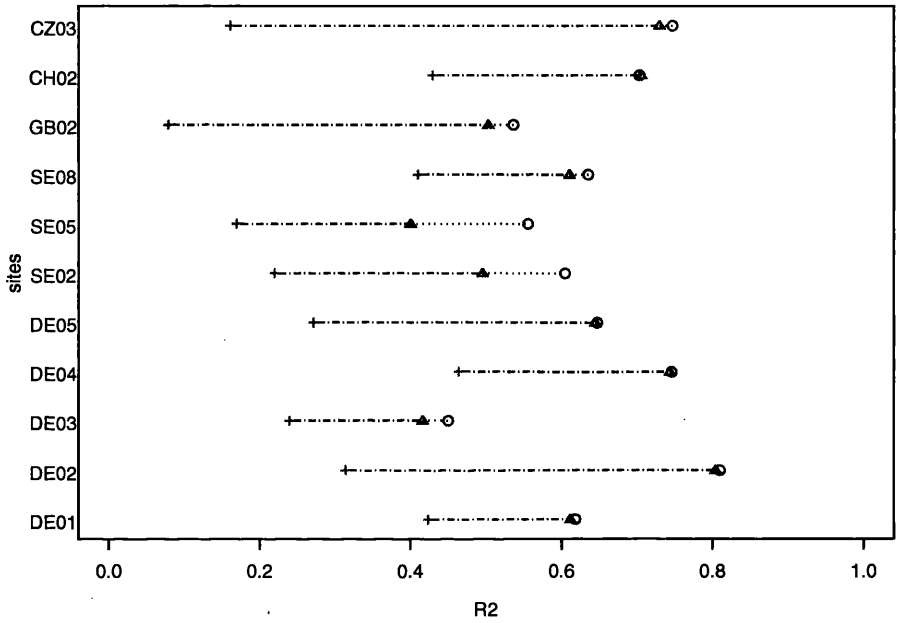

Figure 7.8: R^2s of Model (7.10) (o), Model (7.2) ($\triangle$) and Model (7.4) (+) per each of the 11 sites. The dashed lines shows the differences in R^2s between Model (7.2) ($\triangle$) and Model (7.4) (+), and the dotted lines show the difference in R^2s between Model (7.2) ($\triangle$) and Model (7.10) (o).

Chapter 8

Conclusions

This thesis proposes new nonparametric methodologies for analyzing spatiotemporal data with correlated errors. The flexibility of nonparametric smoothers in modeling trends provides a strong alternative to the usual parametric procedures that rely on assumptions that are often difficult to justify in nature. However, many nonparametric techniques were defined under the assumption of independent data. In addition, computational issues make the use of nonparametric techniques problematic when large data sets are involved. The extension of nonparametric methodologies to deal with correlated data and with large data sets has therefore been the main objective of this work.

The methodologies have been applied to air pollution data monitored in Europe in the last quarter of the 20th century by EMEP (Co-operative Programme for Monitoring and Evaluation of the long Range Transmission of Air Pollutants in Europe), and by OECD (Organization for Economic Co-operation and Development). The issue raised by EMEP and OECD was that from the 1970's co-ordinated international programmes to monitor acidifying air pollution were

initiated in direct response to observed acidification. At the same time, several international protocols on the reduction of acidifying emissions (SO_2, SO_4 etc) were also agreed. The policy question of interest is whether the protocols have resulted in a real improvement in environmental quality and a real change in the acidifying environment. Our spatiotemporal analysis gives strong indications of significant reductions in the SO_2 concentration across Europe in the last quarter of the 20th century.

Chapter 1 presented the data and the scientific questions of interest that have been tackled in this work. Chapter 2 showed some analysis that can be performed with well established nonparametric statistical techniques, some of which rely on assumptions that do not hold with the data we try to analyze.

Chapter 3 proposed a diagnostic for detecting change points (discontinuities) in trends with correlated errors. The procedure was based on that presented by Bowman et al. (2004) and it has been here extended to account for temporal correlation of the data. This is a diagnostic for flagging discontinuities in one-dimensional nonparametric regression. The idea on which this diagnostic is based is to compare at each point two linear smooths of the data. Each smooth is "one-sided" in that it is defined in terms of data lying entirely to the right or entirely to the left of the point at which we wish to test for a discontinuity. If no discontinuity is present we would expect the two smooth estimates to have similar values, but if there is a discontinuity then we might hope to detect a difference between the two. Simulation studies showed good performance of the test, when data are generated from flat, linear, quadratic and sine trends. The proposed discontinuity test has been applied to the SO_2, SO_4 in air and SO_4 in precipitation concentrations across 130 sites in Europe and the results identified

the presence of statistically significant discontinuities, some of which are present in common across sites and compounds.

Univariate and bivariate nonparametric smoothers that account for correlation are proposed in Chapter 4. Local linear regression smoothers (Bowman and Azzalini, 1997) are obtained by solving locally weighted least square problems. In order to account for correlation, the smoothers proposed here are computed by solving locally generalized weighted least squares problems. These generalized local linear regression smoothers produce wider standard error bands. Weights based on the cosine function are used to obtain circular smoothers that can be fitted to seasonal components or directional variables. Estimates of the correlation matrix are obtained analyzing the structures of the residuals from fitting a smoother for independent data.

The generalized local linear regression smoothers are the building blocks of additive models that have been fitted here through a reformulated version of the backfitting algorithm. The algorithm that is proposed here is based on the idea of the backfitting algorithm proposed by Hastie and Tibshirani (1990), but it uses a different matrix formulation in order to obtain the projection matrix of the overall additive model at convergence. The projection matrix and correlation matrix are then used to obtain generalized residual sums of squares and new definitions of degrees of freedom that are needed to compute model selection techniques. The proposed tests are an extension of the approximate F test and the Pseudo Likelihood Ratio test (Bowman and Azzalini, 1997) and simulation studies showed the importance of using these techniques, even when data are characterized by a small amount of correlation. When the correlation is known, the size of the Pseudo Likelihood Ratio test seems to work very well under all

the different settings analyzed. However, the sizes of the approximate F test seem too low, especially when it is used to test for the presence of interaction. When correlation has to be estimated, the size of both tests is dependent on the choice of the smoothing parameters, which affect the estimates of the correlation parameters. However, it seems that the choice of a higher smoothing parameter is a safer approach in terms of not increasing the size of the tests. The power of both tests seems excellent for detecting change of trend of at least 2.5% in 11 years, or changes in amplitude (peak - trough) of more than 1.1 in 11 years, or changes in the phase of the seasonal component of about 2 months across 11 years. On the basis of the approximate F test and the Pseudo Likelihood Ratio test, procedures for testing for no effect and for changes in components between two additive models have also been constructed.

Chapter 5 shows applications of the methodologies presented in Chapter 4 to air pollution data. Trend and seasonal cycles for SO_2 have been studied accounting for correlation and for meteorological effects. Meteorology changes significantly the trend estimates at a few points, but it does not seem to affect the general shape of the time trend. Analysis of the seasonal components show significant changes if meteorology is accounted for, and some sites showed statistically significant changes in seasonality across time. Temporal trend analysis of SO_2 as a function of meteorological variables showed them to have a significant effect in explaining a substantial part of the variability of SO_2. Meteorological variables seem also to be the cause of significant discontinuities detected at 4 of the 11 sites analyzed.

A binned version of the reformulated backfitting algorithm and of the approximate F test that copes with large data sets (such as spatiotemporal ones) and

to account for correlation are presented in Chapter 6. Spatial analysis of sulphur dioxide from 1990 to 2001 at 130 sites across Europe is performed. At each time point, a plane seems to be the most appropriate model for the SO_2 trend surface. Analysis of the resulting residuals showed that the Gaussian variogram model fits better than the Exponential and Spherical models. Time series analysis of the range, the nugget, and the sill showed no evidence that the spatial correlation changes over time. Therefore the average of the estimates for each parameter could be used to define a spatial correlation matrix for the period of analysis.

Time series analysis of each of the 130 sites showed that an AR(1) model is not unrealistic, and that the time correlations seem quite homogeneous over Europe. The average of the estimated correlation coefficients at each of the 130 sites has been used to provide a unique estimate of the temporal correlation across all sites.

A separable space-time covariance structure for SO_2 over Europe across the last 30 years has been used to fit a spatiotemporal additive model. The fit of this spatiotemporal model shows statistically significant changes in the seasonal component across years. The winter values increased and the summer ones reduced from 1990 up to 1993, stressing more the difference between peaks and troughs. From 1994 to 2001 the winter values reduced and the summer ones increased, resulting in a smaller difference between peaks and troughs.

An analysis of the observed SO_2 concentrations as a function of "own country" and "neighbouring countries" emissions (weighted by the neighbouring country distances and wind directions and speeds) has been undertaken in Chapter 7. At each site, a neighbouring countries emissions covariate has been defined as the sum of the neighbouring countries emissions weighted by the distances and by

the wind speeds and directions. Results show a relation between the observed values and the own country emissions. The new neighbouring countries emissions covariate resulted in a statistically significant improvement in the model at more than half of the sites analyzed.

The methodologies proposed in this work relax the assumption of independent errors which some of the most widely used nonparametric techniques rely on. They suit very well the characteristics of the data analyzed, and are general enough to be applied to several spatial and/or temporal data sets. However there are several directions in which the methodologies proposed could be further extended in order to be used in even wider contexts.

An interesting approach could be to derive a test for detecting spatiotemporal discontinuities accounting for correlation. A possible approach could be to use some clustering of the discontinuities that have been obtained in each time series at each site. Another approach could be represented by the two dimensional test proposed by Yap (2004), so that at each time point, discontinuities in space could be detected. A time series analysis of the spatial discontinuities detected could then be applied.

It would be quite useful to extend the additive model fitting and testing techniques proposed here to a multivariate scenario (in order to analyze more pollutants simultaneously for example). In the additive model fitting, it would be quite interesting to implement a smoothing parameter selection tool that accounts for correlation. The reformulated backfitting algorithm that has been presented could be extended in order to account for a cross-validation algorithm that would perform the smoothing parameter selection. Methodologies that deal with fitting and testing non separable spatiotemporal models would relax the

assumption of constant spatial correlation across time and constant time correlation over space. The definition of generalized multivariate local linear regression smoothers that can account for more than two predictors could also give more information on the interactions across all the covariates. Given the computational difficulties of deriving the multivariate version of the local linear regression smoothers, a possible direction of work could be the implementation of different smoothers, such as splines, into the backfitting algorithm. Finally the availability of meteorological variables across all Europe would represent an interesting spatiotemporal analysis.

Bibliography

Aerts, M., G. Claeskens, and M. P. Wand (2002). Some theory for penalised spline generalised additive models. *Journal of Statistical Planning and Inference 103*, 455–470.

Anh, V., R. Wolff, J. Gao, and Q. Tieng (1999). Local linear kernel regression with long-range dependent errors. *Australian & New Zealand J. Statist. 41*(4), 463–479.

Barbieri, A., S. Pozzi, and R. Mosello (2004). Relative contribution of nitrogen and sulphur to deposition acidity and regional modeling in lake maggiore watershed (southern alps, switzerland and italy). *Water, Air, and Soil Pollution 156*, 317–335.

Berge, E., J. Bartnicki, K. Olendrzynsky, and S. Tsyro (1999). Long-term trends in emissions and transboundary transport of acidifying air pollution in europe. *Journal of Environmental Management 57*, 31–50.

Berhane, K. and R. Tibshirani (1998). Generalized additive models for longitudinal data. *The Canadian Journal of Statistics 26*(4), 517–535.

Bhattacharya, P. K. and D. Frierson (1981). A nonparametric control chart for detecting small disorders. *The annals of Statistics 9*(3), 544–554.

Blanchard, C. L., A. Sirois, D. M. Whelpdale, J. Brook, and H. M. Micheals (1996). Evaluation of the capabilities of deposition networks to resolve regional trends and spatial patterns. *Atmospheric Environment 30*(14), 2539–2549.

Bloomfield, P., A. Royle, and L. J. Steinberg (1996). Accounting for meteorological effects in measuring urban ozone levels and trends. *Atmospheric Environment 30*(17), 3067–3077.

Bogaert, P. and G. Christakos (1997). Stochastic analysis of spatiotemporal solute content measurements using a regression model. *Stochastic Hydrology and Hydraulics 11*, 267–295.

Bowman, A. and A. Azzalini (1997). *Applied Smoothing Techniques for Data Analysis: the kernel approach with S-Plus illustrations.* Oxford, U.K.: Oxford University Press.

Bowman, A. and A. Azzalini (2003). Computational aspects of nonparametric smoothing with illustrations from the sm library. *Computational Statistics & Data Analysis 42*(4), 545–560.

Bowman, A., A. Pope, and B. Ismail (2004). Detecting discontinuities in nonparametric regression curves and surfaces. *Technical Report, Department of Statistics, University of Glasgow.*

Brillinger, D. R. (1994). Trend analysis: time series and point process problems. *Environmetrics 5*, 1–19.

Brown, P., K. Kåresen, G. Roberts, and S. Tonellato (2000). Blur-generated non-sperable space-time models. *Journal of the Royal Statistical Society 62 B*(4), 847–860.

Cape, J. N., J. Methven, and L. E. Hudson (2000). The use of trajectory cluster analysis to interpret trace gas measurements at mace head, ireland. *Atmospheric Environment 34*, 3651–3663.

Christakos, G. (1992). *Random Field Models in Earth Sciences.* San Diego, California: Academic Press.

Christakos, G. and M. Serre (2000). Bme analysis of spatiotemporal particular matter distribution in north carolina. *Atmospheric Environment 34*, 3393–3406.

Christakos, G. and V. Vyas (1998). A composite space/time approach to studying ozone distribution over eastern united states. *Atmospheric Environment 32*(16), 2845–2857.

Cleveland, R. B., W. S. Cleveland, J. E. McRae, and I. Terpenning (1990). Stl: A seasonal trend decomposition procedure based on loess. *Journal of Official Statistics 6*, 3–73.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association 74*, 829–836.

Cleveland, W. S. and S. J. Devlin (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association 83*, 596–610.

Cocchi, D., E. Fabrizi, and C. Trivisano (2002). A startified model for the analysis of ozone trends in an urban area. *Working paper GRASPA 13*, 3–24.

Cox, D. and V. Isham (1988). A simple sptial-temporal model of rainfall. *Proceedings of the Royal Society of London 415 A*, 317–328.

Cox, W. and S. Chu (1993). Meteorologically adjusted trends in urban areas: a propbabilistic approach. *Atmospheric Environment 27B*, 425–434.

Cressie, N. (1991). *Statistics for Spatial Data.* USA: Wiley-Interscience.

Cressie, N. and H. Huang (1999). Classes of nonseparable, spatio-temporal stationary covariance function. *Journal of the American Statistical Association 94*(448), 1330–1340.

Davis, J. M., B. K. Eder, D. Nychka, and Q. Yang (1998). Modelling the effects of meteorology on ozone in houston using cluster analysis and generalised additive models. *Atmospheric Environment 32*(14/15), 2505–2520.

Diggle, P., P. Riberio, and O. Christensen (2002). An introduction to model-based geostatistics. *Spatial Statistics and computational methods.*

Dominici, F., A. McDermott, S. L. Zeger, and J. M. Samet (2002). On the use of generalised additive model in time series studies of air pollution and health. *American Journal of Epidemiology 156*(3), 193–203.

Downing, C. E. H., K. J. Vincent, G. W. Campbell, D. Fowler, and R. J. Smith (1995). Trends in wet and dry deposition of sulphur in the united kingdom. *Water Air and Soil Pollution 85*(2), 659–664.

El-Shaarawi, A. H. (1995). Trend detection and estimation with environmental applications. *Mathematics and Computers in Simulation 39*, 441–447.

Esterby, S. R. (1993). Trend analysis methods for environmental data. *Environmetrics 4*, 459–481.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Annals of Statistics 21*, 196–216.

Fan, J. and I. Gijbels (1992). Variable bandwidth and local linear regression smoothers. *Annals of Statistics 20*(4), 2008–2036.

Fisher, N., T. Lewis, and B. Embleton (1993). *Statistical Analysis of Spherical Data*. UK: Cambridge University Press.

Gardner, M. W. and S. R. Dorling (2000a). Meteorologically adjusted trends in uk daily maximum surface ozone concentrations. *Atmospheric Environment 34*, 171–176.

Gardner, M. W. and S. R. Dorling (2000b). Statistical surface ozone models: an improved methodology to account for non-linear behaviour. *Atmospheric Environment 34*, 21–34.

Giraitis, L., R. Leipus, and D. Surgailis (1996). The change-point problem for dependent observation. *Journal of Statistical Planning and Inference 53*, 297–310.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space time data. *Journal of the American Statistical Association 97*(458), 590–600.

Gneiting, T. and M. Schlather (2002). Space-time covariance models. *Encyclopedia of Environmetrics 4*, 2041–2045.

Gombay, E. and L. Horvath (1997). An application of the likelihood method to change-point detection. *Environmetrics 8*, 459–467.

Granger, C. W. J. (1966). The typical spectral shape of an economic variable. *Econometrica 34*(1), 150–161.

Green, P., C. Jennison, and A. Seheult (1985). Analysis of field experiments by least squares smoothing. *Journal of the Royal Statistical Society Series B-Methodological 47*(2), 299–315.

Green, P. and B. W. Silverman (1994). *Nonparametric Regression and Linear Models: A Roughness Penalty Approach.* London: Chapman & Hall.

Guttorp, P., W. Meiring, and P. Sampson (1994). A space-time analysis of ground level ozone data. *Environmetrics 5*(3), 241–254.

Hall, P. and D. M. Titterington (1992). Edge-preserving and peak-preserving smoothing. *Technometrics 38*(4), 429–440.

Hastie, T. and R. Tibshirani (1987). Generalised additive models: some applications. *Journal of American Statistical Association 82*(398), 371–386.

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

Hastie, T. and R. Tibshirani (2000). Bayesian backfitting. *Statistical Sciences 15*(3), 196–223.

Hawkins, D. M. (2001). Fitting multiple change-point models to data. *Computational Statistics and Data Analysis 37*, 323–341.

Healy, M. J. R. (1986). *Matrices for Statistics.* Oxford University Press: Oxford.

Hess, A., H. Iyer, and W. Malm (2001). Linear trend analysis: a comparison of methods. *Atmospheric Environment 35*, 5211–5222.

Hirsch, R. M. and J. R. Slack (2002). A nonparametric trend test for seasonal data with serial dependence. *Water Resources Research 20*(6), 727–732.

Holland, D. M., P. P. Principe, and J. E. Sickles (1999). Trends in atmospheric sulfur and nitrogen species in the eastern united states for 1989 - 1995. *Atmospheric Environment 33*, 37–49.

Horvath, L. (2001). Change-point detection in long-memory processes. *Journal of Multivariate Analysis 78*, 218–234.

Horvath, L. and P. Kokoszka (2002). Change-point detection with non-parametric regression. *Statistics 36*, 9–31.

Huang, L. S. and R. L. Smith (1999). Meteorologically-dependent trends in urban ozone. *Environmetrics 10*, 103–118.

Huerta, G., B. Sanso', and J. Stroud (2004). A spatiotemporal model for mexico city ozone levels. *Applied Statistics 53*(2), 231–248.

Hùnová, I., J. Šantroch, and Ostatnická (2004). Ambient air quality and deposition trends at rural stations in the czeck republic during 1993-2001. *Atmospheric Environment 38*, 887–898.

Jaruskova, D. (1997). Some problems with application of change-point detection methods to environmental data. *Environmetrics 8*(5), 469–483.

Jaruskova, D. (1998). Testing appearance of linear trend. *Journal of Statistical Planning and Inference 70*, 263–276.

Johnson, N. and S. Kotz (1972). *Distributions in Statistics: Continuous Univariate Distributions, Vol. II.* New York: Wiley.

Jones, R. and Y. Zhang (1997). Models for continuous space-time process. *Modelling Longitudinal and Spatially correlated Data*, 289–298.

Kammann, E. and M. Wand (2003). Geoadditive models. *Applied Statistics 52*(1), 1–18.

Koo, J. Y. (1997). Spline estimation of discontinuous regression functions. *Journal of Computational and Graphical Statistics 6*(3), 266–284.

Kopka, H. and P. Daly (1995). *A guide to LaTeX2ε*. Reading, U.K.: Addison-Wesley Publishing Company.

Kyriakidis, P. and A. Journel (1999). Geostatistical space-time models: A review. *Mathematical Geology 31*(6), 651–684.

Kyriakidis, P. and A. Journel (2001a). Stochastic modeling of atmospheric pollution: a spatial time-series framework. part i: methodology. *Atmospheric Environment 35*, 2331–2337.

Kyriakidis, P. and A. Journel (2001b). Stochastic modeling of atmospheric pollution: a spatial time-series framework. part ii: application to monitoring

monthly sulfate deposition over europe. *Atmospheric Environment 35*, 2339–2348.

Libiseller, C. (2003). *Considering Meteorological Variation in Assessments of Environmental Quality Trends.* Ph. D. thesis, Linköpings Universitet.

Libiseller, C. and A. Grimvall (2002). Performance of partial mann-kendall test for trend detection in the presence of covariates. *Environmetrics 13*, 71–84.

Libiseller, C. and A. Grimvall (2003). Model selection for local and regional approaches to meteorological normalisation of background concentrations of tropospheric ozone. *Atmospheric Environment 37*(28), 3895–4035.

Libiseller, C., A. Grimvall, J. Walden, and J. Paatero (2003). Meteorological normalization of tropospheric ozone using back trajectories. *Technical Report, Department of Mathematics, Linköpings Universitet.*

Libiseller, C. and A. Nordgaard (2003). Variance reduction for trend analysis of hydrochemical data in brackish waters. *LIU-MAT-R-2003-02.*

Loader, C. (1999). *Local Regression and Likelihood.* New York: Springer-Verlag.

Loader, C. R. (1996). Change point estimation using nonparametric regression. *The Annals of Statistics 24*(4), 1667–1678.

Lombard, F. (1988). Detecting change points by fourier analysis. *Technometrics 30*(3), 305–310.

Luo, Z., G. Wahba, and D. Johnson (1998). Spatial-temporal analysis of temperature using smoothing spline anova. *Journal of Climate 11*(1), 18–28.

Lynch, J. A., V. C. Bowersox, and J. W. Grimm (2000). Changes in sulfate deposition in eastern usa following implementation of phase i of title iv of the clean air act amendments of 1990. *Atmospheric Environment 34*, 1665–1680.

Lynch, J. A., J. M. Grimm, and V. C. Bowersox (1995). Trends in precipitation chemestry in the united states: a national perspective, 1980 - 1992. *Atmospheric Environment 29*(11), 1231–1246.

Mardia, B. and C. Goodall (1993). Spatial-temporal analysis of multivariate environmental monitoring data. *Multivariate Environmental Statistics*, 347–386.

Mardia, K. V. (1972). *Statistics of directional data.* London: Academic Press.

McDonald, J. A. and A. B. Owen (1986). Smoothing with split linear fits. *Technometrics 28*(3), 195–208.

McLeod, A., K. Hipel, and B. Bodo (1991). Linear trend analysis: a comparison of methods. *Environmetrics 2*, 169–200.

McMullan, A. (2004). *Non-linear and nonparametric modelling of seasonal environmental data.* Ph. D. thesis, The University of Glasgow.

McMullan, A., A. Bowman, and E. Scott (2005). Nonparametric modeling using additive models with correlated data. *Technical Report, Department of Statistics, The University of Glasgow.*

Meinhold, R. and N. Singpurwalla (1983). Understanding the kalman filter. *The American Statistiscian 37*(2), 123–127.

Meiring, W., P. Guttorp, and P. Sampson (1998). Space-time estimation of grid cell hourly ozone levels for assessment of a deterministic model. *Environmental and Ecological Statistics 5*, 197–222.

Müller, H. (1992). Change-points in nonparametric regression analysis. *The Annals of Statistics 20*(2), 737–761.

Müller, H. G. and K. S. Song (1997). Two-stage change-points estimators in smooth regression models. *Statistics and Probability Letters 34*, 323–335.

Niu, X. F. (1996). Nonlinear additive models for environmental time series, with applications to ground-level ozone data analysis. *Journal of the American Statistical Association 91*(435), 1310–1321.

Odgen, T. and E. Parzen (1996). Data dependent wavelet thresholding in non-parametric regression with change-points applications. *Computational Statistics & Data Analysis 22*, 53–70.

Oehlert, G. (1993). Regional trend in sulfate wet deposition. *Journal of the American Statistical Association 88*(422), 390–399.

Opsomer, J. (2000). Asymptotic properties of backfitting estimators. *Journal of multivariate Analysis 73*, 166–179.

Opsomer, J. and D. Ruppert (1997). Fitting a bivariate additive model by local polynomial regression. *The Annals of Statistics 25*(1), 186–211.

Opsomer, J. and D. Ruppert (1998). A fully automated bandwidth selection method for fitting additive models. *Journal of the American Statistical Association 93*(442), 605–619.

Opsomer, J., Y. Wang, and Y. Yang (2001). Nonparametric regression with correlated errors. *Statistical science 16*(2), 134–153.

Qiu, P. and B. Yandell (1997). Jump detection in regression surface. *Journal of Computational and Graphical Statistics 6*(3), 332–354.

Qiu, P. and B. Yandell (1998). A local polynomial-jump algorithm in non parametric regression. *Technometrics 40*(2), 141–152.

Rao, S. and I. Zurbenko (1994). Detecting and tracking changes in ozone air quality. *Journal of Air and Waste Management Association 44*, 1089–1092.

Reinsel, G. C. and G. C. Tiao (2002). Impact of chlorofluoromethanes on stratospheric ozone. *Journal of the American Statistical Association 82*, 20–30.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Annals of Statistics 12*, 1215–1230.

Ruppert, D. and M. P. Wand (1994). Multivariate locally weighted least squares regression. *Annals of Statistics 22*, 1346–1370.

Sampson, P. and P. Guttorp (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association 87*(417), 108–119.

Searle, S. (1971). *Linear Models.* New York: John Wiley & Sons.

Seber, G. (1977). *Linear Regression Analysis.* New York: John Wiley & Sons.

Shannon, J. D. (1999). Regional trends in wet deposition of sulfate in the

united states and so2 emissions from 1980 through 1995. *Atmospheric Environment 33*, 807–816.

Shively, T. S. and T. W. Sager (1999). Semiparametric regression approach to adjusting for meteorological variable in air pollution trends. *Environmental Sciences and Technology 33*(21), 3873–3880.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics.* Springer-Verlag: New York.

Sirois, A. (1997). Temporal variation of oxides of sulphur and nitrogen in ambient air in eastern canada: 1979-1994. *Tellus 49B*, 270–291.

Thompson, M. L., J. Reynolds, L. H. Cox, P. Guttorp, and P. D. Sampson (2001). A review of statostical methods for meteorological adjustment of tropospheric ozone. *Atmospheric Environment 35*, 617–630.

Tiao, G. C., G. C. Reinsel, D. Xu, J. H. Pedrick, X. Zhu, A. J. Miller, J. J. DeLuisi, C. L. Mateer, and D. J. Wuebbles (1990). Effects of autocorrelation and temporal sampling schemes on estimates of trend and spatial correlation. *Journal of Geographical Research 95*, 20507–20517.

Tørset, K., A. Wenche, and S. Solberg (2001). Trends in airborne sulphur and nitrogen compounds in norway during 1985-1996 in relation to air mass origin. *Water, Air, and Soil Pollution 130*, 1493–1498.

Venables, W. N. and B. D. Ripley (1994). *Modern Applied Statistics with S-PLUS.* Springer-Verlag: New York.

Vuorenmaa, J. (2004). Long-term changes of acidifying deposition in finland (1973-200). *Environmental Pollution 128*, 351–362.

Vyas, V. and G. Christakos (1997). Spatiotemporal analysis and mapping of sulfate deposition data over eastern usa. *Atmospheric Environment 31*(21), 3623–3633.

Wahba, G. (1990). *Spline models for observational data.* Philadelphia: SIAM.

Wang, Y. (1995). Jump and sharp cusp detection by wavelets. *Biometrika 82*(2), 385–397.

Wang, Y. (1998). Change curve estimation via wavelets. *Journal Of The American Statistical Association 93*(441), 163–172.

Weatherhead, E. C., G. C. Reinsel, G. C. Tiao, X.-L. Meng, M. Choi, and W. a. Cheang (1998). Factors affecting the detection of trends: Statistical considerations and applications to environmental data. *Journal of Geophysical Research 103*(D14), 17149–17161.

Wikle, C., L. Berliner, and N. Cressie (1998). Hierarchical bayesian space-time models. *Environmental and Ecological Statistics 5*, 117–154.

Wikle, C. and N. Cressie (1999). A dimension-reduced approach to space-time kalman filtering. *Biometrika 86*(4), 815–829.

Wood, S. and N. Augustin (2002). Gams with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling 157*, 157–177.

Yap, C. (2004). *Detecting discontinuities using nonparametric techniques with correlated data.* Ph. D. thesis, The University of Glasgow.

Yue, S., P. Pilon, B. Phinney, and G. Cavadias (2002). The influence of the autocorrelation on the ability to detect trend in hydrological series. *Hydrological Processes 16*, 1807–1829.