Statistical Modelling Of Environmental Trends Over Both Time And Space

Francisco Andrés Rincón

A Dissertation Submitted to the
University of Glasgow
for the degree of
Master of Science

Department of Statistics

September 2009

Abstract

The analysis of environmental data represents an opportunity to use statistical tools to provide a better understanding of changes over time and space, making it easier to tackle problems such as pollution, water quality or climate change.

The analysis of environmental data requires methodologies that allow us to fit models capable of explaining seasonal patterns and changes observed over time and space.

The work developed in this thesis is centred on the modelling of trends over time and space simultaneously. The analysis of this information could be carried out in a marginal manner over time and space; however the main objective of this thesis is to fit a model using both time and space simultaneously to be able to provide a closer representation of environmental data.

The data used in this thesis were provided by the Environmental Change Network (ECN), the Acid Water Monitoring Network (AWMN) and the Macaulay Institute. These data sets correspond to water quality where the main interest is to assess change over time in the case of the ECN and the AWMN and changes over time and space for the Macaulay Institute.

A brief description of the problems of water quality, a description of linear and nonparametric models and the main goals of the analysis are given in chapter 1.

Chapter 2 provides a detailed analysis of the information provided by the ECN and AWMN through the use of linear models, indicating the advantages and disadvantages of this approach. The aim is to explain changes over time for 11 variables related to surface water chemistry and to assess differences between both sources of information.

Chapter 3 shows the modelling of a catchment using additive models to capture the trend over time and space for water quality measures for rivers. The modelling of the trend over space uses 17 sites under the assumption that Euclidean distance is a sensible measure.

A test to assess the need for a linear effect opposed to a nonparametric effect was carried out as well as a sensitivity analysis, assessing stability in the conclusions under different degrees of freedom. From residuals the need for a covariance structure over time and/or space is assessed. A comparison between models to evaluate the improvement under the inclusion of river flow information is also included in this chapter.

Chapter 4 provides an introduction to the problem of modelling river networks indicating the difficulty in obtaining an adequate spatial model. The inclusion of this idea into an additive model to tackle this problem, allows us to capture the trend over space assuming that not all the points are flow connected, using upstream distance.

A test to assess the need for a linear effect rather than a nonparametric effect as well as a sensitivity analysis to evaluate stability under different degrees of freedom, were carried out with the final models. Finally a comparison between a model using Euclidean distance and upstream distance to model the trend over space is made to choose the model with better performance.

The scientific questions and a discussion of the main outputs observed for each data set is presented at the end of each chapter. Chapter 5 summarises the main outputs and findings of the thesis, indicating advantages and disadvantages of each methodology used as well as some ideas for future work.

Throughout the development of this thesis all the analysis were carried out using R software [Venables & Smith (2009)]. The packages used provided a suitable tool to cover descriptive analysis, modelling process and graphic displays. The large majority of the analysis were carried out using the packages Stats, nlme, sm, lattice and geoR.

Acknowledgements

The information used in this thesis was provided by the Environmental Change Network (ECN), the Acid Water Monitoring Network (AWMN) and the Macaulay Institute. I would like to thanks for the cooperation and feedback from these institutions through this project.

This has been a challenging process for many different reasons, nevertheless I was lucky and had two excellent tutors to guide me over this year. I would like to thank Marian Scott and Adrian Bowman for sharing all their experience and knowledge, allowing that this learning experience to turn out exactly as I thought one year before when I chose Glasgow to do my Master program abroad.

Finally but not less important, to my family in Colombia: Es bastante duro dejar a las personas que quieres y que te quieren, al final todas las metas personales solo son relevantes cuando las compartes con alguien a quien quieres y extrañas.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Nowadays environmental changes have become a critical topic in the agenda for academics and governments around the world. Answers that allow us to better understand climate change, pollution, overpopulation or renewable and efficient energy, are demanded now more than ever.

From a statistical point of view this represents a challenging task to provide proper guidance to policy makers, assessing how effective the decisions taken in the past have been or how they can change current policies.

The analysis of environmental data presents a complex problem for several reasons: there is usually a large number of variables involved and the collection process can be challenging, with spatial and temporal components. These two reasons make it harder to identify the existence, the magnitude and the factors associated with environmental changes. For these reasons it is necessary to move from classical statistical approaches to modern statistical methodologies and in some cases, depending on the problem, to use more than one methodology simultaneously.

1.1 Environmental Organizations

The development of this project was made possible with the support and information provided by the Environmental Change Network (ECN), the Acid Waters Monitoring Network (AWMN) and the Macaulay Institute.

1.1.1 ECN

The Environmental Change Network [ECN (Webpage)] is the UK's long term environmental monitoring programme. The main aim is to analyse long term data based on a set of physical, chemical and biological variables, which drive and respond to environmental changes at a range of terrestrial and fresh water sites across the UK. Established in 1992, nowadays the ECN programme has 12 terrestrial and 45 fresh water sites divided into river sites (29) and lake sites (16).

The ECN is a multi-agency programme sponsored by a consortium of 14 UK government departments and agencies with four main objectives (Information obtained from the website of the ECN).

- To establish and maintain a selected network of sites within the UK from which to obtain comparable long-term data sets through the monitoring of a range of variables, identified as being of major environmental importance.
- To provide for the integration and analysis of these data, so as to identify natural and man-induced environmental changes and improve understanding of the causes of change.
- To distinguish short-term fluctuations from long-term trends, and predict future changes.
- To provide, for research purposes, a range of representative sites with good instrumentation and reliable environmental information.

1.1.2 AWMN

The Acid Waters Monitoring Network [AWMN (Webpage)] was established in 1988 to monitor the effect of acid deposits in the UK. The information collected provides a long term data set of water chemistry and biology, obtained from a network corresponding to 11 lakes and 11 streams across the UK. Samples for extensive analysis of chemical determinants, including pH, DOC, conductivity and a standard suite of base cations, anions and metals are collected regularly over all the sites.

1.1.3 Macaulay Institute

The Maculay Institute [M.I. (Webpage)], founded in 1930, is an international research group with a main interest in sustainable uses of land and its natural resources. One of their research interests is the exploration of the relationships between land use and catchment management, aiming to understand how pollutants move throughout the environment, assessing the impact of pollution on soil and water, developing methodologies to predict the effect of human activities in the environment and to provide scientific evidence to develop and implement government policies.

1.2 Water Quality

The increasing demand to provide clean water for human consumption, agriculture and industry is an important topic around the world [Bates et al. (2008)]. The fact that 70% of the surface of the earth is water seems to be a good reason not to be too concerned about this issue; however the reality is that only 3% corresponds to fresh water while the remainder corresponds to salt water.

There are several sources of pollution that might affect the quality of the water; some, such as litter, can be observed easily. Some, including pathological agents like bacteria, require the use of a microscope to identify their presence in

water, while others such as chemical polluting agents, require more sophisticated chemical analysis to detect them.

The main sources of water pollution are:

- Sewage
- Oil
- Fertilizers and Pesticides
- Soil Sediments
- Industrial Waste

The negative effect that comes with the deterioration of water can have an impact on human health. The increase of illnesses such as diarrhoea, malaria or cholera, illnesses which affect mainly elderly and young children, along with the effect of oils, plastics, pesticides, detergents and personal care products, which are related to nervous system damage and some specific types of cancers, are examples of how change in water quality can affect us.

To be able to tackle this problem the best strategy is to monitor the biological and chemical characteristics of water aiming to:

- Evaluate water quality for recreation
- Evaluate water quality for fishing
- Evaluate trends and policies
- Evaluate current technologies for use in water treatment plants

An ever growing human population, industry and agriculture production are some of the main concerns with respect to water quality. Therefore, the European Union designed a new legal framework for the protection, improvement and sustainable uses of all water bodies.

The Water Framework Directive (WFD) [European Commission (2000)], approved on 22 December 2000, establishes a long term policy for the management of rivers, lakes, groundwater and coastal beaches. The WFD not only define quality standards for certain types of water (bathing water, fish and shellfish water, and water used for drinking water abstraction), the new innovation is that it covers all water bodies, not only those for human uses, specifying a quality status "good status" which is measurable and specific for each type of water body, with agreed deadlines which reflect regional diversity.

1.3 Statistical Methodology

The methodology presented in this thesis includes linear regression and nonparametric regression applied to environmental data with the aim of identifying trends over time and space.

Linear regression allows us to identify trends over time and space only when a single hyperplane captures the information provided for all the covariates. When the information is collected over time and space simultaneously, it is also necessary to assess the assumptions of independence over time and space for the residuals.

In environmental data in some cases it is too simplistic to assume that trends over time and space follow a linear pattern. Seasonal patterns over time and changes over space are common, indicating that a linear parameter suggesting an upward or downward trend is not enough to capture the variability involved.

Nonparametric regression provides a useful tool to identify the presence of trends when the information collected over time and space is not properly explained

by a linear model. Nonparametric regression follows the same idea of linear regression and the same assumptions, although rather than linear parametric effect the nonparametric effect correspond to smooth curves, or surfaces in the case of interactions.

The advantage of nonparametric regression is the flexibility provided by the smoothing parameter or equivalent degrees of freedom [Hastie & Tibshirani (1990)]. This allows a wide range of possibilities, but there is a cost to this flexibility in a trade off between bias and variance. Large smoothing parameter values decrease the variance but tend to increase the bias, and conversely.

1.4 Aims

The aim of this research project is to present statistical methodology applied to three environmental data sets to identify trends over both time and space, dealing with missing observations, uneven sampling and non-linear relationships.

Through the development of this thesis, the main objective is to present different approaches to the analysis of environmental data, looking for suitable methodology. The analysis made of the information provided by the ECN and the AWMN starts with a linear approach to evaluate the strengths and weaknesses of this methodology.

The analysis of the Macaulay Institute information involves the use of additive models to capture non-linear trends over time and space, assessing the need to include a time and space covariance structure. In addition, a methodology to analyse river network information when the data observed has been collected over time and space simultaneously is also explored.

The objectives of this thesis are:

1. To evaluate the presence of trend over time for the information provided for the ECN and AWMN data.

2. To identify any differences between the two sources of information, ECN and AWMN.

3. To assess whether there is an improvement in water quality over time and space for the information provided by the Macaulay Institute, using standard spatial methods.

4. To assess whether there is an improvement in water quality over time and space using river distance, including the fact that not all the sites are flow-connected and to evaluate whether river distance provides a more suitable form of model.

The main interest in fulfilling these aims is to provide useful tools to analyse environmental data. Since statistical models do not reproduce reality perfectly, the idea is to obtain statistical models which capture the behaviour of environmental events well, while allowing useful conclusions to be drawn.

Chapter 2

AWMN and ECN

This chapter describes the analysis of two data sets provided by the Environmental Change Network (ECN) and the Acid Waters Monitoring Network (AWMN) from a location in the Cairngorms (Figure 2.1). The Cairngorms site joined the ECN network in 1999. It lies on the western flank of the Cairngorms and is the catchment of the Allt a' Mharcaidh, one of the freshwater sites of the ECN. With an area of 1000 ha and an altitude of 1110m, this place has relatively low levels of air, water and soil pollution compared to other places in the UK, making it a good control place for monitoring changes in the environment.

The ECN data have been collected from Aug 1999 to Dec 2006. The information was collected unevenly and some months have more observations than others. The AWMN data have been collected from July 1988 to March 2007. The frequency for the collection process was one observation per month, although December 2001, July 2002 and December 2002 have a second observation.

Throughout this chapter the main aim is to identify trends and seasonal components over time and to assess whether there are differences between both sources of information. This analysis was carried out for 11 variables related to surface water chemistry (pH, DOC, Sodium (Na), Calcium (Ca), Magnesium (Mg), Iron (Fe), Chloride (Cl), NO_3, Aluminium (Al), Potassium (K) and SO_4(S)).

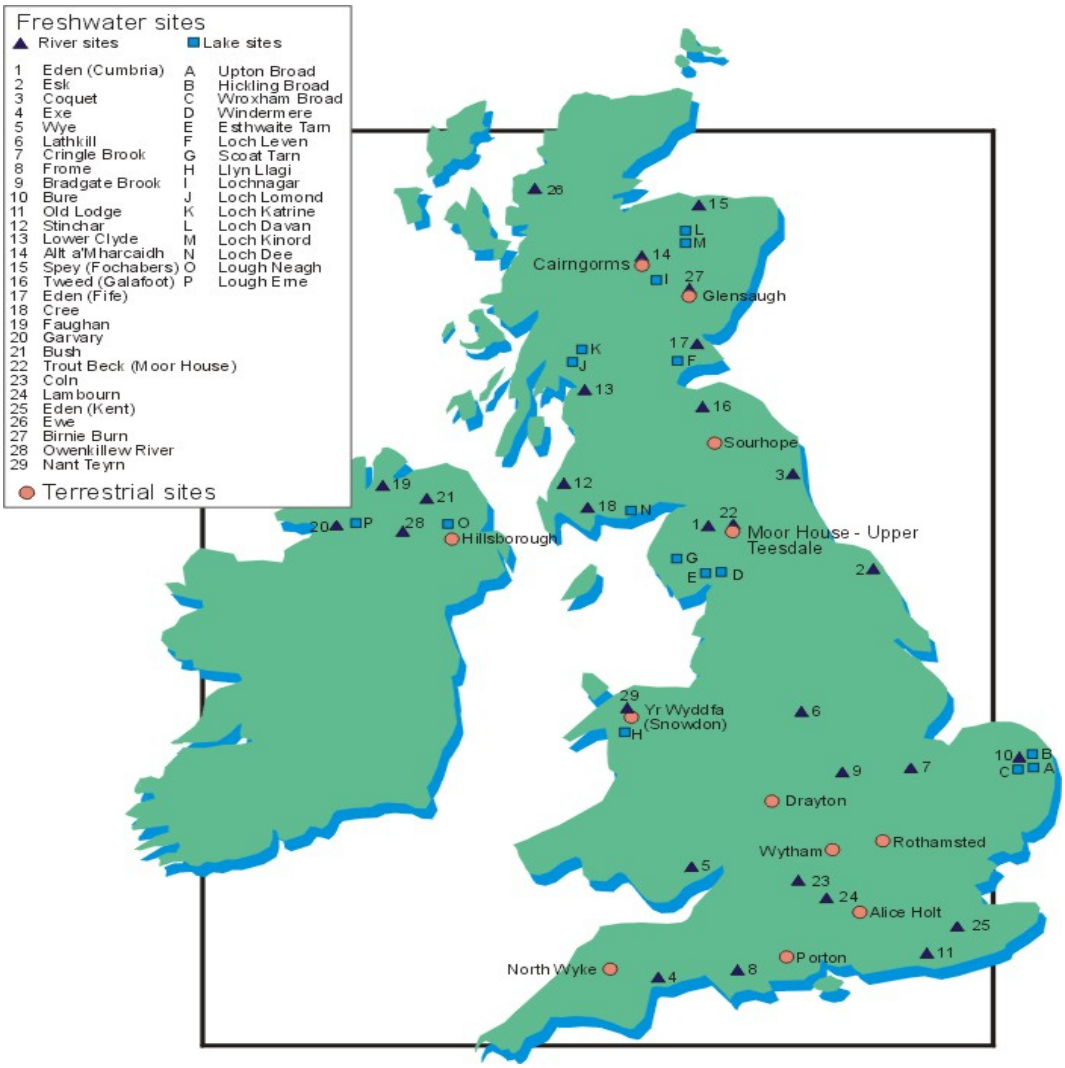

Figure 2.1. ECN network sites

Throughout this thesis, limits of detection values (LOD) were included in the analysis. There are several methods to deal with this problem [Cothern & Ross (1994)], such as use the detection limit value, half the detection limit or replace them by zero. The method used to handle the limit of detection values, was the one adopted by the ECN and the AWMN, where a value equal to zero was assigned for those observation classified as LOD .

2.1 Data Description and Transformations

A log transformation was applied to DOC to remove skewness and to obtain a more normally distributed scale. For Sodium, Calcium, Magnesium, Iron, Chloride, NO_3, Aluminium, Potassium and SO_4(S) a log transformation was also applied, but a constant was added as $log(x + c)$ to deal with values close to zero. The constants added for each of the variables were: Sodium c=5, Calcium c=2, Magnesium c=0.5, Iron c=0.5, Chloride c=3, NO_3 c=0.5, Aluminium c=0.5, Potassium c=0.5 and SO_4(S) c=1.

Figures 2.2 and 2.3 show the variables after the transformation was applied. Only pH remained on the original scale. The time series show that the AWMN data were collected for a longer period of time. The ECN data shows greater variability than AWMN data. A downward trend for $log(DOC)$, $log(Na+5)$ and $log(Ca+2)$ is observed in the ECN data, while only for $log(DOC)$ in the AWMN data is there a clear upward trend.

The larger variability in the ECN data set compared to the AWMN may be explained by the collection process. The ECN data capture not only the variability of the variables over the year, but also capture the monthly variability with a higher number of observations per month. This characteristic is not observed in the AWMN data, which provide only one observation per month and so the

variability within each month in not observed.

Lower variability for variables such as $log(Fe+0.5)$, $log(NO_3+0.5)$ and $log(Al+0.5)$ in the AWMN can be explained by the presence of limit of detection values. Values 0.075, 0.018, and 0.02 are observed, corresponding to 43.75% (98), 93.27% (203) and 64.57% (144) of the total number of observations respectively.

Evidence of differences between the locations where the information was collected may suggest that the physical characteristics for the two catchments are different, according to the descriptive analysis made to the data at both sites.

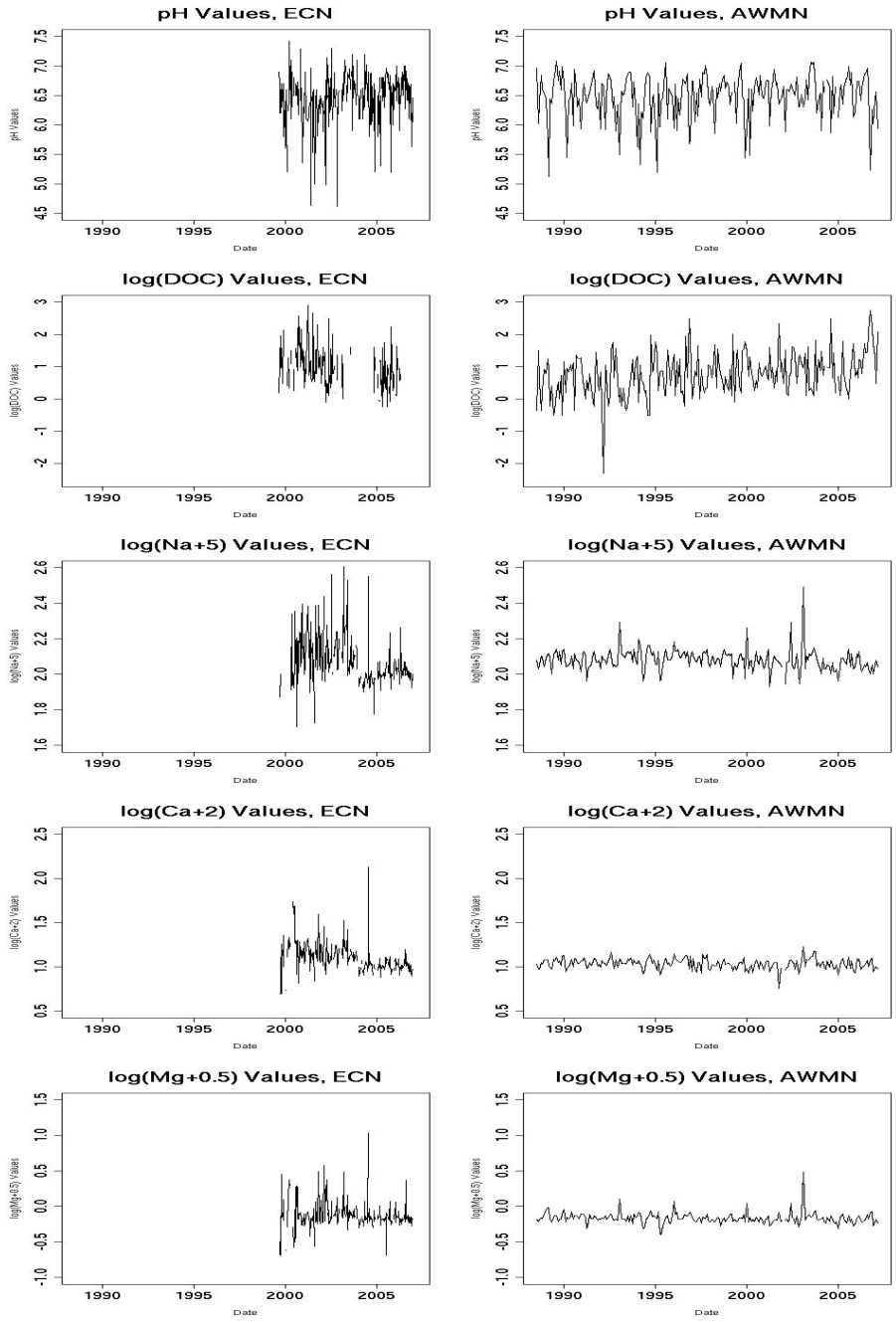

Figure 2.2. Time series graphs for variables pH, $log(DOC)$, $log(Na + 5)$, $log(Ca + 2)$ and $log(Mg + 0.5)$ at ECN and AWMN sites

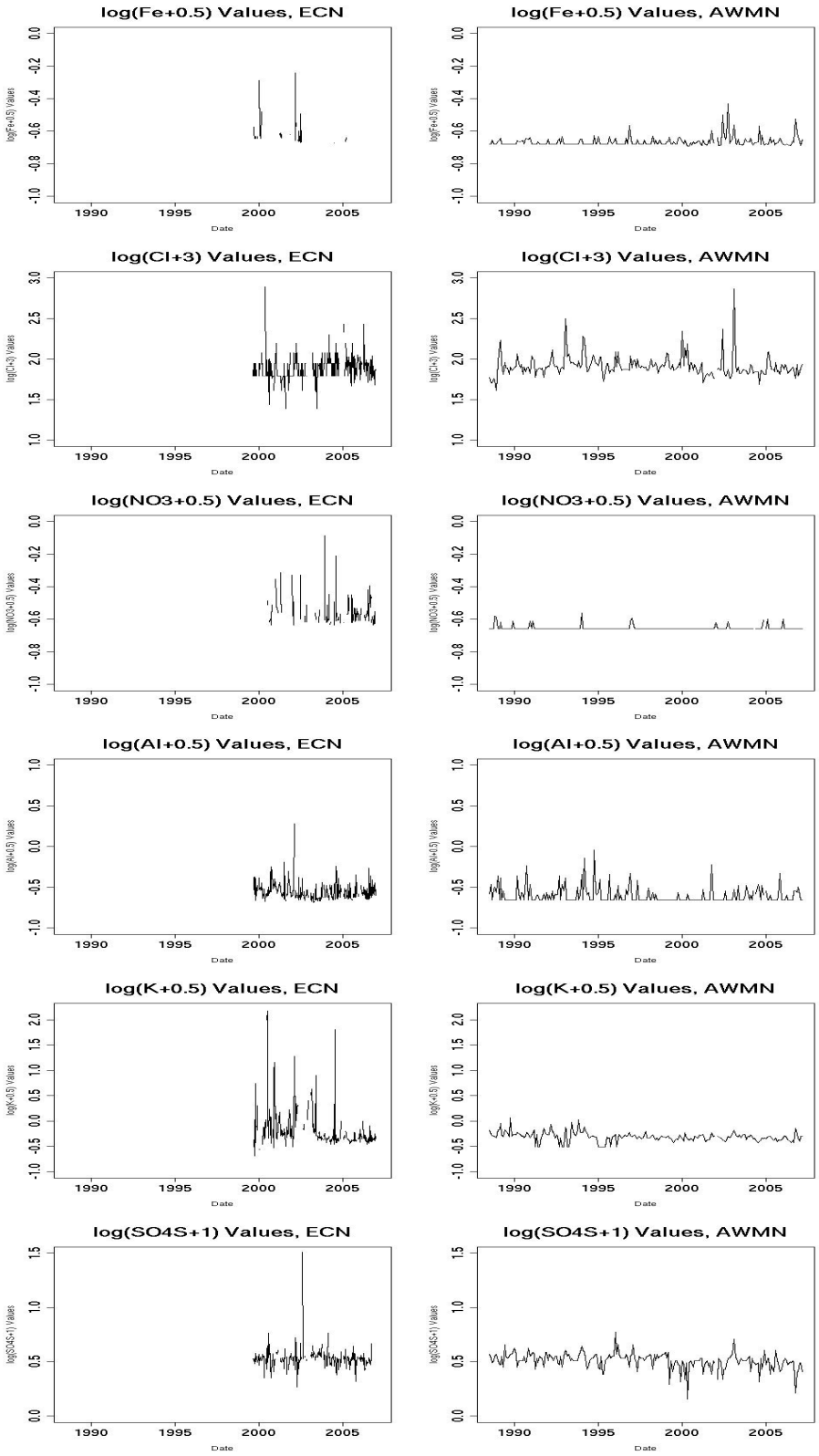

Figure 2.3. Time series graphs for variables $log(Fe{+}0.5)$, $log(Cl{+}3)$, $log(NO_3{+}0.5)$, $log(Al+0.5)$, $log(K+0.5)$ and $log(SO_4S+1)$ at ECN and AWMN sites

2.2 Descriptive Analysis

Table 2.1 shows a summary for each variable for both data sets. For pH and $log(DOC)$ both data sets lie in the same range of values. For the other variables, there is more variability for the information provided by the ECN, although 75% of the distribution for both data sets lie between the same values.

DESCRIPTIVE ANALYSIS								
Variable	Min	1stQ	Median	Mean	3rdQ	Max	SD	n
pH ECN	4.610	6.300	6.500	6.478	6.700	7.420	0.401	313
pH AWMN	5.120	6.333	6.560	6.484	6.740	7.080	0.380	222
$log(DOC)$ ECN	-0.248	0.530	0.916	0.946	1.308	2.912	0.605	177
$log(DOC)$ AWMN	-2.303	0.262	0.788	0.766	1.238	2.754	0.681	223
$log(Na+5)$ ECN	1.702	1.993	2.029	2.063	2.104	2.608	0.124	285
$log(Na+5)$ AWMN	1.932	2.041	2.079	2.081	2.116	2.493	0.059	224
$log(Ca+2)$ ECN	0.694	0.994	1.061	1.087	1.151	2.129	0.164	292
$log(Ca+2)$ AWMN	0.756	0.993	1.040	1.035	1.075	1.230	0.059	224
$log(Mg+0.5)$ ECN	-0.691	-0.205	-0.164	-0.141	-0.105	1.033	0.191	293
$log(Mg+0.5)$ AWMN	-0.400	-0.198	-0.174	-0.164	-0.127	0.488	0.07	224
$log(Fe+0.5)$ ECN	-0.673	-0.644	-0.631	-0.596	-0.583	-0.242	0.09	59
$log(Fe+0.5)$ AWMN	-0.691	-0.678	-0.678	-0.665	-0.659	-0.430	0.02	224
$log(Cl+3)$ ECN	1.386	1.792	1.914	1.888	1.946	2.890	0.149	280
$log(Cl+3)$ AWMN	1.609	1.841	1.887	1.911	1.946	2.868	0.132	224
$log(NO_3+0.5)$ ECN	-0.634	-0.607	-0.579	-0.555	-0.534	-0.085	0.083	146
$log(NO_3+0.5)$ AWMN	-0.657	-0.657	-0.657	-0.654	-0.657	-0.562	0.014	223
$log(Al+0.5)$ ECN	-0.687	-0.627	-0.588	-0.558	-0.519	0.282	0.101	297
$log(Al+0.5)$AWMN	-0.653	-0.653	-0.653	-0.600	-0.579	-0.040	0.099	223
$log(K+0.5)$ ECN	-0.691	-0.371	-0.301	-0.171	-0.162	2.170	0.414	285
$log(K+0.5)$ AWMN	-0.510	-0.356	-0.314	-0.313	-0.274	0.067	0.091	224
$log(SO_4(S)+1)$ ECN	0.267	0.499	0.528	0.528	0.550	1.511	0.084	268
$log(SO_4(S)+1)$ AWMN	0.154	0.490	0.510	0.510	0.550	0.773	0.072	224

Table 2.1. Descriptive Analysis for ECN and AWMN sites

Figures 2.4 and 2.5 show the monthly boxplots for each variable. There is more variability per month for the ECN compared to the AWMN, as a result of the collection process with more observations per month in the ECN data while only one observation per month in the AWMN data. According to the graphs, pH, $log(DOC)$, $log(Ca+2)$ and $log(Cl+3)$ show evidence of a seasonal component based on the monthly boxplots.

To identify any linear relationship and to evaluate the strength of correlation between the variables, Figure 2.6 shows a scatterplot between each pair of variables for ECN and AWMN separately. The upper panels display the correlation coefficients while the lower panels display a scatterplot. Based on the correlation coefficient there is evidence that the relationship between variables is not the same in both data sources. Specific examples are pH and $log(DOC)$ where the correlation coefficient is negative in both data sources, while for $log(DOC)$ and $log(Na+0.5)$ the correlation coefficient is positive for the ECN data and negative for the AWMN.

To explore the level of agreement between both sources of information, Figure 2.7 shows the Bland-Altman plots. The aim of this graph [Bland & Altman (1986)] is to assess the level of agreement between two methods that measure the same process, or in this case the same variable in two different locations.

From each of the data sets for each variable, the mean of the two measurements (x-axis) is plotted against the difference between both values (y-axis). Each point on the graphs corresponds to

$$\left(\frac{X_{ECN}+Y_{AWMN}}{2}, (X_{ECN}-Y_{AWMN})\right).$$

Upper and lower limits of agreement can be added. These are defined as $\bar{d}\pm 2sd(d)$,

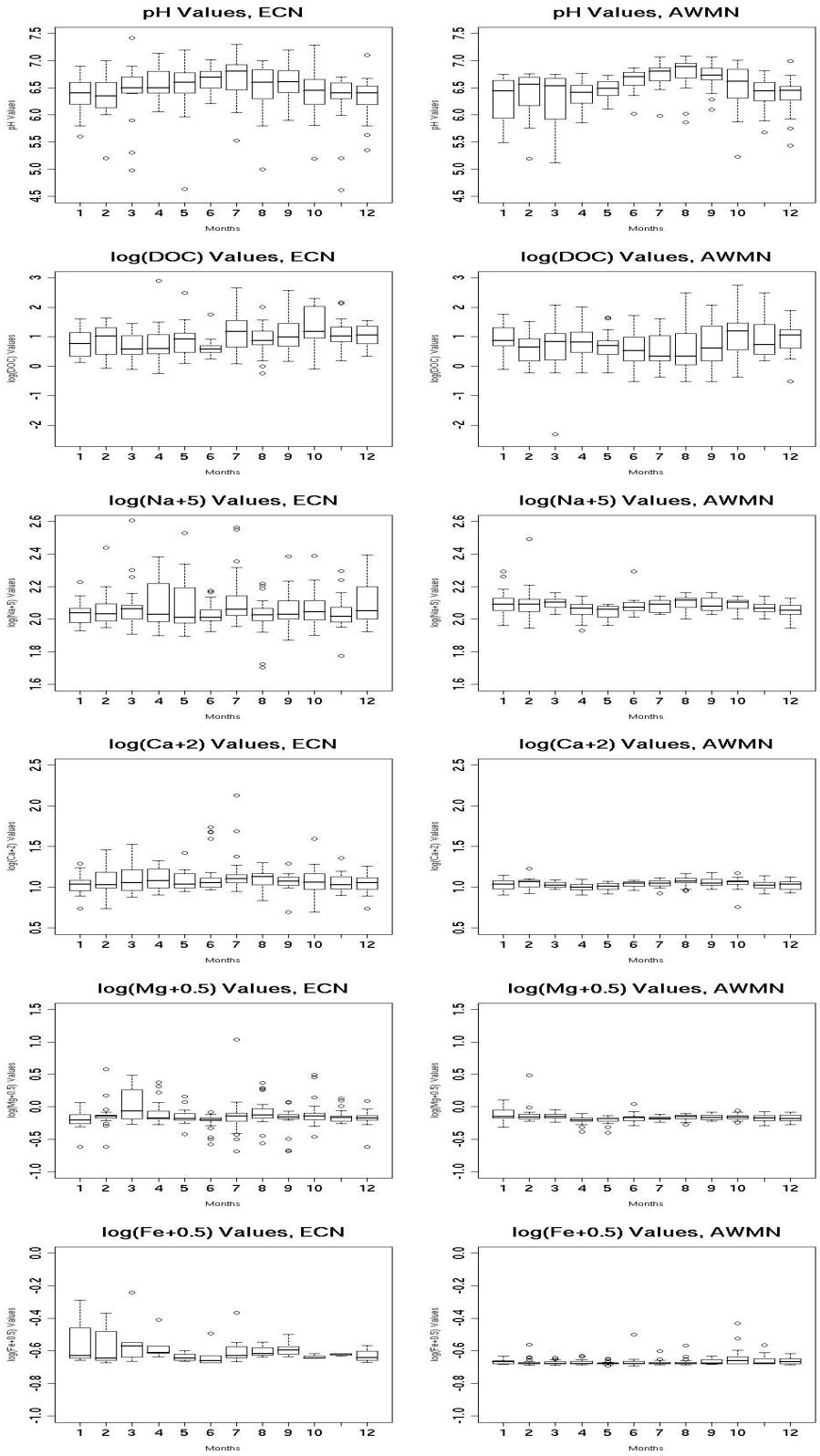

Figure 2.4. Monthly Boxplot for variables pH, $log(DOC)$, $log(Na+5)$, $log(Ca+2)$, $log(Mg+0.5)$ and $log(Fe+0.5)$ at ECN and AWMN sites

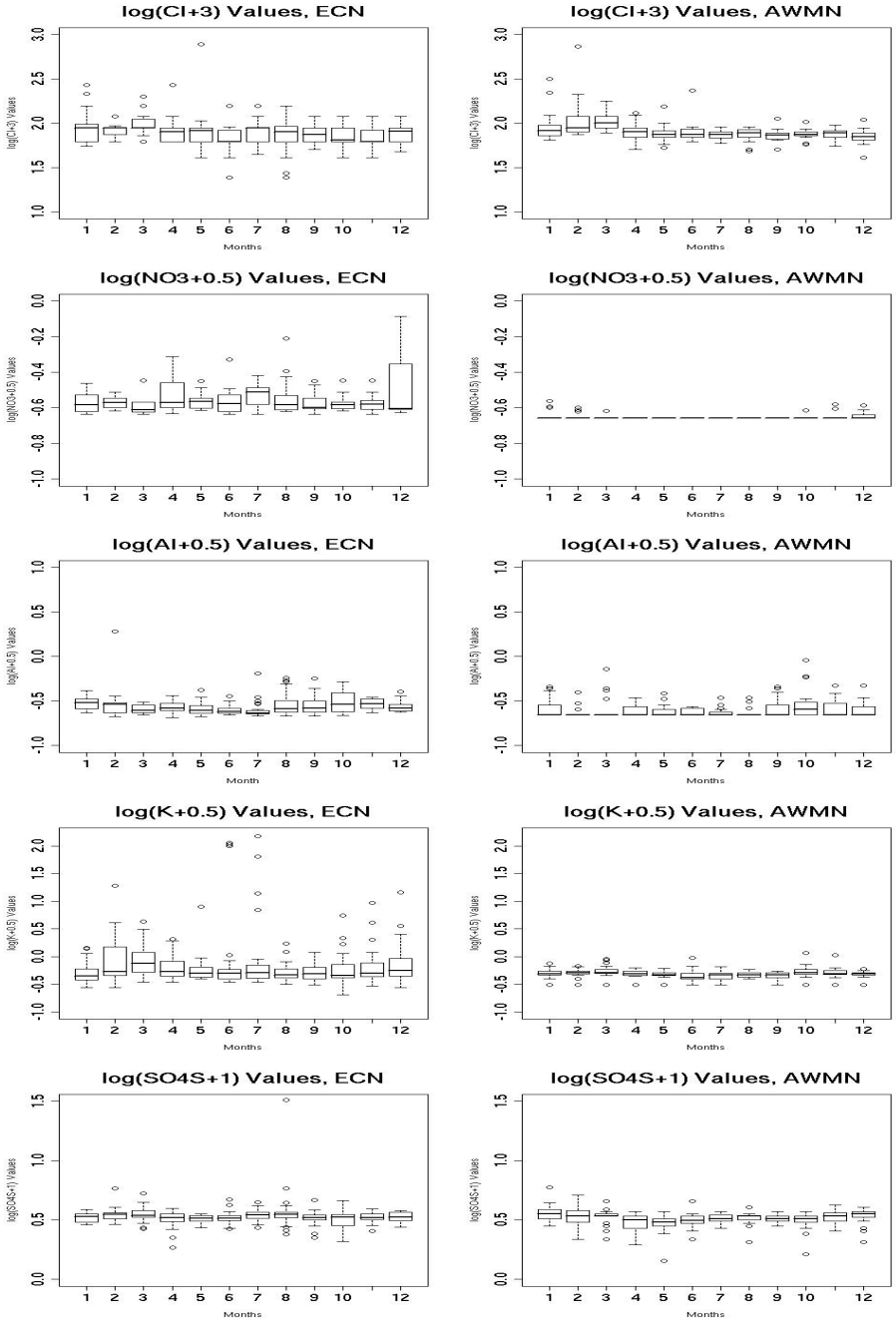

Figure 2.5. Monthly Boxplot for variables $log(Cl+3)$, $log(NO_3+0.5)$, $log(Al+0.5)$, $log(K+0.5)$ and $log(SO_4S+1)$ at ECN and AWMN sites

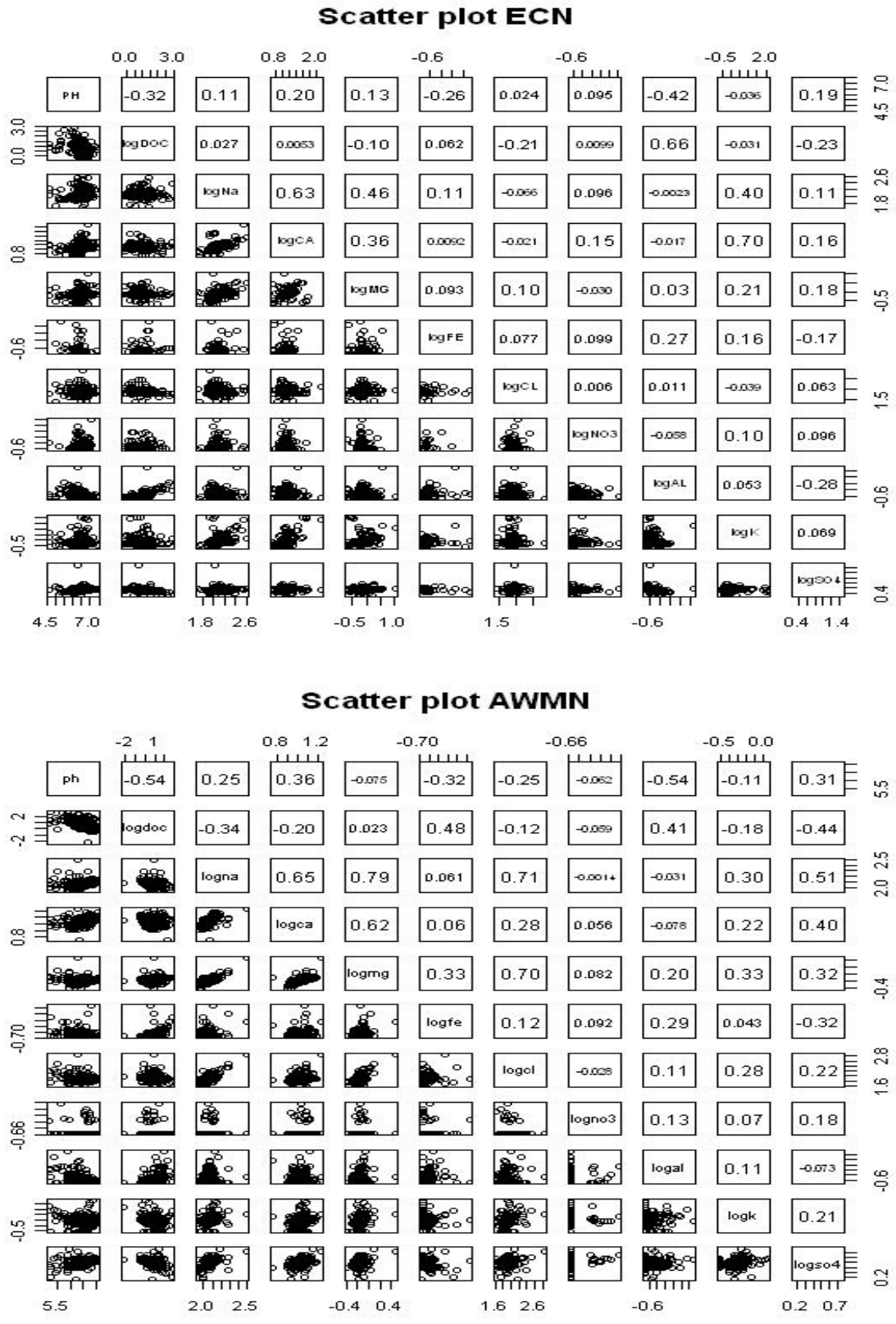

Figure 2.6. Scatterplot with Correlation Matrix for ECN and AWMN

under the assumption of normality, with d as the difference between the two measurement, $\bar{d}$ the sample average and $sd(d)$ the standard deviation.

According to Figure 2.7 only pH, $log(DOC)$ and $log(SO_4(S)+1)$ show a good level of agreement, with few points below and above the limits of agreement. $log(SO_4(S)+1)$ shows an outlier with a value in the ECN data which is three times the corresponding value in the AWMN data.

For the others variables, a clear positive relationship between the difference and the average can be observed, indicating that the measurement for ECN tends to provide higher values than for AWMN. On average, the values provided for the ECN are 25% higher for $log(NO_3+0.5)$, 16% for $log(Mg+0.5)$, 13% for $log(Fe+0.5)$, 11% for $log(Al+0.5)$, 10% $log(K+0.5)$ and 3% for $log(Ca+2)$.

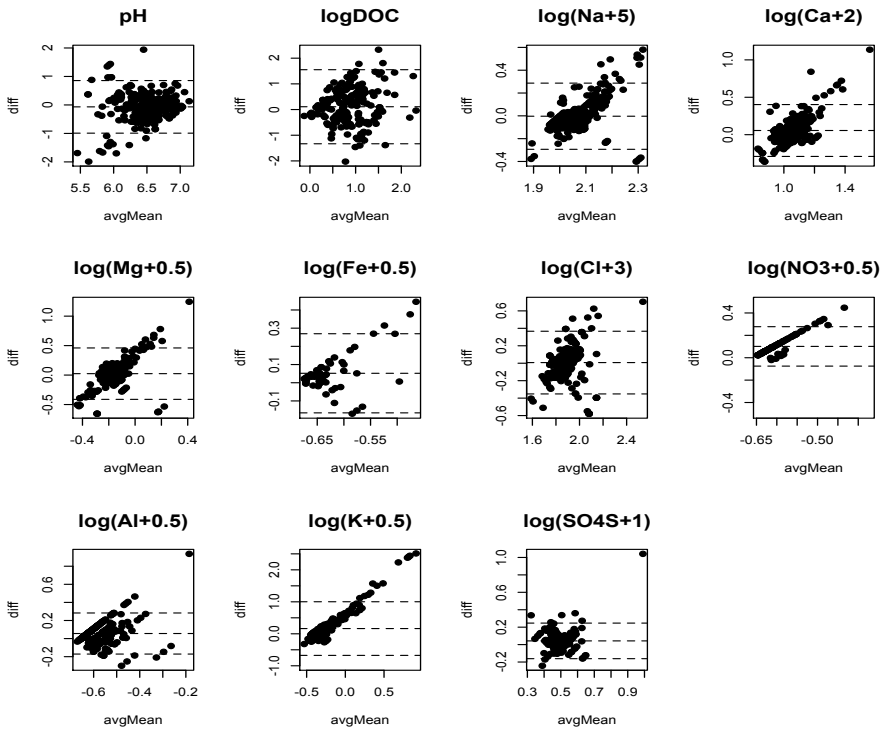

Figure 2.7. Bland-Altman plots, level of agreement between both sites

2.3 Linear Model

The main aim in this section is to identify the presence of trend and seasonality over time through a non-linear model to capture both components.

$$y = \beta_0 + \beta_1 year + \beta_2 cos\left(2\pi\left(\frac{days-\gamma}{366}\right)\right) + \varepsilon_i \quad i = 1, ..., n \tag{2.1}$$

Model (2.1) includes the year and the day, where the term $\beta_2 cos\left(2\pi\left(\frac{days-\gamma}{366}\right)\right)$ describes a seasonal component [Esterby et al. (1991)]. Using the trigonometric identity $cos(a-b) = cos(a)sin(b)+sin(a)cos(b)$, the seasonal term in model (2.1) can be expressed as

$$\beta_2 cos\left(2\pi\left(\frac{days-\gamma}{366}\right)\right) = \beta_2\left[cos\left(2\pi\left(\frac{days}{366}\right)\right)sin\left(2\pi\left(\frac{\gamma}{366}\right)\right) + sin\left(2\pi\left(\frac{days}{366}\right)\right)cos\left(2\pi\left(\frac{\gamma}{366}\right)\right)\right]$$

To estimate the parameter γ, model (2.1) can be re-expressed in a linear form as (2.2) where ε_i are assumed independent with mean 0 and constant variance σ^2.

$$y = \beta_0 + \beta_1 year + \beta_2' cos\left(2\pi\left(\frac{days}{366}\right)\right) + \beta_2'' sin\left(2\pi\left(\frac{days}{366}\right)\right) + \varepsilon_i \quad i = 1, ..., n \tag{2.2}$$

where β_2' and β_2'' are estimated as

$$\hat{\beta}_2' = \hat{\beta}_2 sin\left(2\pi\left(\frac{\hat{\gamma}}{366}\right)\right) \qquad \text{and} \qquad \hat{\beta}_2'' = \hat{\beta}_2 cos\left(2\pi\left(\frac{\hat{\gamma}}{366}\right)\right) \tag{2.3}$$

Model (2.2) can be fitted by ordinary least squares (OLS) providing an estimate for β_2' and β_2'', allowing an estimate of $\hat{\gamma}$ to be obtained by solving the equations in (2.3). Model (2.2) was fitted by OLS for all variables, although in those cases

where there was evidence of autocorrelation in the residuals using the Durbin-Watson statistic, these models were fitted by GLS including a continuous AR(1) for a continuous time covariate. This is a suitable approach when the errors are separated by a s unit of time, where the correlation between error is $\rho(s) = \phi^{|s|}$ with $0 \leq \phi \leq 1$. [Pinheiro & Bates (2000)].

Table 2.2 provides a summary of the trend and seasonal parameters with their corresponding p-values for non-linear model (2.1) under a null hypothesis that these parameters are statistically equal to zero for each variable in both data sets. For $log(Na + 5)$, $log(Ca + 2)$ and $log(K + 0.5)$ in the ECN data, model (2.1) was fitted with autocorrelated errors. The conclusion obtained does not change if uncorrelated errors are assumed, although the AIC shows a slight improvement.

According to the results observed in Table 2.2 is possible to identify if both variables exhibit similar behaviour with respect to the presence of a linear trend and/or seasonal component in both sources of information.

- Only $log(Ca + 2)$ and $log(Al + 0.5)$ show the same pattern in both sources of information, with the presence of a seasonal component and a downward trend.

- For $log(DOC)$ there is a seasonal component and trend in both sources of information but there is a downward trend for the ECN while for the AWMN there is an upward trend. However, the size of these parameters indicates that the trend is not strong.

- $log(Na+5)$ and $log(K+0.5)$ show in both sources of information downward trend but only $log(K + 0.5)$ has a seasonal component in the AWMN data.

- pH and $log(Cl + 3)$ show a seasonal component in both sources of information but only the ECN data has an upward trend in both variables.

- $log(Fe + 0.5)$ and $log(SO_4(S) + 1)$ show a seasonal component with an

upward and downward trend respectively for the AWMN data.

- $log(Mg+0.5)$ and $log(NO_3+0.5)$ show a seasonal component for the AWMN data.

Only $log(DOC)$, $log(Na+5)$, $log(Ca+2)$, $log(Al+0.5)$ and $log(K+0.5)$ show a significant parameter for trend in both data sets while pH, $log(Fe+0.5)$, $log(Cl+3)$ and $log(SO_4(S)+1)$ only in one of them. However the size for these parameters indicates that despite a statistically significant difference from zero there is not a strong trend in all the variables.

Figure 2.8 provides the same conclusions observed in Table 2.2, although it allows us to assess whether the parameters for linear trend or/and seasonal component are statistically equal in both sources of information, using a 95% confidence interval for the estimated parameters. The left hand side graph provides the C.I. for trend $\hat{\beta}_1$, while the right hand side provides the C.I. for seasonal component $\hat{\beta}_2$.

- Only for $log(Al+0.5)$ the parameters for trend are not statistically different in both sources of information.
- None of the parameters for trend or seasonality are statistically equal for the variable $log(DOC)$ since the confidence intervals do not overlap.
- The trend parameter for $log(Na+5)$ and $log(K+0.5)$ are statistically different in both sources of information.
- None of the parameters for seasonality are statistically equal for pH and $log(Cl+3)$ since the confidence intervals do not overlap.

Figures 2.9 and 2.10 show the residuals versus fitted values. The left panel corresponds to the ECN while the right panel shows the AWMN. These graphs allow assessment of whether the linear model works well for these variables.

PARAMETERS ESTIMATED				
Variable/Parameters	ECN	AWMN	p-value (ECN)	p-value(AWMN)
$pH\ \beta_1$	0.033	0.006	0.001	0.112
$pH\ \beta_2$	-0.162	-0.234	< 0.001	< 0.001
$log(DOC)\ \beta_1$	-0.090	0.041	< 0.001	< 0.001
$log(DOC)\ \beta_2$	0.159	0.188	0.007	0.002
$log(Na + 5)\ \beta_1$	-0.018	-0.001	< 0.001	0.019
$log(Na + 5)\ \beta_2$	0.012	0.004	0.383	0.421
$log(Ca + 2)\ \beta_1$	-0.025	-0.001	< 0.001	0.043
$log(Ca + 2)\ \beta_2$	-0.054	-0.017	0.005	0.001
$log(Mg + 0.5)\ \beta_1$	-0.001	-0.001	0.850	0.086
$log(Mg + 0.5)\ \beta_2$	0.019	0.019	0.214	0.009
$log(Fe + 0.5)\ \beta_1$	0.006	0.0009	0.358	0.008
$log(Fe + 0.5)\ \beta_2$	0.021	0.008	0.213	0.001
$log(Cl + 3)\ \beta_1$	0.011	-0.002	0.003	0.08
$log(Cl + 3)\ \beta_2$	0.045	0.069	< 0.001	< 0.001
$log(NO_3 + 0.5)\ \beta_1$	-0.004	-0.0002	0.157	0.174
$log(NO_3 + 0.5)\ \beta_2$	-0.007	0.006	0.464	< 0.001
$log(Al + 0.5)\ \beta_1$	-0.006	-0.002	0.025	0.018
$log(Al + 0.5)\ \beta_2$	0.034	0.025	< 0.001	< 0.001
$log(K + 0.5)\ \beta_1$	-0.060	-0.004	0.005	< 0.001
$log(K + 0.5)\ \beta_2$	-0.046	0.023	0.439	0.004
$log(SO_4(S) + 1)\ \beta_1$	-0.002	-0.004	0.273	< 0.001
$log(SO_4(S) + 1)\ \beta_2$	-0.003	0.023	0.679	< 0.001

Table 2.2. Parameter Estimates and p-values for Trend and Seasonal Component (SC) at ECN and AWMN sites under model (2.1)

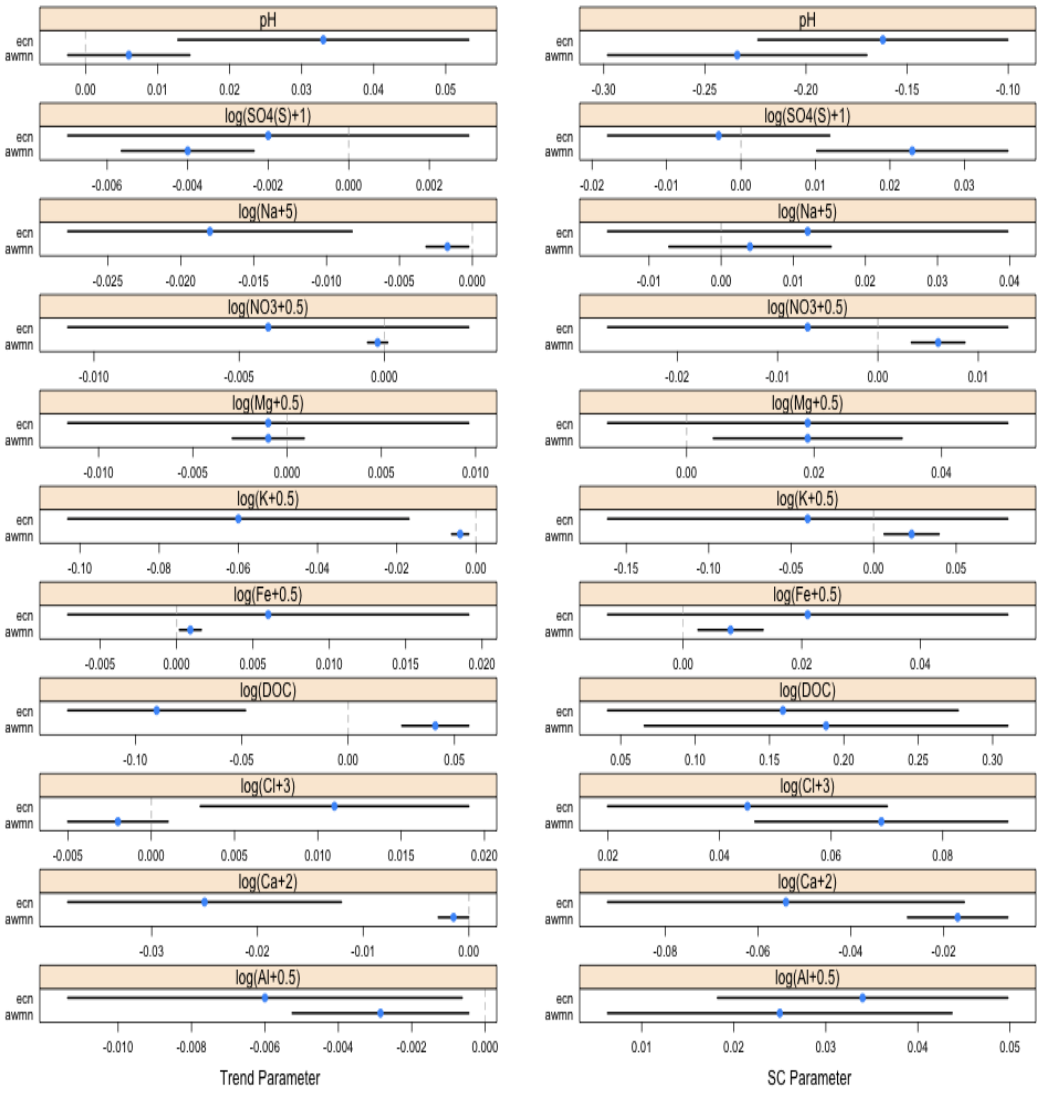

Figure 2.8. Confidence Intervals and estimated parameter for trend (left hand side) and seasonal component (right hand side) under model (2.1) at ECN and AWMN sites

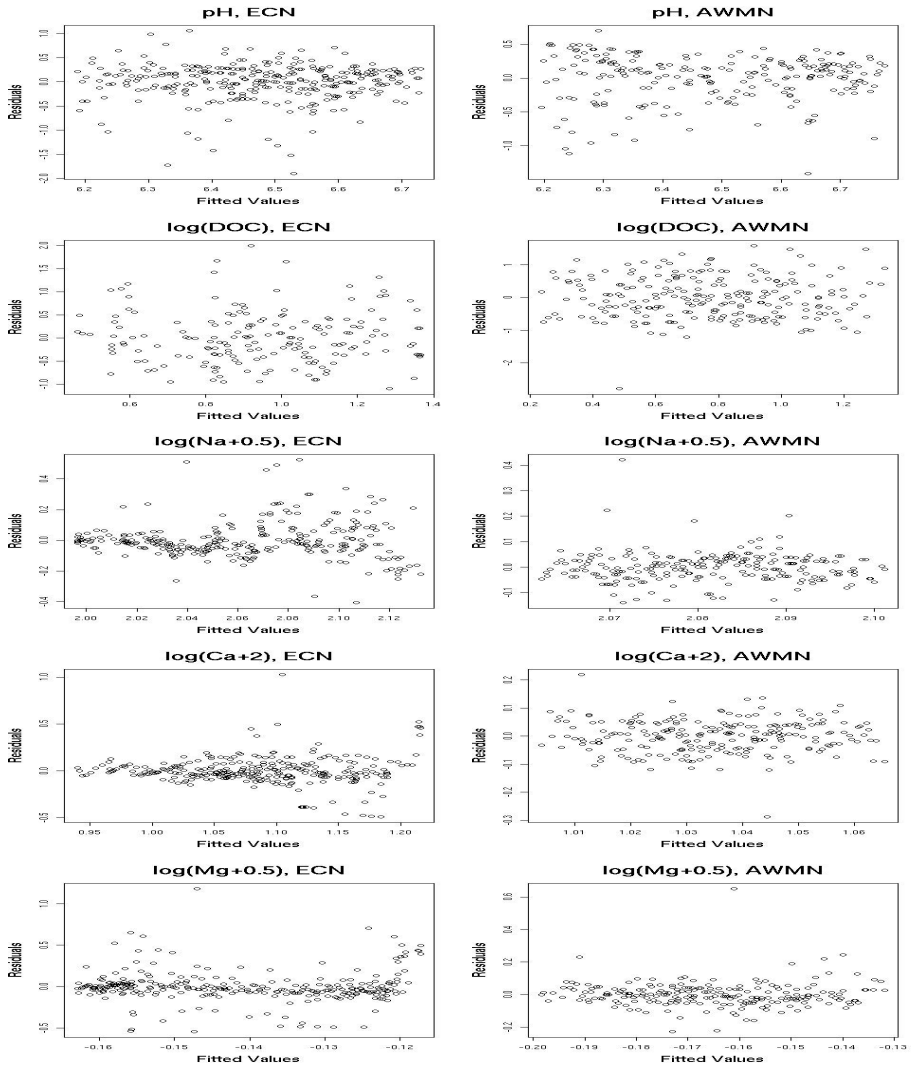

Figure 2.9. Residuals versus Fitted Values for variables pH, $log(DOC)$, $log(Na + 5)$, $log(Ca + 2)$ and $log(Mg + 0.5)$ at ECN and AWMN sites under model (2.1)

The patterns observed for $log(Fe + 0.5)$, $log(NO_3 + 0.5)$ and $log(Al + 0.5)$ in the AWMN data correspond to limits of detection values with a frequency of 98, 208 and 144 times respectively. For $log(Cl + 3)$ in the ECN data, corresponds to values 1.79 and 1.94 with a frequency of 82 and 74 times respectively.

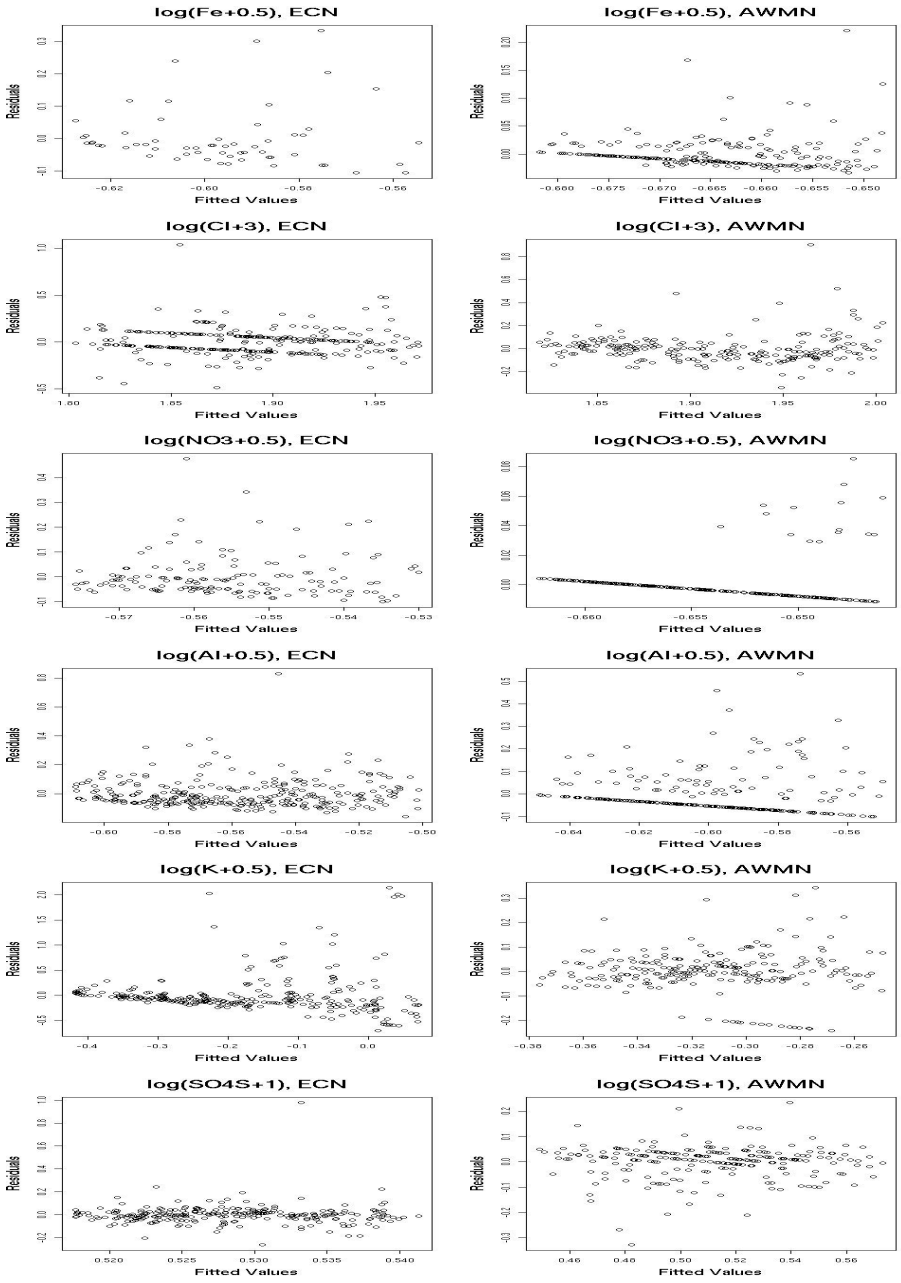

Figure 2.10. Residuals versus Fitted Values for variables $log(Fe+0.5)$, $log(Cl+3)$, $log(NO_3+0.5)$, $log(Al+0.5)$, $log(K+0.5)$ and $log(SO_4S+1)$ at ECN and AWMN sites under model (2.1)

The results observed indicate that a linear approach works well for this information; however the question is also raised of whether a nonparametric regression would be preferable. The interest here is to evaluate the need for a nonparametric effect, over three relevant variables related to environment pollution ($log(DOC)$, $log(NO_3 + 0.5)$ and $log(SO_4(S) + 1)$) to assess if there is an improvement in the model or to confirm the use of a linear approach.

2.4 Nonparametric Regression

When the data observed are not easily described by a linear model, a suitable approach is to fit a nonparametric regression model of the form

$$y_i = m(x_i) + \varepsilon_i \qquad i = 1, ..., n \tag{2.4}$$

where $m(x_i)$ corresponds to a smooth function, $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$.

There are different ways to obtain an estimate for $\hat{m}(x)$, one such approach is to use kernel estimators. Some of the most common are kernel smoothers, local regression, smoothing splines, regression splines, orthogonal series and wavelets [Green & Silverman (1994), Wood (2006) Fan & Gijbels (1996)].

Throughout this thesis a kernel smoother and local regression approach are introduced in detail, where the similarities with standard linear models lead to many useful statistical properties.

An estimate for $\hat{m}(x)$ can be obtained by a local mean estimator [Nadaraya (1964*b*), Watson (1964)] as

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} w(x_i - x; h) y_i}{\sum_{i=1}^{n} w(x_i - x; h)}, \tag{2.5}$$

where $w(x_i - x; h)$, the weight function chosen, corresponds to a normal density centred on zero with standard deviation equal to h [Bowman & Azzalini (1997)]. The solution for expression (2.5) arises through the process of minimising the weighted least squares over α.

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

In the case of cyclical variables or seasonal effects, quite common in environmental information, an estimate for $\hat{m}(x)$ can be obtain using a local mean approach, where the weight function chosen corresponds to $w(x_i - x; h) = exp\left[\frac{r}{h} cos\left(\frac{2\pi(x_i - x)}{r}\right)\right]$, allowing us to obtain an estimate with period r [Bowman et al. (2009)].

The selection of h, called the smoothing parameter or bandwidth is the key step here to establish the influence of the data points on the estimate. As the value of h is increased, the number of observations that contribute to the estimate increases, reducing the flexibility of the estimate. The opposite effect is observed when small values of h are chosen, increasing the flexibility of the estimate thus reproducing the data more closely.

An alternative approach to obtain an estimate is local linear regression, where an estimate for $\hat{m}(x)$ can be obtained by weighted least squares minimising the expression

$$min_{\alpha,\beta} \sum_{i=1}^{n} [y_i - \alpha - \beta(x_i - x)]^2 w(x_i - x; h). \tag{2.7}$$

The weight function $w(x_i - x; h)$, as in local mean estimation, corresponds to a normal density centred on zero with standard deviation equal to h.

The solution for (2.7) corresponds to $\hat{m}(x) = v^T y$ where the ith element of v

can be express as

$$v_i = \frac{1}{n}\frac{\Big[s_2(x;h) - s_1(x;h)(x_i - x)\Big]w(x_i - x;h)}{s_2(x;h)s_0(x;h) - s_1(x;h)^2},$$

where $s_r(x;h) = \Big[\sum (x_i - x)^r w(x_i - x;h)\Big]/n$

Despite that the local linear regression is prefer over the local mean because its superior properties [Fan & Gijbels (1996)], the local mean provides a better approach in the case of seasonal effect, thereby throughout this thesis the kernel estimator chosen corresponds to a local mean estimator.

The estimate $\hat{m}(x)$ at a set of values x for these two estimators can be defined as Sy, where S corresponds to the smoothing matrix whose rows contain the vectors v to obtain the estimate at a particular value of x. This provides an important result, indicating that the estimation process is linear in the response data y.

Based on this result it is possible to define a relationship between the smoothing parameter and the degrees of freedom of a nonparametric model. The degrees of freedom under a linear model can be defined as the trace (tr) of the projection matrix P, where $\hat{y} = Py$. In the case of a nonparametric model, the degrees of freedom can be defined as $df = tr(S)$, with S the smoothing matrix.

2.5 Smoothing parameter selection

The selection of the value for the smoothing parameter h, or degrees of freedom df, is still a matter of discussion, although it is clear that we can not simply choose the value to minimise the residual sum of squares, as this would lead to an overfitted solution.

In this section the idea is to provide an outline of a method called *cross-validation*,

choosing a value of degrees of freedom to minimise $\sum_{i=1}^{n}(y_i - \hat{m}_{-i}(x_i))^2$ under the idea of use a training set and testing set. Additional methods with detailed explanations can be found in Hastie & Tibshirani (1990) and Bowman & Azzalini (1997).

Leaving a point (y_i, x_i) out, the idea is to estimate the smooth curve at x_i based on the $(n-1)$ remaining points, evaluating how accurate the prediction is under different values of degrees of freedom. This procedure is replicated for each x_i $i = 1, ..., n$, looking for a value of degrees of freedom which minimise $\sum_{i=1}^{n}(y_i - \hat{m}_{-i}(x_i))^2$ to obtain an adequate value.

This method provides a method of obtaining an optimal value for the degrees of freedom. However, *cross-validation* and other automatic selection methods, tend to be less reliable and far more expensive to implement for additive models [Hastie & Tibshirani (1990)] which are introduced in the next section, mainly because several degrees of freedom must be selected simultaneously. In addition, these methods are not suitable for information collected over time and/or space with a temporal [Hart (1991)] or spatial correlation structure.

The selection of degrees of freedom throughout this thesis is performed by a subjective method, as the main objective is to assess different models to capture trends over time and space rather than choose a model based on automatic methods. In addition, the assessment of the partial residuals to evaluate the effect of each variable allows us to explore whether the degrees of freedom chosen is capturing well the relationship between the covariates and the dependent variable.

2.6 Additive Models

Additive models developed by Hastie and Tibshirani (1990) follow the same characteristics as linear models, although they are written as

$$E(Y|x_1, x_2, ..., x_p) = m_1(x_1) + m_2(x_2) + ... + m_p(x_p) + \varepsilon_i \qquad i = 1, ..., n \quad (2.8)$$

where $m_j(x_j)$ $j = 1, ..., p$ corresponds to a smooth function that describes the effect of covariate j on Y and $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$. This has the advantage that each of these smooth functions is not restricted in shape, which means that even if the relationship between our dependent and independent variables is non-linear, the smooth function is able to capture this relationship.

The smooth functions are obtained by use of the backfitting algorithm which corresponds to an iterative process that follows three steps [Hastie & Tibshirani (1990)].

1. Initialize: $m_j = m_j^{(0)}, j = 1, ..., p$
2. Cycle:$j = 1, ..., p, 1...p, ...$ $m_j = S_j(y - \sum_{k \neq j} m_j | x_j)$
3. Continue (2) until the individual functions do not change

Once the algorithm converges, model (2.8) is written as

$$E(Y|x_1, x_2, ..., x_p) = \hat{\beta}_0 + \hat{m}_1(x_1) + \hat{m}_2(x_2) + ... + \hat{m}_p(x_p), \quad (2.9)$$

where β_0 is $\bar{y}$ and $\hat{m}_j(x_j)$ correspond to numerical vectors corresponding to the smooth functions. It is important to recall that $S_j(y|x_j)$ denotes a smoothing matrix for the response y against the predictor x_j.

2.7 Comparison of Models

Once a model has been fitted, a natural step is to assess if there is any possibility to improve that model. Following the idea developed by Hastie and Tibshirani (1990), this comparison is possible through an approximate F-test.

This test statistic does not follow the exact F distribution, although results based on simulations [Hastie & Tibshirani (1990)] provide enough evidence to support it as a guide to choose between different models. The approximate F-test is defined as

$$\frac{(RSS_1 - RSS_2)/(df_2 - df_1)}{RSS_2/(n - df_2)} \sim F_{df_2-df_1,n-df_2},$$

where RSS_1 and RSS_2 are the residual sum of squares and df_1 and df_2 are the degrees of freedom of the models fitted.

Having fitted an additive model the RSS is defined as $RSS = \sum_{i=1}^{n}(y_i - \hat{m}(x_i))^2$ or as a quadratic form as $RSS = y^tQy$ where $Q = (I - P)^t(I - P)$. Each of the smooth functions can be expressed as a set of n x n projection matrices, providing the fitted values for an additive model as $Py = (\sum_{k=0}^{p} P_k)y$, where P_0 corresponds to a matrix with the value $1/n$ to estimate $\bar{y}$.

In the same way as in a linear model, it is possible to obtain an analogous definition of approximate degrees of freedom for an additive model, where the approximate degrees of freedom for error can be defined as $df = tr[(I-P)^t(I-P)]$, with $P = \sum_{k=0}^{p} P_k$. [Bowman & Azzalini (1997)].

In the case of correlated data, the main effect of correlation is in the calculation of standard error and in the implementation of model comparison [Giannitrapani et al. (2005)]. A modification of the RSS through the generalised least squares

criterion allows this structure to be included as

$$RSS = y^t(I - P)^t V^{-1}(I - P)y,$$

where V corresponds to the estimate of the correlation matrix. In the same way, the degrees of freedom can be defined as $df = tr[(I - P)^t V^{-1}(I - P)]$, allowing a comparison to be made between different models using an approximate F-test [McMullan et al. (2007)].

2.8 Testing for No Effect and Sensitivity Analysis

Using the approximate F-test, the main aim in this section is to test a null hypothesis that a linear effect is adequate compared to a nonparametric effect, so assessing whether a nonparametric effect is required. The sensitivity analysis also allows us to assess possible changes in the conclusions under different values of degrees of freedom.

Table 2.3 provides the p-values to assess the need for a nonparametric effect for the three variables under 6 degrees of freedom, indicating that a nonparametric effect is required for day and year for $log(SO_4(S) + 1)$.

Tables 2.4 to 2.6 show the results under different degrees of freedom, indicating that the results are stable. According to these results there is no doubt that a linear approach works well for $log(DOC)$ and $log(NO_3 + 0.5)$. The opposite conclusion is observed in the variable $log(SO_4(S) + 1)$ where a nonparametric effect for year and day is required.

Figure 2.11 depicts the estimate for year and day for $log(SO_4(S) + 1)$, the upper panel displays the ECN, while the lower panel displays the AWMN. In each

	$log(DOC)$		$log(NO_3 + 0.5)$		$log(SO_4(S)) + 1)$	
Parameter	ECN	AWMN	ECN	AWMN	ECN	AWMN
year df(6)	0.120	0.211	0.409	0.063	0.002	0.001
day df(6)	0.804	0.290	0.167	0.496	0.005	0.010

Table 2.3. p-values for the test of the need for a nonparametric effect opposed to linear effect at ECN and AWMN sites

p values							
ECN	df=4	df=6	df=8	AWMN	df=4	df=6	df=8
year	0.066	0.120	0.162	year	0.248	0.211	0.246
day	0.871	0.804	0.751	day	0.247	0.290	0.453

Table 2.4. p-values sensitivity analysis for variable $log(DOC)$ to assess stability under different degrees of freedom for year and day

p values							
ECN	df=4	df=6	df=8	AWMN	df=4	df=6	df=8
year	0.512	0.409	0.272	year	0.047	0.063	0.069
day	0.220	0.167	0.110	day	0.429	0.496	0.733

Table 2.5. p-values sensitivity analysis for variable $log(NO_3 + 0.5)$ to assess stability under different degrees of freedom for year and day

p values							
ECN	df=4	df=6	df=8	AWMN	df=4	df=6	df=8
year	0.002	0.002	0.003	year	0.012	0.001	0.001
day	0.006	0.005	0.007	day	0.010	0.010	0.027

Table 2.6. p-values sensitivity analysis for variable $log(SO_4(S) + 1)$ to assess stability under different degrees of freedom for year and day

graph the points corresponds to partial residuals, the solid line corresponds to the smooth function fitted and the dashed line corresponds to $\pm$ 2 standard error band [Bowman & Young (1996)]. The outliers in the ECN corresponds to 1.5 which is also observed in the time series graphs.

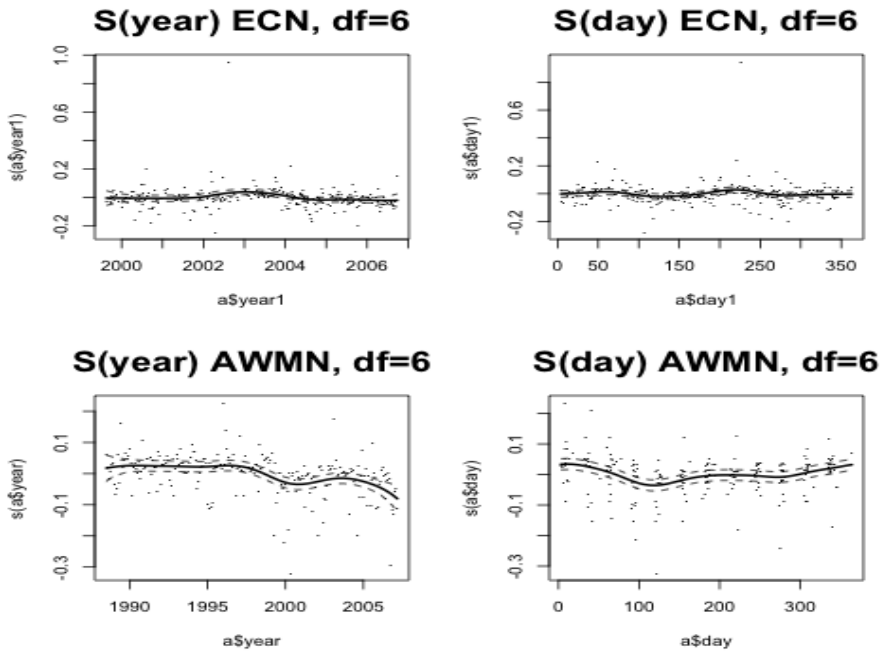

Figure 2.11. partial residuals (points), fitted smooth function (solid line) and $\pm$ 2 standard error band (dashed line) for additive model variable $log(SO_4(S) + 1)$ at ECN and AWMN sites

2.9 Summary

The descriptive analysis indicates that 75% of the values for all the variables lie in the same range in both data sets, although a higher variability for the ECN data is observed in the descriptive analysis as well as in the time series graphs.

The analysis of the behaviour of the variables over the year through monthly boxplots to identify seasonal patterns, the relationship between variables through the scatterplot matrix and the Bland-Altman plots, indicates that

- there are variables with a seasonal pattern such as pH, $log(DOC)$, $log(Ca+2)$ and $log(Cl+3)$ in both data sets
- the linear relation between variables is different in both data sets.
- only pH, $log(DOC)$ and $log(SO_4(S)+1)$ show a good level of agreement while for the other variables the ECN data provide higher values in average than the AWMN data.

Regarding the scientific questions formulated at the beginning of this chapter the analysis of the parameters in Table 2.2 allows the presence of linear trend and seasonal components over time to be identified. In addition, it also allows us to establish differences between the two data sets.

With respect to the assessment of trend over time, variables such as $log(Na+5)$, $log(Ca+2)$, $log(Al+0.5)$ and $log(K+0.5)$ show a downward trend over both data sets, $log(DOC)$ shows an upward trend for the AWMN data while the ECN data show a downward trend. Variables such as pH and $log(Cl+3)$ exhibit an upward trend in the ECN data while $log(Fe+0.5)$ and $log(SO_4(S)+1)$ exhibit an upward and downward trend respectively in the AWMN data. Only $log(Mg+0.5)$ and $log(NO_3+0.5)$ did not show a trend either in the ECN or in the AWMN data.

Looking to identify similar behaviour in both data sets, Table 2.2 allows us to establish that only $log(Ca+2)$ and $log(Al+0.5)$ show the same behaviour with a downward trend and seasonal component in both data sets while $log(Na+5)$ shows a downward trend but not a seasonal component in both data sets.

For the other 8 variables the results indicate different behaviour, showing a seasonal component in both data sets but only an upward trend in the ECN data (pH and $log(Cl+3)$), trend and seasonal components but only for the AWMN data ($log(Fe+0.5)$ and $log(SO_4(S)+1)$), trend in both data sets but only a seasonal component for the AWMN data ($log(K+0.5)$), seasonal components in

both data sets but an opposite trends, upward trend for the ECN and downward trend for the AWMN ($log(DOC)$) and only a seasonal component for the AWMN data ($log(NO_3 + 0.5)$ and $log(Mg + 0.5)$).

The differences between the two locations could be explained by several possible reasons. It is clear that there are differences in the collection process for both sources of information, as well as differences between locations, indicating that the physical characteristics of the two catchments are different. The fact that the information from the ECN data provides higher values than the AWMN, allows us to confirm that the level of agreement is not good for the large majority of the variables in the two catchments. Only pH, $log(DOC)$ and $log(SO_4(S) + 1)$ showed a good level of agreement.

The introduction of additive models provides an opportunity to work with environmental data using nonparametric regression. The variables chosen indicate that a linear approach was suitable, although for $log(SO_4(S) + 1)$ an additive model using a smooth function for year and day is required.

In the next chapter, the aim is to explore in greater detail the use of these models over environmental data when information over time and space is collected simultaneously.

Chapter 3

Catchment Modelling

The main objective of this chapter is to evaluate the quality of the water in the Tarland catchment located in the north of Scotland (Figure 3.1). The information provided by the Macaulay Institute corresponds to nitrates and phosphates measured in 6 variables (NH4.N, Total N, NO3.N, PO4.P, Total P and Suspended Solids) collected regularly at 17 sites from April 2004 to November 2008. Site 33 does not have information for the period between April 2007 and November 2008 and the information for suspended solids is missing between 19th of February 2007 and the 4th of July 2007.

This chapter provides a descriptive analysis of the data and a simple model including time (trend and seasonal components) and space as covariates. Initially, standard spatial methods based on Euclidean distance have been used.

Given that the location for each site was not provided by the Macaulay Institute, coordinates in kilometres were calculated to be able to use the location of each site. According to the new coordinates the maximum distance is 7 kilometres corresponding to site 8 and 1 which are the most separated sites located in the map.

3.1 Descriptive Analysis

The first step was to evaluate the need for a transformation of the scale for the 6 variables and to identify possible outliers. For NH4.N, Suspended Solids (SuSo), PO4.P and Total P, a $log(x)$ transformation was applied, while for NO3.N and Total N the transformation applied was $log(x + c)$ with $c = 1$.

Figure 3.1 shows a map with the locations of the 17 sites. Figures 3.2 to 3.7 show the time series by site for each of the variables; these shows a diminution in the sampling frequency since 2006.

Figure 3.8 shows the boxplot by site for each of the six variables.

- For $log(NH4.N)$, sites 1, 27, 30 and 31 shows the highest values.
- For $log(NO3.N + 1)$ and $log(TotalN + 1)$, sites 6, 13 and 16 shows the highest values while sites 7, 8 and 30 shows the lowest values.
- For $log(PO4.P)$ and $log(TotalP)$, sites 1,13, 14 and 27 shows the highest values while sites 8 and 10 shows the lowest.
- For $log(SuSo)$, sites 14, 20 and 30 shows the highest values while site 13 show the lowest value.

3.2 Model for time and space effects

3.2.1 Linear Model

The data provided by the Macaulay Institute corresponds to observations collected over time and space. One possible way of analysing these data is to treat the information in a marginal way, separating time and space. A second possibility, which is the main aim of this chapter, is to fit a model including both time and space simultaneously, although this means that a proper covariance structure

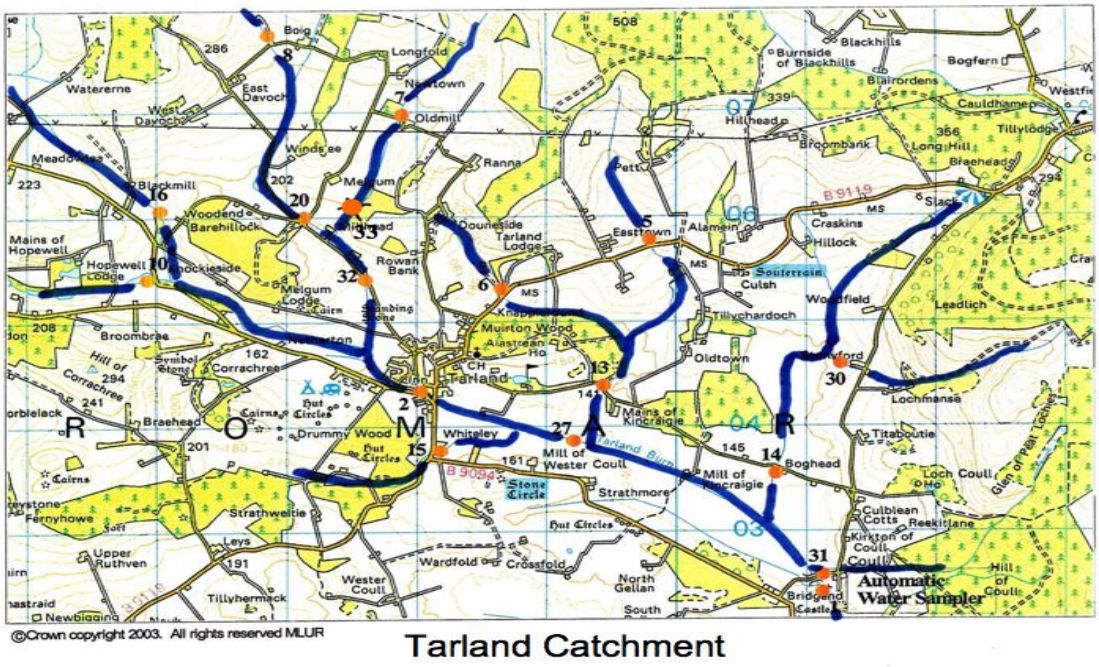

Figure 3.1. Location for the 17 sites in the Tarland Catchment

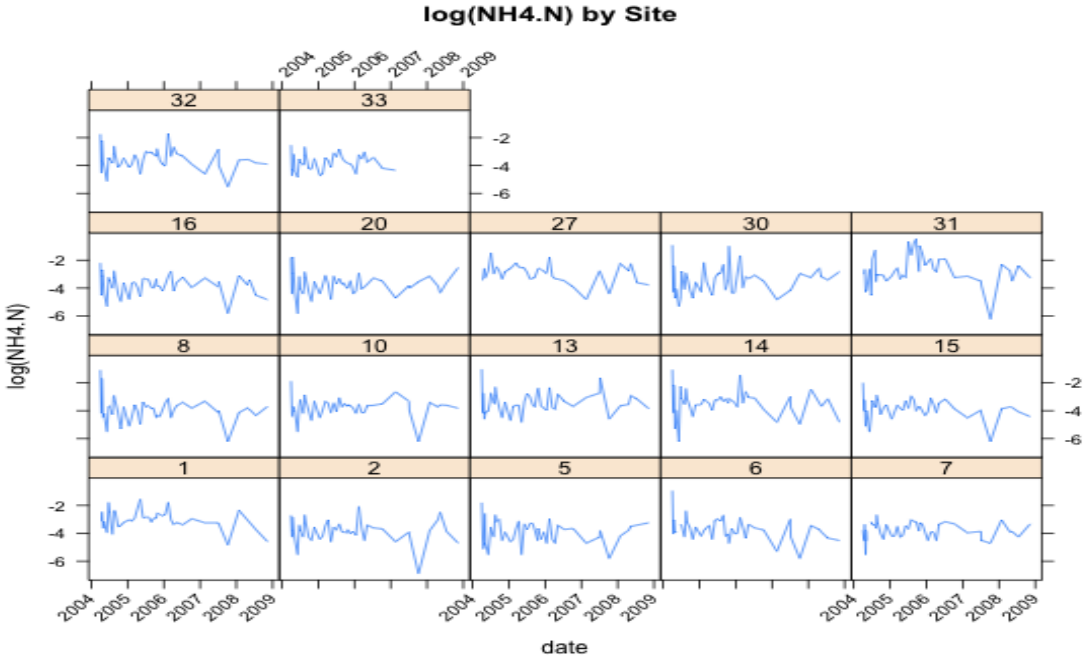

Figure 3.2. Time Series for $log(NH4.N)$ by Site

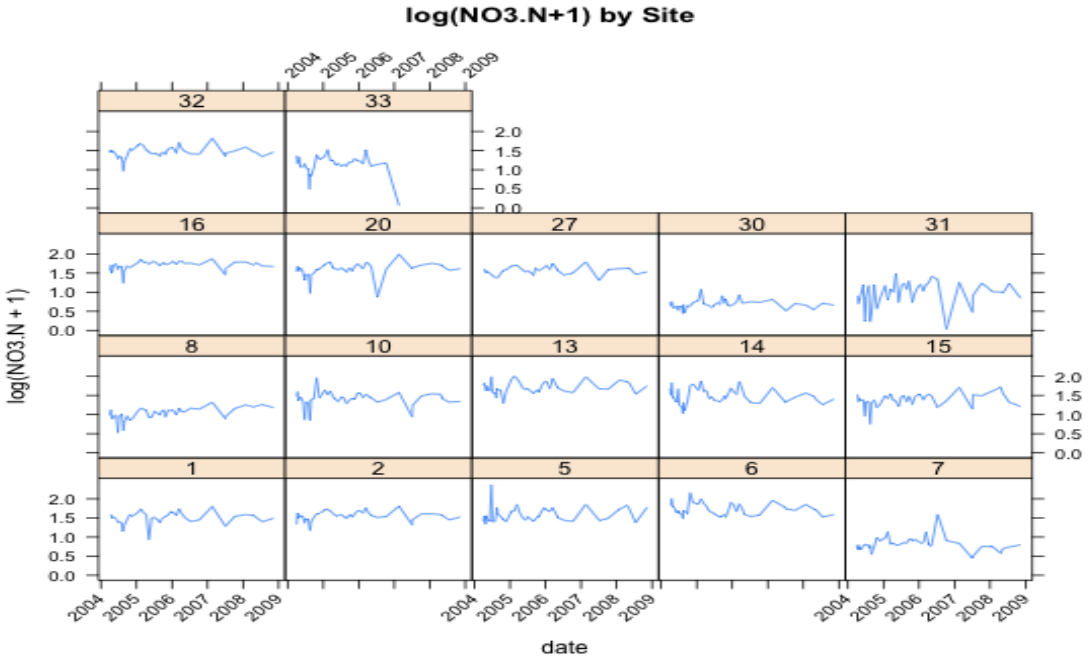

Figure 3.3. Time Series for $log(NO3.N + 1)$ by Site

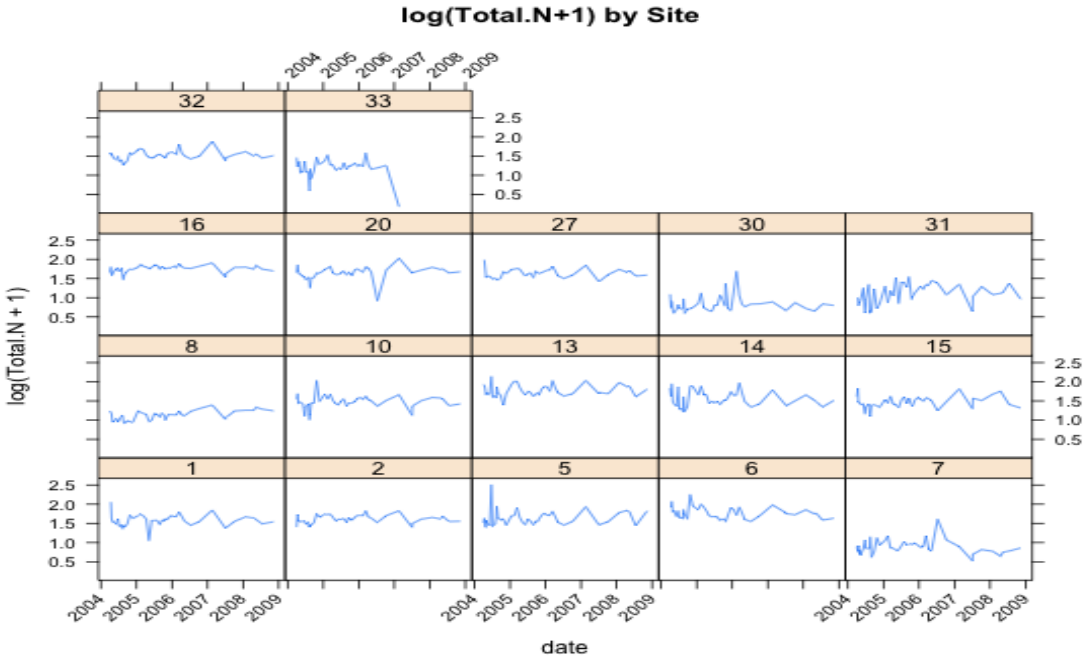

Figure 3.4. Time Series for $log(Total.N + 1)$ by Site

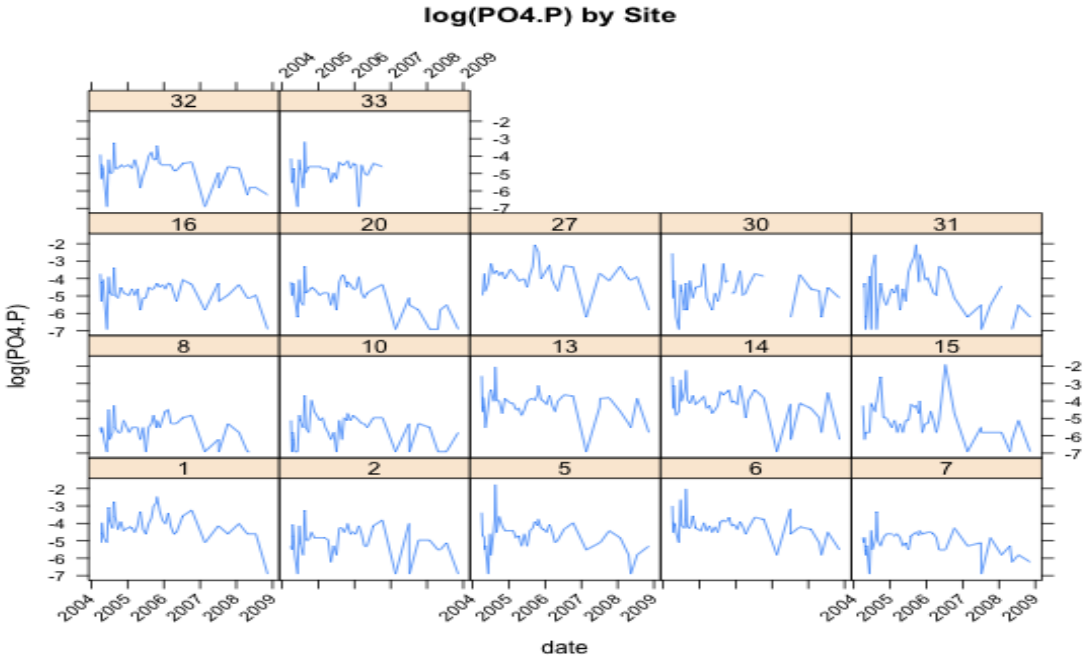

Figure 3.5. Time Series for $log(PO4.P)$ by Site

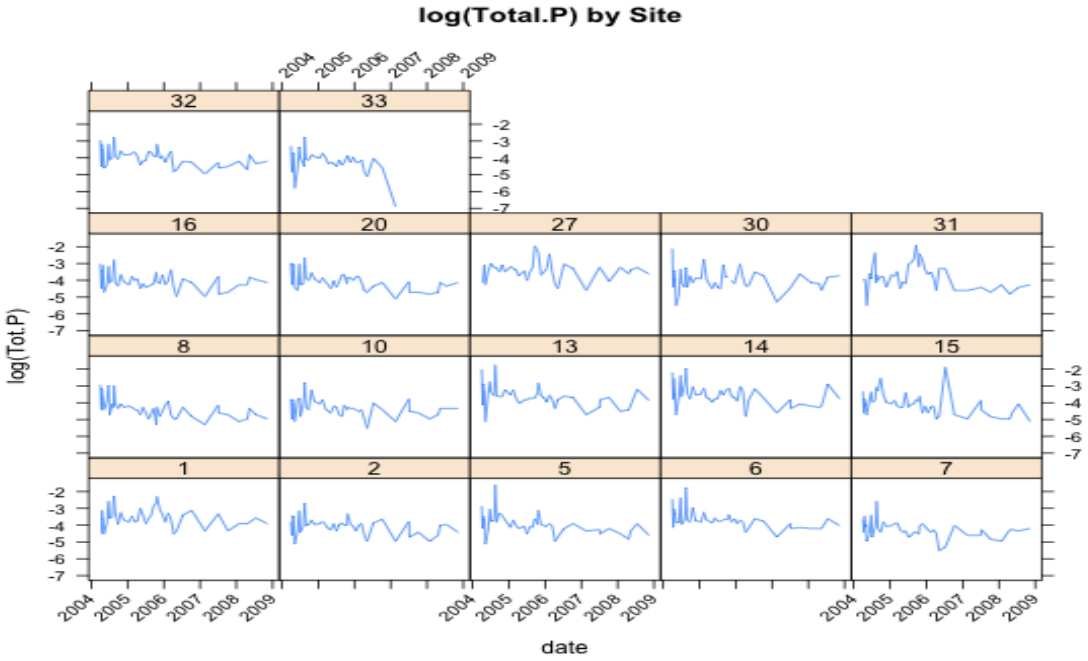

Figure 3.6. Time Series for $log(TotalP)$ by Site

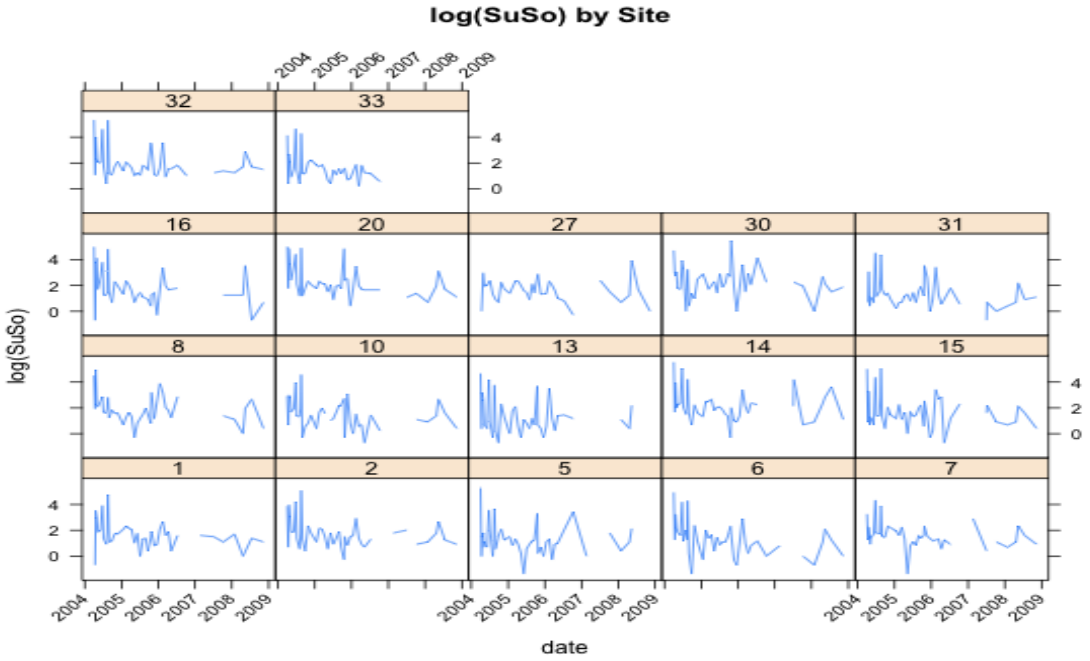

Figure 3.7. Time Series for $log(SuSo)$ by Site

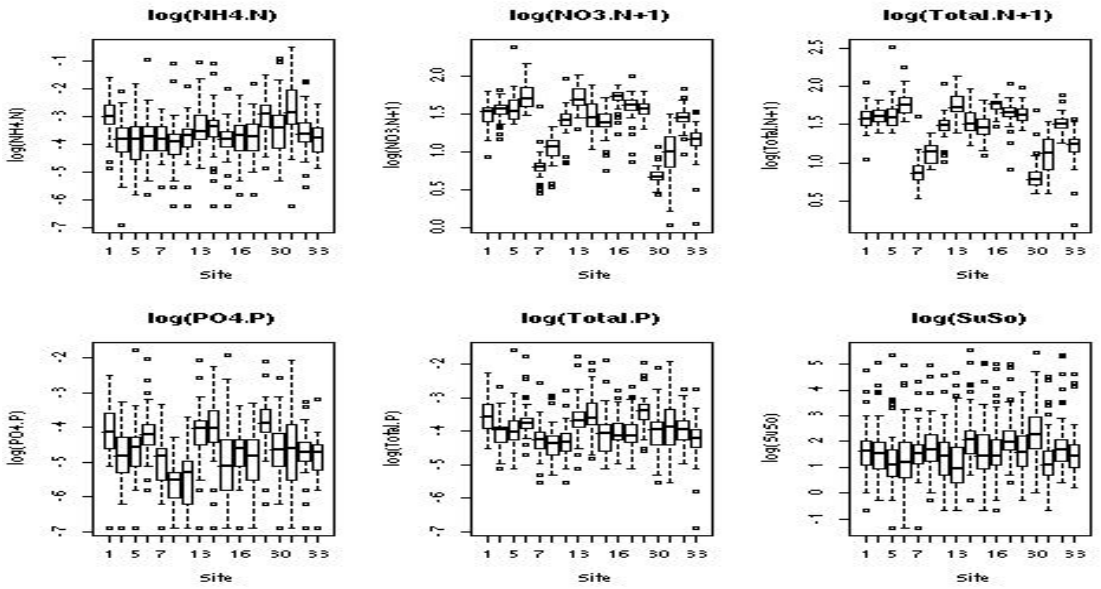

Figure 3.8. Boxplot for variables $log(NH4.N)$, $log(NO3.N+1)$, $log(TotalN+1)$, $log(PO4.P)$, $log(TotalP)$ and $log(SuSo)$ by site

for both time and space must be included if required.

A very simple approach is available in the model below, which fits a linear effect over year, a sinusoidal seasonal effect plus a linear spatial trend, where X (North-South) and Y (West-East) corresponds to the coordinates of the 17 sites. This model is fitted under the assumption that ε_i are independent with mean 0 and constant variance σ^2.

$$y = \beta_0 + \beta_1 year + \beta_2 cos\left(2\pi\left(\frac{days}{366}\right)\right) + \beta_3 sin\left(2\pi\left(\frac{days}{366}\right)\right) + \beta_4 X + \beta_5 Y + \varepsilon_i \quad i = 1, ..., n \tag{3.1}$$

Table 3.1 shows the estimated parameters for the six variables under model (3.1). For $log(NH4.N)$, $log(PO4.P)$, $log(TotalP)$ and $log(SuSo)$ there is a trend over time, although according to the size of this parameter it is not strong. All the variables show a seasonal pattern that is captured by the sine and cosine terms. The coordinates X and Y indicate higher values for $log(NH4.N)$, $log(PO4.P)$ and $log(TotalP)$ in the direction of X while the opposite behaviour is observed in the direction of Y. For $log(NO3.N + 1)$ and $log(TotalN + 1)$, X and Y indicate lower values in both directions while for $log(SuSo)$ neither parameter is significant.

Figure 3.9 indicates that a linear approach may be not adequate for variables such as $log(NO3N + 1)$, $log(TotalN + 1)$ and $log(SuSo)$ where the plots of the residuals against fitted values show a systematic pattern.

For variables such as $log(NH4N)$, $log(TotalP)$ and $log(PO4.P)$ a linear approach seems to work well, although the idea in the next section is to test it is suitability by comparing it with a nonparametric effect.

Linear Model Parameter						
	$log(NH4.N)$		$log(NO3.N+1)$		$log(TotalN+1)$	
Parameter	Estimate	p-value	Estimate	p-value	Estimate	p-value
year	-0.049	0.031	0.016	0.067	0.010	0.222
cos	-0.086	0.055	0.107	<0.001	0.092	<0.001
sin	0.055	0.169	0.075	<0.001	0.055	<0.001
X	0.059	0.023	-0.123	<0.001	-0.113	<0.001
Y	-0.142	<0.001	-0.135	<0.001	-0.135	<0.001
	$log(PO4.P)$		$log(TotalP+1)$		$log(SuSo)$	
Parameter	Estimate	p-value	Estimate	p-value	Estimate	p-value
year	-0.216	<0.001	-0.168	<0.001	-0.203	<0.001
cos	0.133	0.004	0.001	0.994	-0.153	0.019
sin	-0.364	<0.001	-0.205	<0.001	0.148	0.010
X	0.139	<0.001	0.062	<0.001	0.018	0.626
Y	-0.093	0.003	-0.096	<0.001	0.028	0.521

Table 3.1. Parameters for linear model for all variables

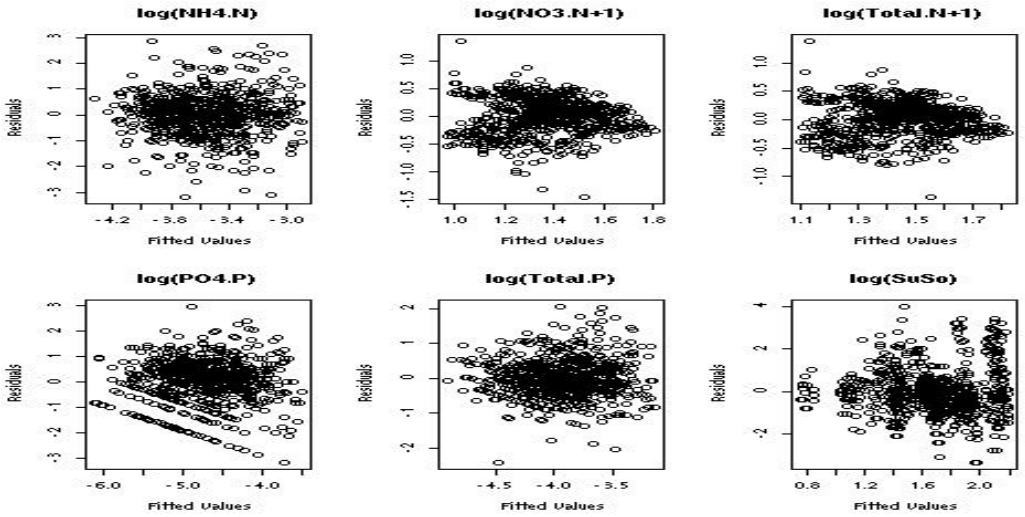

Figure 3.9. Residuals versus fitted values for all the variables under a linear model

3.2.2 Use of Additive models in the Tarland Catchment

The results observed in the previous section indicate that a linear model as a first approach was not adequate to capture the trend over time and space. In this section the aim is to use additive models [Hastie & Tibshirani (1990)] to capture in a better way the trend over time and space. The additive model is fitted under the assumption that the errors are independent, although later in this chapter the need to include a covariance structure over time and/or space is assessed in the residuals.

In this case the model to be fitted is model (3.2), where the $m_j(x_j)$ $j = 1, ..., p$, correspond to smooth functions estimated in a nonparametric manner and the ε_i are assumed to be independent with mean 0 and constant variance σ^2.

$$y = \beta_0 + m_1(year) + m_2(days) + m_3(X, Y) + \varepsilon_i \qquad i = 1, ..., n \tag{3.2}$$

For model (3.2), each of the $m_j(x_j)$ $j = 1, ..., p$ is fitted through the backfitting algorithm while β_0 is estimated by $\bar{y}$.

To illustrate how each function is estimated, we rewrite model (3.2) as $y - \beta_0 - m_1(year) - m_2(days) = m_3(X, Y) + \varepsilon_i$, obtaining $\hat{m}_3(X, Y)$ by smoothing the residuals of the model after fitting $\hat{m}_1(year)$ and $\hat{m}_2(days)$, i.e. $\hat{m}_3(X, Y) = S(y - \bar{y} - \hat{m}_1(year) - \hat{m}_2(days))$. In the same way, we can obtain $\hat{m}_1(year) = S(y - \bar{y} - \hat{m}_2(days) - \hat{m}_3(X, Y))$ and $\hat{m}_2(days) = S(y - \bar{y} - \hat{m}_1(year) - \hat{m}_3(X, Y))$. The algorithm is an iterative process which terminates after the individual functions do not change appreciably.

Figure 3.10 depicts the graphs for each of the variables over the 17 sites. These correspond to the plot of the average of each site over time. The colours over the maps indicates how the variables change over space.

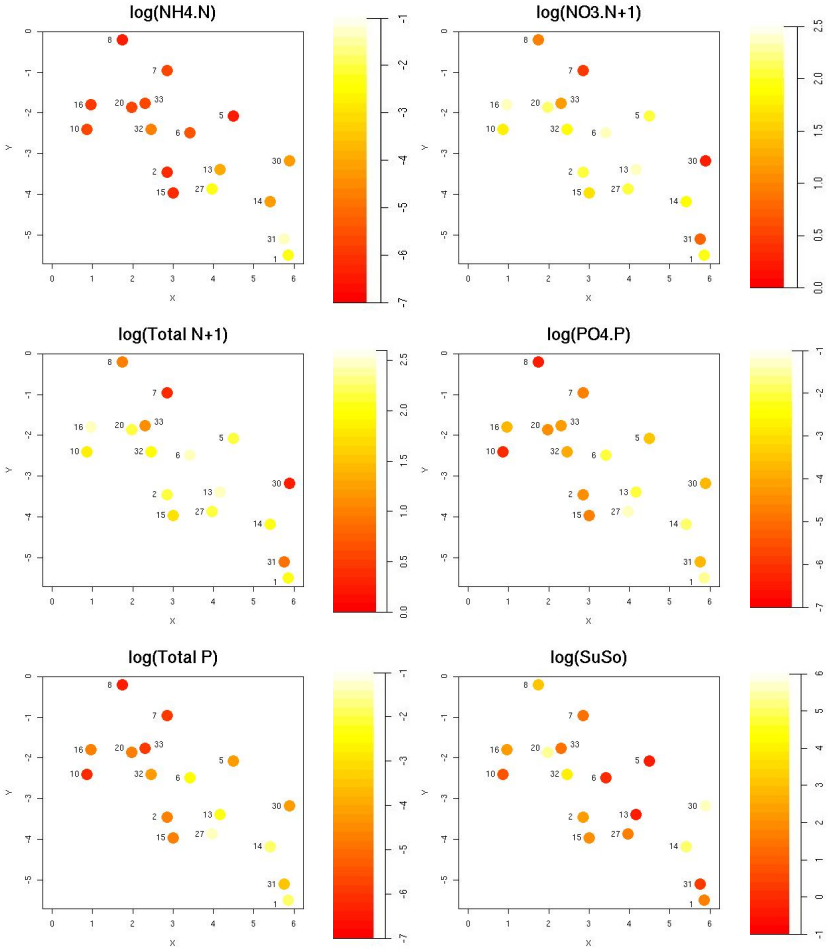

Figure 3.10. Distribution over space for the average at each site over time

Figures 3.11 to 3.16 show the fitted components of an additive model over the six variables. Each term shows the relationship between the dependent variable and the covariates. The solid line corresponds to the smooth function fitted, the dashed line corresponds to a $\pm$ 2 standard error band and the surface corresponds to a smoothing function in two dimensions to capture the trend for each variable over space; the last corresponds to $\hat{m}_3(X, Y)$ in model (3.2).

The degrees of freedom value chosen for each single covariate was 6, while a

value of 12 was chosen to obtain a smooth function over two covariates simultaneously, in this case the location of each site. This provides enough flexibility beyond a linear shape while the relatively modest value used, ensures that we capture large scale trend rather than small scale fluctuations.

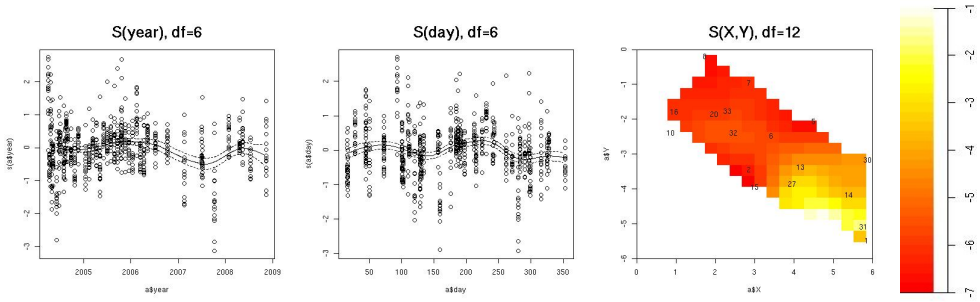

Figure 3.11. Plot of the components of additive models for $log(NH4.N)$

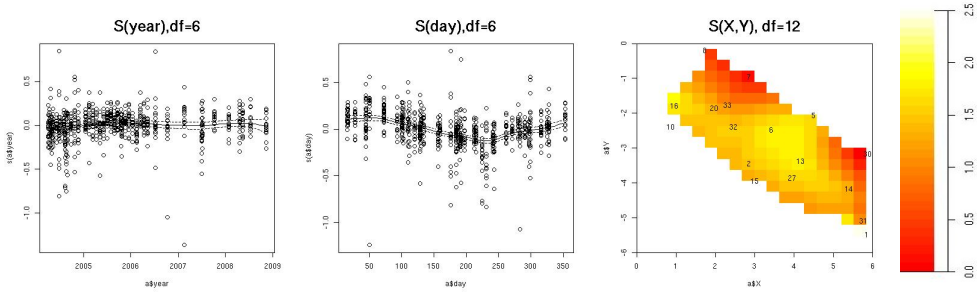

Figure 3.12. Plot of the components of additive models for $log(NO3.N + 1)$

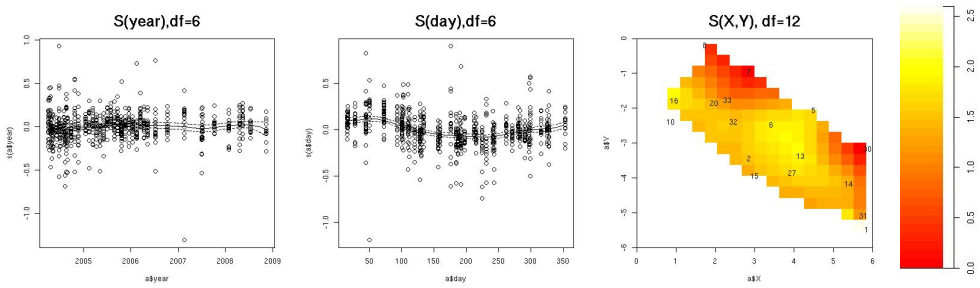

Figure 3.13. Plot of the components of additive models for $log(TotalN + 1)$

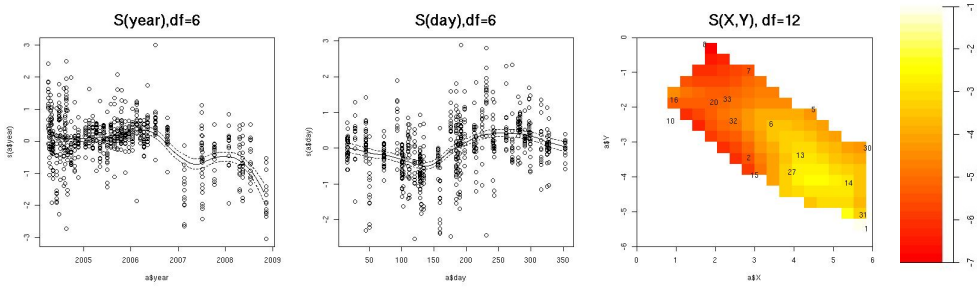

Figure 3.14. Plot of the components of additive models for $log(PO4.P)$

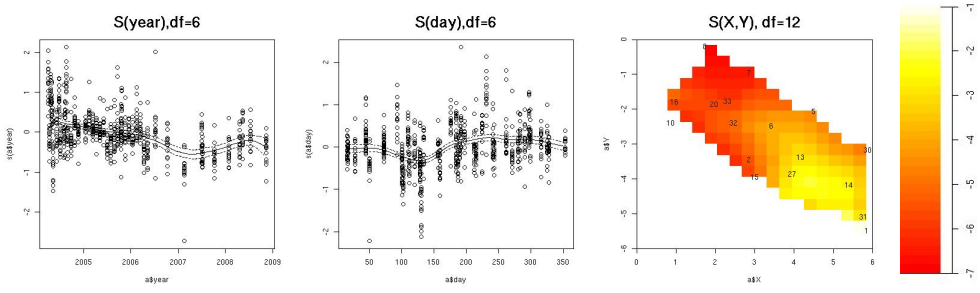

Figure 3.15. Plot of the components of additive models for $log(TotalP)$

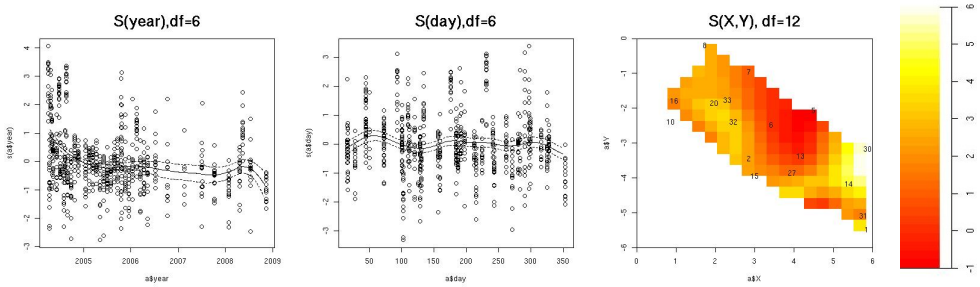

Figure 3.16. Plot of the components of additive models for $log(SuSo)$

3.3 Diagnostic Check

Figure 3.17 shows the graphs of residuals versus fitted values for all variables indicating that the additive models fit well, with only $log(PO4.P)$ showing unusual behaviour. Likely due to observations of limit of detection values.

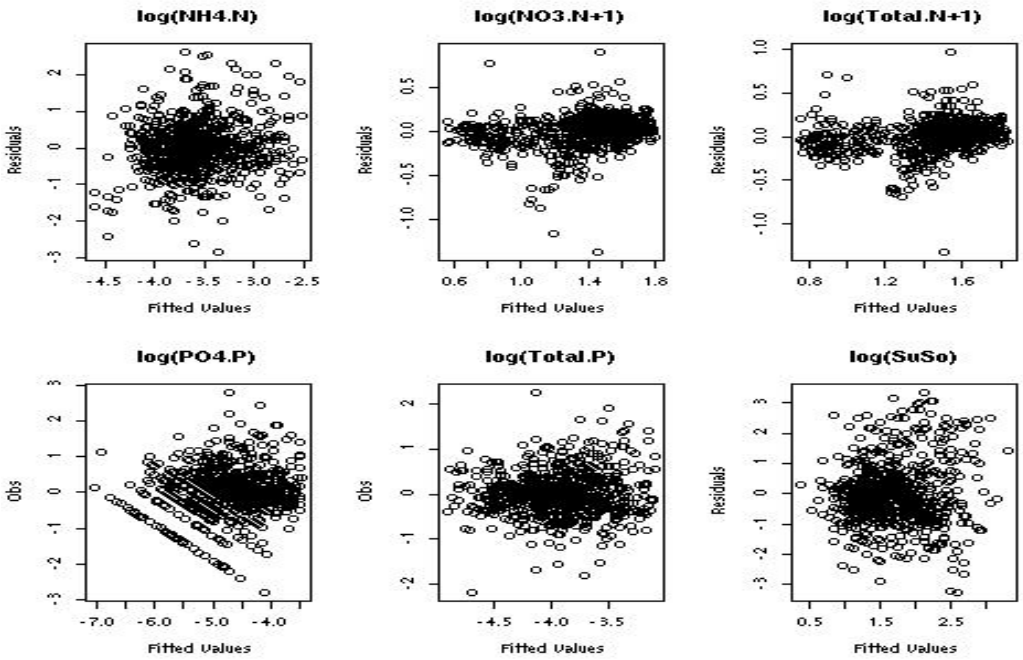

Figure 3.17. Residuals versus fitted values for all the variables under an additive model

Since the data are not equally spaced, it is not possible to use an autocorrelation

function to evaluate the presence of autocorrelation over time. One possibility explained by Diblasi and Bowman (2001) is to build a variogram for the residuals. This test was originally developed to evaluate indepedence over space for a single sample but it is also useful as a diagnostic check for regression models by examining the residuals.

The test evaluates the evidence that the empirical variogram changes as a function of the distance h (in this case h stands for distance between location rather than smoothing parameter), using $\hat{\gamma}(h) = \frac{1}{2}\frac{1}{|N(h)|}\sum_{N(h)}|Y(s_i) - Y(s_j)|^{\frac{1}{2}}$ as an estimator, where $N(h)$ denotes the collection of pairs of observations separated by a distance h. Independence over time or space is reflected in a constant variogram of the form $\gamma(h) = \sigma^2$, where $\gamma(h)$, the theoretical variogram, is a function that describes the degree of dependence in two dimensions of a set of observations collected over space or over time in one dimension.

Under the assumption that the distribution does not change over space, known as stationarity and uniformity in all direction, known as isotropy [Cressie (1993)], a model for the data can be expressed as

$$Y(s) = \mu + \varepsilon(s),$$

where $\varepsilon(s)$ are assumed to be independent with variance equal to $\gamma(h)$. To assess if the variogram changes as a function of the distance, the idea is to use nonparametric regression models, using a linear approach to provides an smooth function from the differences pairs $(|s_i - s_j|, |Y(s_i) - Y(s_j)|^{\frac{1}{2}})$, denoted by (h_{ij}, d_{ij}), where $i < j$. The estimate of the variogram is defined as $\hat{\gamma}(h) = \sum_{i<j} w_{ij}d_{ij}$, where w_{ij} the weights are derived from a nonparametric regression.

This approach corresponds to a special case for checking a linearity assumption in regression models [Azzalini & Bowman (1993)], allowing us to establish a

similar idea; if the errors are independent the variogram $\gamma(h)$ is constant, otherwise any significant evidence provided by the nonparametric regression, indicates the presence of spatial correlation. Diblasi and Bowman (2001), provide a test to assess the presence of spatial correlation, allowing us to obtain a p-value under a null hypothesis that $\gamma(h) = \sigma^2$.

The inclusion of references bands [Bowman & Young (1996)], provides a graphical tool to assess changes in the variogram as a function of the distance, where these bands display pointwise standard errors at each value of h, as $2[\widehat{var}(\hat{\gamma}(h) - \bar{d})]^{\frac{1}{2}}$, where $\bar{d}$ is the mean value of d_{ij}.

To be able to construct the variogram to test autocorrelation over time, the date when the data was collected was taken in Julian format as a distance variable. Figure 3.18 depicts the variogram for each variable with their respective p-values under a hypothesis of independence over time, indicating no evidence of autocorrelation over time for all six variables.

Following the same idea Figure 3.19 depicts the variograms for each of the six variables with their corresponding p-values, under a hypothesis of independence over space. It is important to highlight here that the reason why the test developed by Diblasi and Bowman (2001) can be applied is that there is no evidence of autocorrelation over time. According to the results observed in the variograms, only $log(TotalP)$ show evidence of spatial correlation with a p-value=0.038, although it is not a strong evidence in accordance with the output for the other variables.

To confirm the previous statement Figure 3.20 shows the variogram proposed by

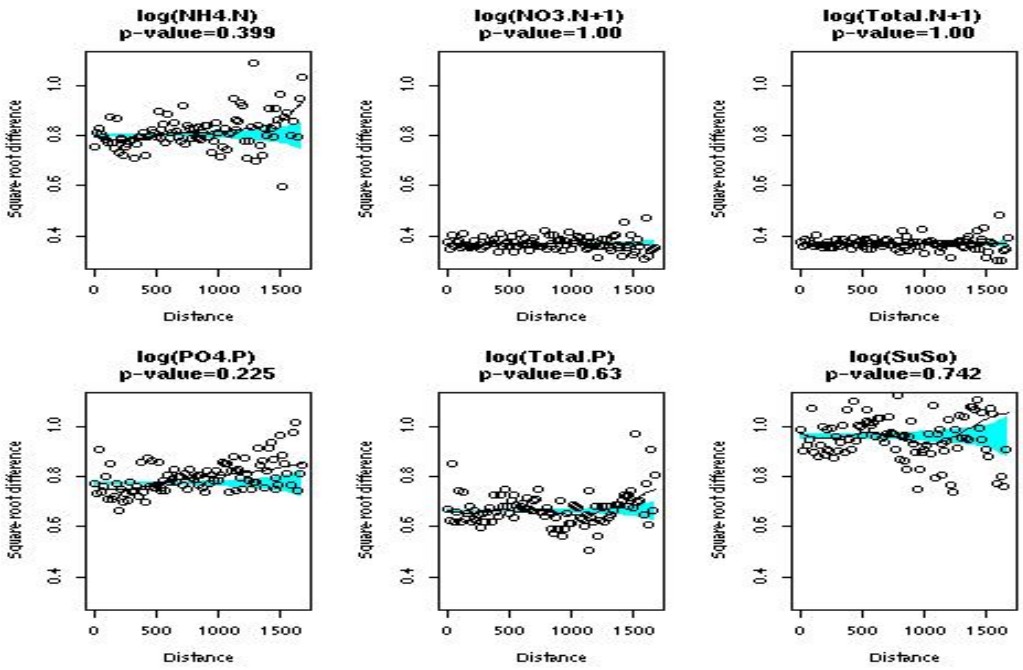

Figure 3.18. Independence test over time for residuals under an additive model

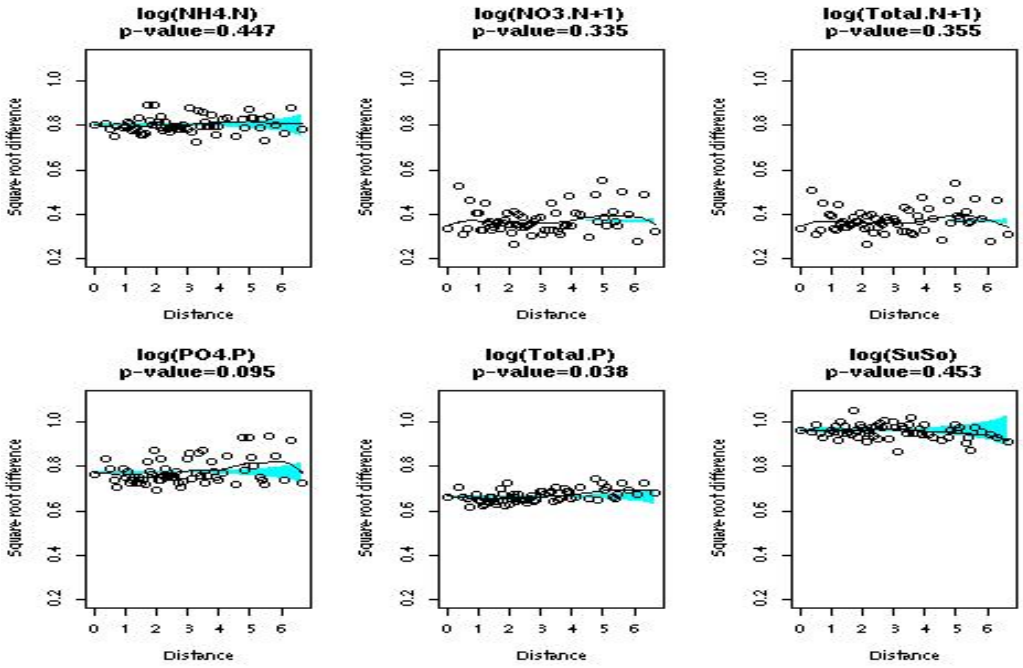

Figure 3.19. Independence test over space for residuals under an additive model

Cressie and Hawkins (1980) fitted for all variables.

$$\gamma(h) = \frac{\left(\frac{1}{N_h}\sum_{i=1}^{N_h}|Y(s_{i+h}) - Y(s_i)|^{1/2}]^4\right)}{\left(0.914 + \frac{0.988}{N_h}\right)}$$

Only $log(TotalP)$ shows a shape that might fit with some spatial model [Webster & Oliver (2007)] while for the rest of the variables it confirms the results observed in figure 3.19 indicating no correlation over space.

Despite the fact that the evidence for correlation over space is not strong for the variable $log(TotalP)$, it is possible to assess further whether a spatial model could be fitted. Table 3.2 depicts the results for different models (Exponential,

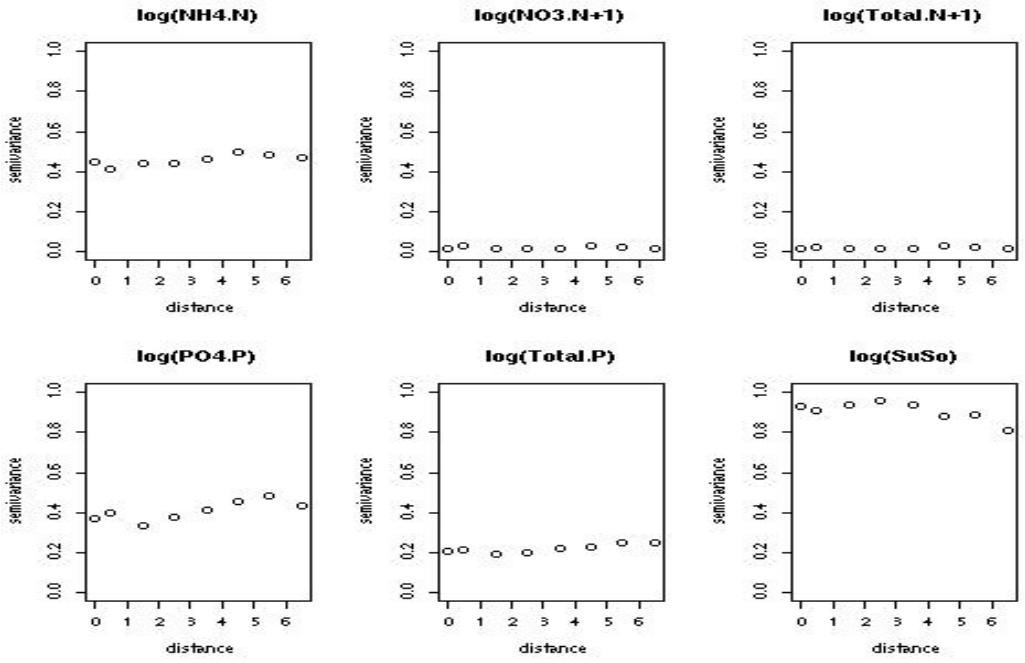

Figure 3.20. Cressie and Hawkins variogram for residuals under an additive model

Gaussian, Spherical and Pure Nugget), showing the estimated parameter for the nugget, the sill, the range and in addition the sum of squares for each model fitted to the residuals. The estimated parameters where obtained using the variofit

option in the package geoR [Pinheiro & Diggle (2009)], using the weighted least squares approach as suggested by Cressie (1980).

According to the results for the parameters, the model that fits best is a pure nugget effect. Recalling that the maximum distance between the two most separated sites in the map is 7 kilometres, the parameter for the range is unacceptably large mainly for the Exponential and Spherical models, making these models unsuitable to describe the residuals of model (3.2).

Spatial model over residuals model 3.2				
Model	Nugget	Sill	Range	Sum of Squares
Pure Nugget	0.2209	0	0	0.0031
Exponential	0.1988	20.61	2709.809	0.0009
Spherical	0.1988	2.336	462.07	0.0009
Gaussian	0.2040	0.5141	19.7707	0.0005

Table 3.2. Spatial model for residuals under an additive model for $log(TotalP)$

3.3.1 Testing for No Effect and Sensitivity Analysis

Having fitted an additive model for the six variables the next step is to assess the need for a nonparametric effect rather than a linear effect, following the same idea developed by Hastie and Tibshirani (1990) to compare different models based on an approximate F-test. The test was performed under the assumption of independence based on the earlier results. According to Table 3.3 there is clear evidence that a nonparametric effect is required for all the variables.
The sensitivity analysis allows us to assess changes in the conclusions under different values of degrees of freedom, showing that the evidence of the need for a nonparametric term is stable.

$log(NH4.N)$		$log(NO3.N+1)$		$log(TotalN+1)$	
Parameter	p-value	Parameter	p-value	Parameter	p-value
year df(6)	<0.001	year df(6)	<0.001	year df(6)	0.030
day df(6)	<0.001	day df(6)	0.028	day df(6)	0.023
(X,Y) df(12)	<0.001	(X,Y) df(12)	<0.001	(X,Y) df(12)	<0.001
$log(PO4.P)$		$log(TotalP)$		$log(SuSo)$	
Parameter	p-value	Parameter	p-value	Parameter	p-value
year df(6)	<0.001	year df(6)	<0.001	year df(6)	<0.001
day df(6)	<0.001	day df(6)	<0.001	day df(6)	<0.001
(X,Y) df(12)	<0.001	(X,Y) df(12)	<0.001	(X,Y) df(12)	<0.001

Table 3.3. p-values for the test of the need for a nonparametric effect opposed to linear effect for year, day and (X,Y) for all variables

Table 3.4 shows the conclusions over different values of degrees of freedom. According to these results, for $log(NH4.N)$, $log(PO4.P)$, $log(TotalP)$ and $log(SuSo)$ there is a clear evidence that nonparametric effects are required for year, day and (X,Y).

For $log(NO3.N+1)$ and $log(TotalN+1)$ there is a marginal result observed for day, year and day respectively, nevertheless the overall result suggests that a nonparametric effect is required for these two variables where a non-linear pattern can be observed in Figures 3.12 and 3.13.

3.3.2 Additive Model including river flow information

River flow information provides a measure of the overall water resources of a region, reflecting regional rainfall and evaporation patterns. In addition, it is sensitive to climatic and other factors, such as land uses and pollutants, allowing us to capture more information about the dynamics of the catchment. It is therefore a natural candidate to include in models for water quality,

p values $log(NH4.N)$							
	df=4	df=6	df=8		df=10	df=12	df=14
year	<0.001	<0.001	<0.001	(X,Y)	<0.001	<0.001	<0.001
day	<0.001	<0.001	<0.001				
p values $log(NO3.N + 1)$							
	df=4	df=6	df=8		df=10	df=12	df=14
year	0.001	<0.001	<0.001	(X,Y)	<0.001	<0.001	<0.001
day	0.297	0.028	<0.001				
p values $log(TotalN + 1)$							
	df=4	df=6	df=8		df=10	df=12	df=14
year	0.104	0.030	<0.001	(X,Y)	<0.001	<0.001	<0.001
day	0.161	0.023	<0.001				
p values $log(PO4.P)$							
	df=4	df=6	df=8		df=10	df=12	df=14
year	<0.001	<0.001	<0.001	(X,Y)	<0.001	<0.001	<0.001
day	<0.001	<0.001	<0.001				
p values $log(TotalP)$							
	df=4	df=6	df=8		df=10	df=12	df=14
year	<0.001	<0.001	<0.001	(X,Y)	<0.001	<0.001	<0.001
day	<0.001	<0.001	<0.001				
p values $log(SuSo)$							
	df=4	df=6	df=8		df=10	df=12	df=14
year	<0.001	<0.001	<0.001	(X,Y)	<0.001	<0.001	<0.001
day	<0.001	<0.001	<0.001				

Table 3.4. p-values sensitivity analysis under different degrees of freedom

To be able to include the flow information as a covariate, there are particular issues which must be addressed. The information provided by the Macaulay Institute contains flow information at one single site. It would be more informative to have a measure of flow as each site of the river. However, based on previous discussion with the Macaulay Institute experts, it was decided to use the information available, assuming that this information reflects the behaviour over all the catchment, given that we are working with a small catchment.

For this exercise we have daily river flow information from 2000 to 2006 and the six variables with information from 2004 to 2008. We took only the information from 2004 to 2006 and matched the river flow information with the specific dates when our six variables were collected.

Figure 3.21 depicts the time series for river flow, indicating a trough between 2005 and 2006. Figure 3.22 depicts plots of flow against all the six variables, indicating non-linear patterns.

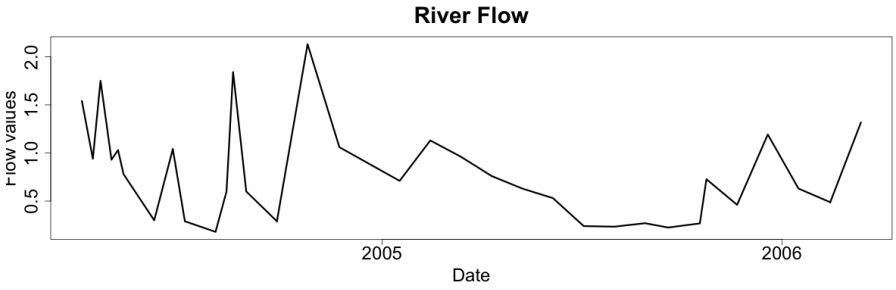

Figure 3.21. Time Series of river flow information

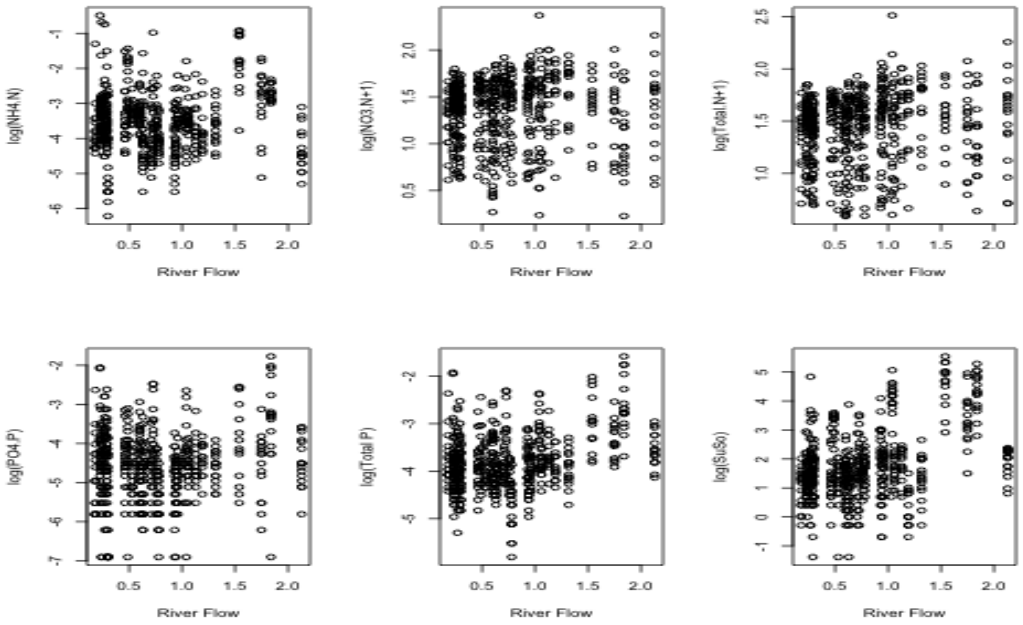

Figure 3.22. River flow against all six variables

The first step is to assess the improvement achieved by including the flow information in an additive model. The additive model to be fitted corresponds to model (3.3), where $m_4(flow)$ corresponds to a smooth function for river flow, where ε_i are assumed independent with mean 0 and constant variance σ^2.

$$y = \beta_0 + m_1(year) + m_2(days) + m_3(X, Y) + m_4(flow) + \varepsilon_i \qquad i = 1, ..., n \quad (3.3)$$

Table 3.5 shows the p-values under a null hypothesis that a model without flow is an adequate description of the data. The results of the approximate F-test confirm that the model which includes flow is superior for all six variables.
Figures 3.23 to 3.28 depict the smooth functions fitted for all six variables under model (3.3). Each plot shows the relationship of the dependent variable against the covariates to assess trend over time and space.

ANOVA between model with and without flow		
$log(NH4.N)$	$log(NO3.N+1)$	$log(TotalN+1)$
<0.001	<0.001	<0.001
$log(PO4.P)$	$log(TotalP)$	$log(SuSo)$
<0.001	<0.001	<0.001

Table 3.5. Comparison between models including flow

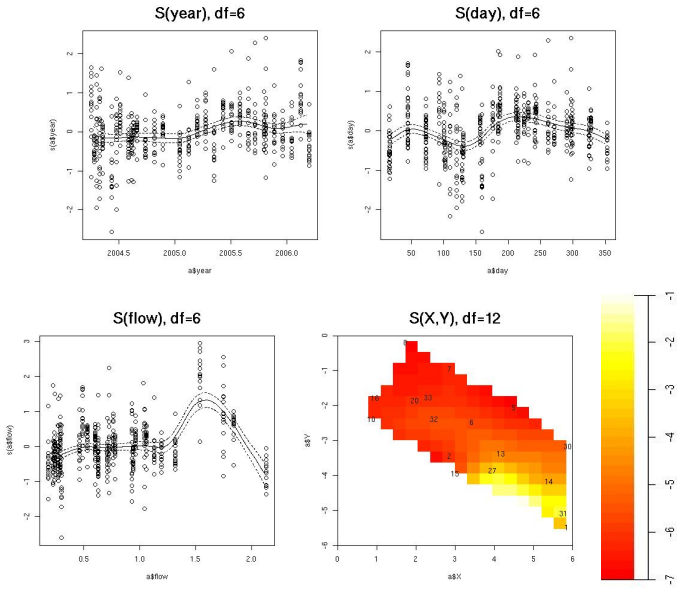

Figure 3.23. Plot of the components additive models for $log(NH4.N)$ including flow

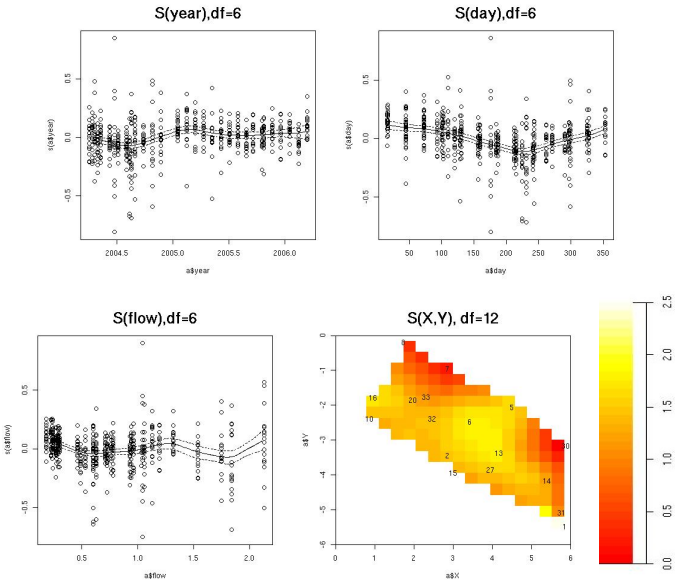

Figure 3.24. Plot of the components additive models for $log(NO3.N + 1)$ including flow

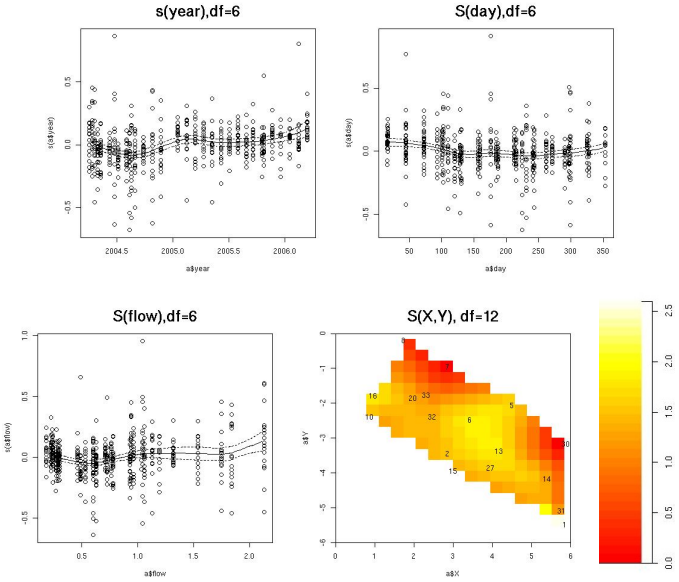

Figure 3.25. Plot of the components additive models for $log(TotalN + 1)$ including flow

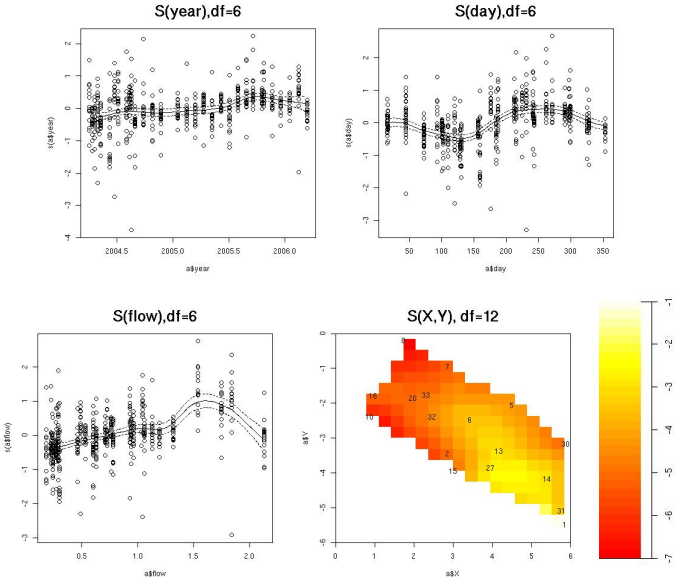

Figure 3.26. Plot of the components additive model for $log(PO4.P)$ including flow

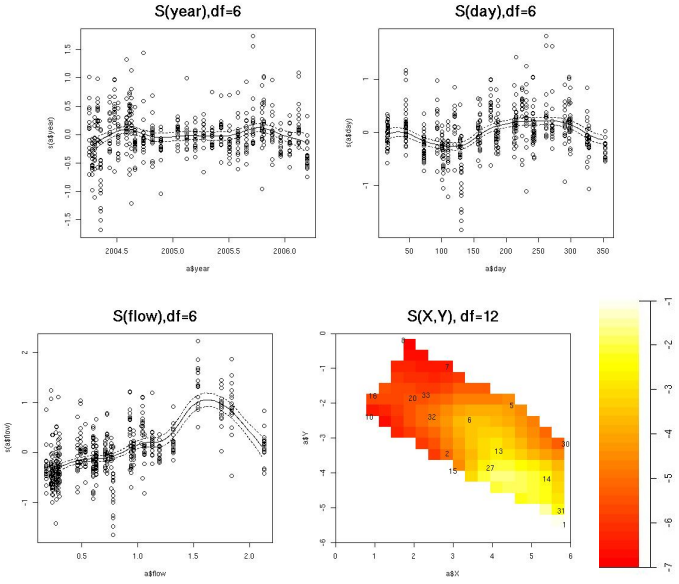

Figure 3.27. Plot of the components additive model for $log(TotalP)$ including flow

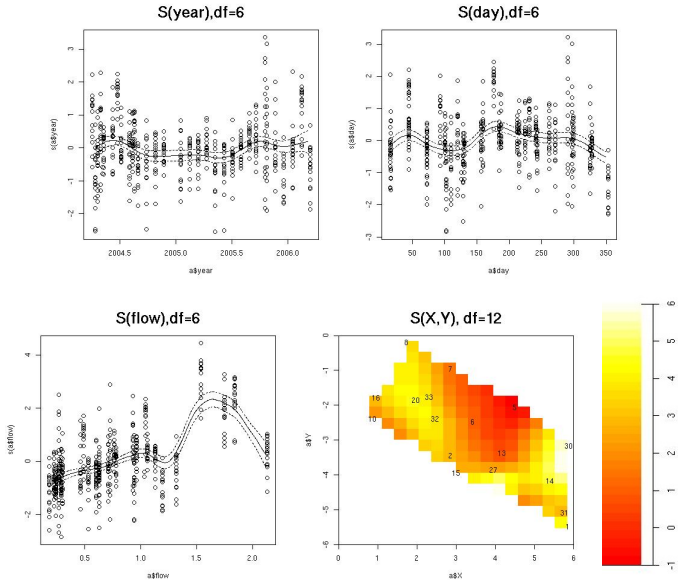

Figure 3.28. Plot of the components additive model for $log(SuSo)$ including flow

Figures 3.29 to 3.34 provide a comparison of the estimates for year, day and (X,Y) under an additive model without flow (upper panel) and the additive model including flow (lower panel). Both models were fitted using the same period of information from 2004 to 2006. The upper panel corresponds to the additive model without river flow, while the lower panel corresponds to the additive model including river flow.

According to these results only $log(NO3.N + 1)$ does not show changes in the estimates for all the three covariates. With respect to year, the estimate for $log(TotalN + 1)$ is the same in both models, for $log(NH4.N)$, $log(TotalP)$ and $log(SuSo)$, the estimates of the model including river flow are better, showing less dispersion for the fitted values in respect to the residuals. For $log(PO4.P)$ the estimate under the model without river flow is better.

For the covariate day, the estimate for $log(NH4.N)$ and $log(SuSo)$ do not show

any change. For $log(TotalN + 1)$ and $log(TotalP)$, the estimates are better under the model including river flow. For $log(PO4.P)$ the estimate under model without flow river is better.

With respect to (X,Y), the estimate of the trend over space, there is no change in the estimates irrespective of whether a model with or without river flow is used.

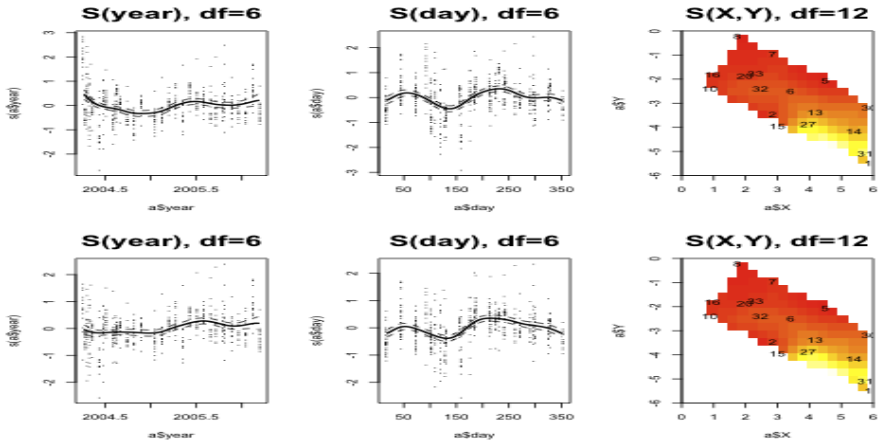

Figure 3.29. $log(NH4.N)$ comparison of the estimates for year, day and (X,Y) under an additive model without flow and an additive model including flow

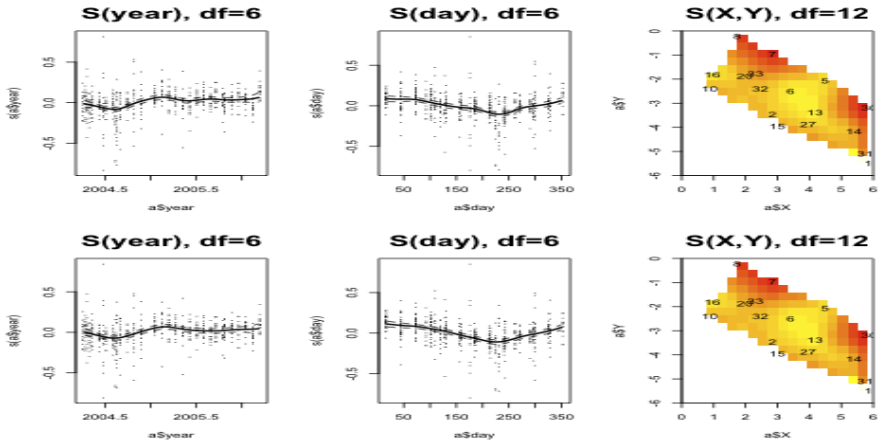

Figure 3.30. $log(NO3.N + 1)$ comparison of the estimates for year, day and (X,Y) under an additive model without flow and an additive model including flow

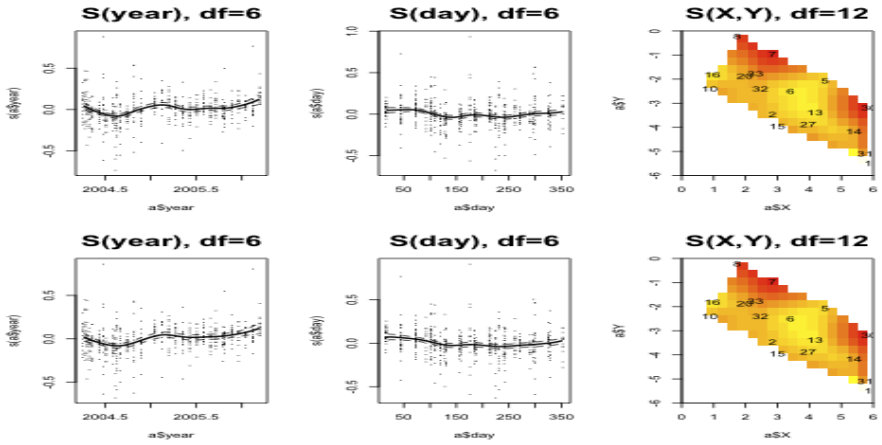

Figure 3.31. $log(TotalN + 1)$ comparison of the estimates for year, day and (X,Y) under an additive model without flow and an additive model including flow

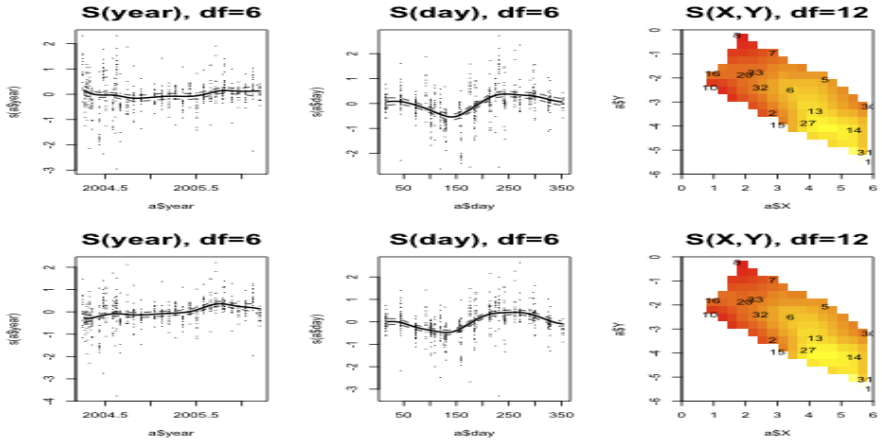

Figure 3.32. $log(PO4.P)$ comparison of the estimates for year, day and (X,Y) under an additive model without flow and an additive model including flow

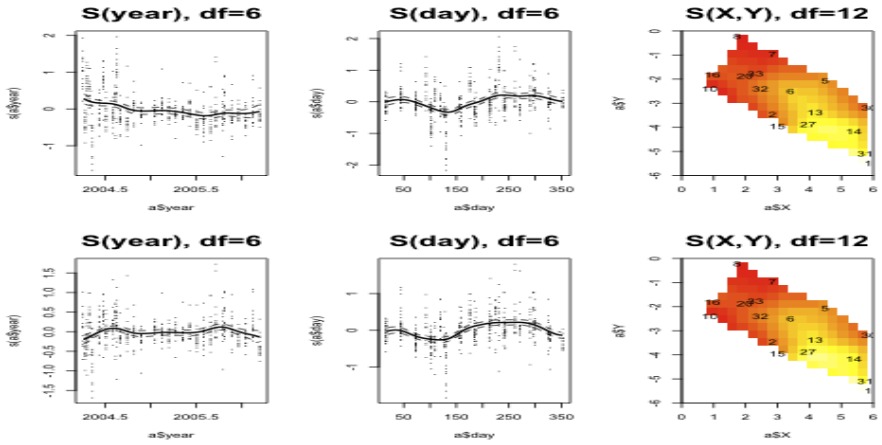

Figure 3.33. $log(TotalP)$ comparison of the estimates for year, day and (X,Y) under an additive model without flow and an additive model including flow

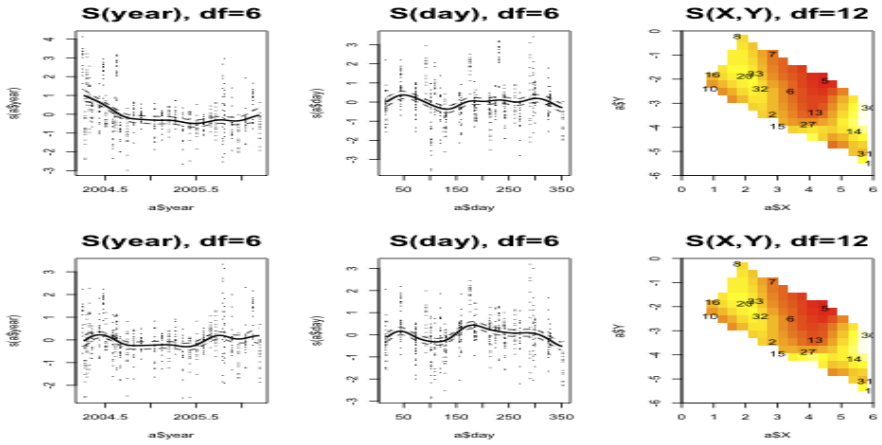

Figure 3.34. $log(SuSo)$ comparison of the estimates for year, day and (X,Y) under an additive model without flow and an additive model including flow

3.4 Testing for No Effect and Sensitivity Analysis including flow.

Having shown the improvement by including flow, the next step is to assess the need for a nonparametric effect. According to Table 3.6, there is clear evidence that a nonparametric effect is not required for the variable day for $log(NO3.N+1)$ and for the variable year for $log(TotalN+1)$. For the other variables there is evidence that a nonparametric effect is required.

$log(NH4.N)$		$log(NO3.N+1)$		$log(TotalN+1)$	
Parameter	p-value	Parameter	p-value	Parameter	p-value
year df(6)	<0.001	year df(6)	0.003	year df(6)	0.064
day df(6)	<0.001	day df(6)	0.551	day df(6)	0.026
(X,Y) df(12)	<0.001	(X,Y) df(12)	<0.001	(X,Y) df(12)	<0.001
flow df(6)	<0.001	flow df(6)	<0.001	flow df(6)	0.002
$log(PO4.P)$		$log(TotalP)$		$log(SuSo)$	
Parameter	p-value	Parameter	p-value	Parameter	p-value
year df(6)	<0.001	year df(6)	<0.001	year df(6)	<0.001
day df(6)	<0.001	day df(6)	<0.001	day df(6)	<0.001
(X,Y) df(12)	<0.001	(X,Y) df(12)	<0.001	(X,Y) df(12)	<0.001
flow df(6)	<0.001	flow df(6)	<0.001	flow df(6)	<0.001

Table 3.6. p-values for test of the need for a nonparametric effect opposed to a linear effect including flow

Following the same idea discussed earlier, the sensitivity analysis allows us to assess the sensitivity of the test under different values of degrees of freedom. According to Table 3.7, for $log(NH4.N)$, $log(PO4.P)$, $log(TotalP)$ and $log(SuSo)$ a nonparametric effect is better than a linear effect over different values of degrees of freedom. For $log(NO3.N+1)$ and $log(TotalN+1)$ the conclusion is that a nonparametric effect is not required for day and year, indicating that a semiparametric model could be the best approach.

p values $log(NH4.N)$							
	df=4	df=6	df=8		df=10	df=12	df=14
year	<0.001	0.005	<0.001	(X,Y)	<0.001	<0.001	<0.001
day	<0.001	<0.001	<0.001				
flow	<0.001	<0.001	<0.001				
p values $log(NO3.N + 1)$							
	df=4	df=6	df=8		df=10	df=12	df=14
year	0.015	0.003	0.001	(X,Y)	<0.001	<0.001	<0.001
day	0.996	0.550	0.190				
flow	<0.001	<0.001	<0.001				
p values $log(TotalN + 1)$							
	df=4	df=6	df=8		df=10	df=12	df=14
year	0.081	0.064	0.019	(X,Y)	<0.001	<0.001	<0.001
day	0.117	0.026	0.006				
flow	<0.001	0.002	0.005				
p values $log(PO4.P)$							
	df=4	df=6	df=8		df=10	df=12	df=14
year	0.006	0.003	0.004	(X,Y)	<0.001	<0.001	<0.001
day	<0.001	<0.001	<0.001				
flow	<0.001	<0.001	<0.001				
p values $log(TotalP)$							
	df=4	df=6	df=8		df=10	df=12	df=14
year	0.015	<0.001	<0.001	(X,Y)	<0.001	<0.001	<0.001
day	<0.001	<0.001	<0.001				
flow	<0.001	<0.001	<0.001				
p values $log(SuSo)$							
	df=4	df=6	df=8		df=10	df=12	df=14
year	<0.001	<0.001	<0.001	(X,Y)	<0.001	<0.001	<0.001
day	<0.001	<0.001	<0.001				
flow	<0.001	<0.001	<0.001				

Table 3.7. p-values sensitivity analysis under different degrees of freedom including flow information

According to the results in Table 3.7, a model with linear effect for days for $log(NO3.N+1)$ and year for $log(TotalN+1)$ will be adequate. In this case the linear terms added for day corresponds to $sin\left(\frac{2\pi\ day}{366}\right)$ and $cos\left(\frac{2\pi\ day}{366}\right)$ which capture the seasonality effect suitably [Esterby (1993)].

For $log(NO3.N+1)$ the model chosen is $y = \beta_0+\beta_1 sin\left(\frac{2\pi\ day}{366}\right)+\beta_2 cos\left(\frac{2\pi\ day}{366}\right)+ m_1(year)+m_2(X,Y)+m_3(flow)$.

For $log(TotalN+1)$ the model chosen is $y = \beta_0+\beta_1 year+m_1(days)+m_2(X,Y)+ m_3(flow)$.

3.5 Diagnostic Check

Figure 3.35 depicts the residuals versus fitted values for all the variables indicating that the models fit well. The linear pattern for $log(PO4.P)$ correspond to limit of detection values. The results for $log(NO3.N+1)$ and $log(TotalN+1)$ correspond to the semi-parametric models.

Figure 3.36 and 3.37 show the variograms to evaluate the independence over time and space showing no autocorrelation over time. Based on that result the test for independence over space can be applied, and this indicates that only $log(TotalP)$ shows correlation over space, with a p-value=0.035, although it is not strong.

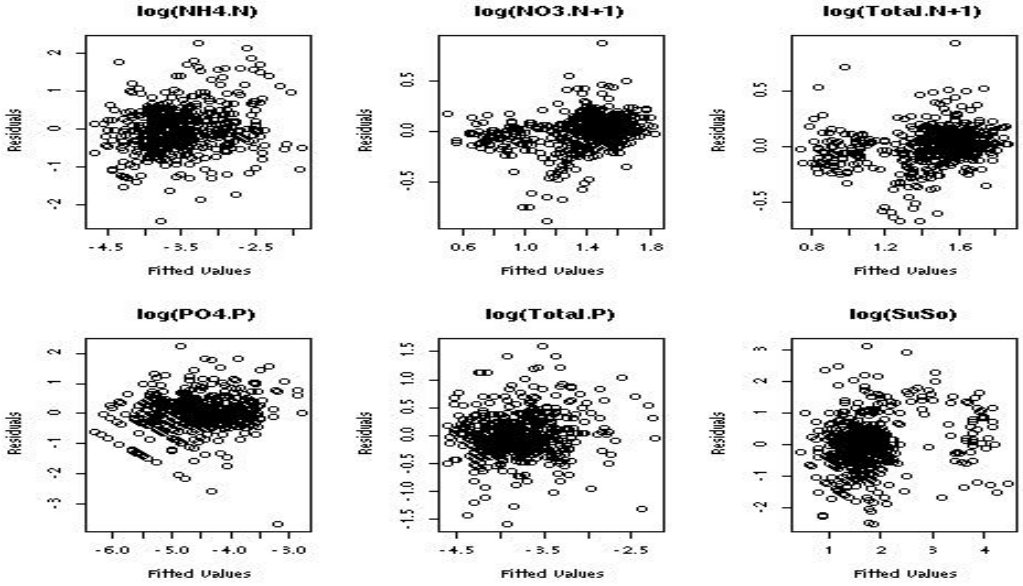

Figure 3.35. Residuals versus fitted values under an additive model including flow

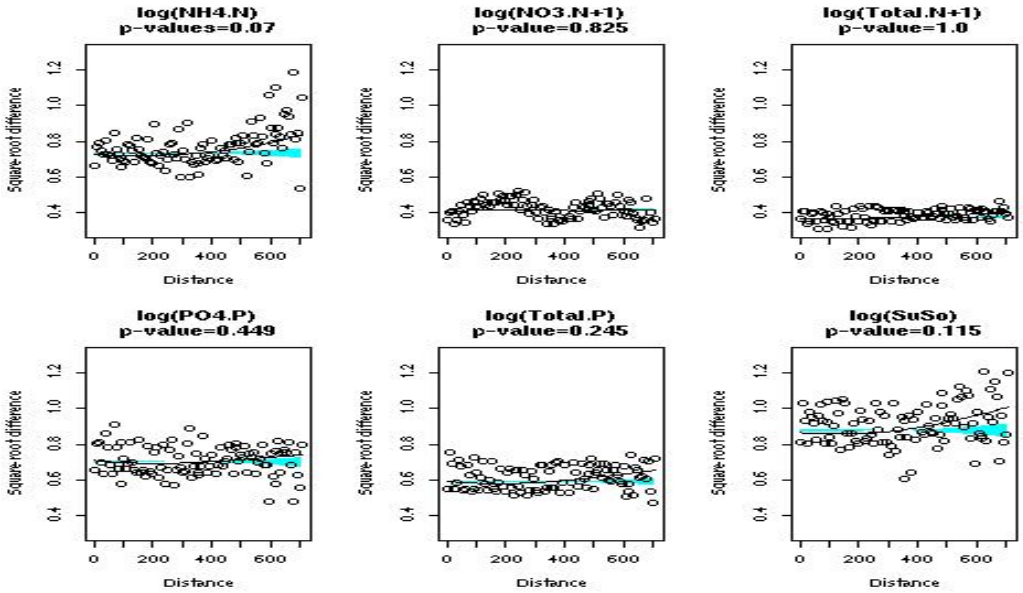

Figure 3.36. Independence test over time for residuals under an additive model including flow

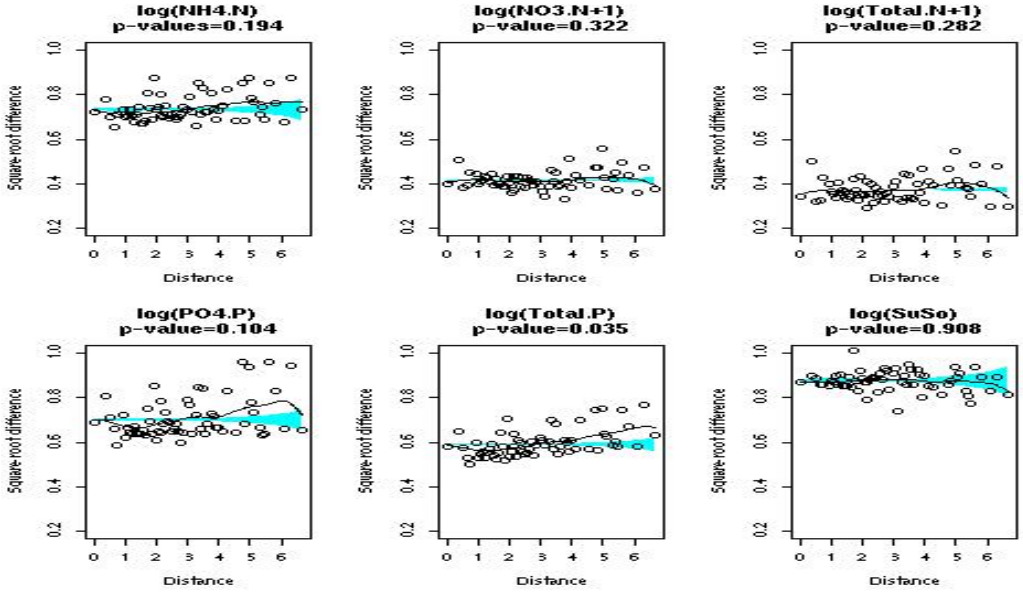

Figure 3.37. Independence test over space for residuals under an additive model including flow

Figure 3.38 shows the variogram suggested by Cressie and Hawkins (1980) for all the six variables. As well as in the previous section only $log(TotalP)$ shows a shape that might fit with some spatial model.

In the same way as in the previous section, Table 3.8 depicts the results of different spatial models (Pure Nugget, Exponential, Spherical and Gaussian) for the residuals of model (3.3) for $log(TotalP)$. According to this result a Gaussian model could be used to explain the spatial covariance structure of $log(TotalP)$. However, based on the previous test of independence over space carried out over the residuals of model (3.2), where the amount of information used was twice (observed period 2004 to 2008) the amount of information that we are using in model (3.3) (observed period 2004 to 2006), the fact that $log(NH4.N)$, $log(NO3.N+1)$, $log(TotalN+1)$, $log(PO4.P)$ and $log(SuSo)$, did not show evidence of correlation over space and a weak evidence of spatial correlation with a p-value=0.035, the decision was to assume independence over space for $log(TotalP)$.

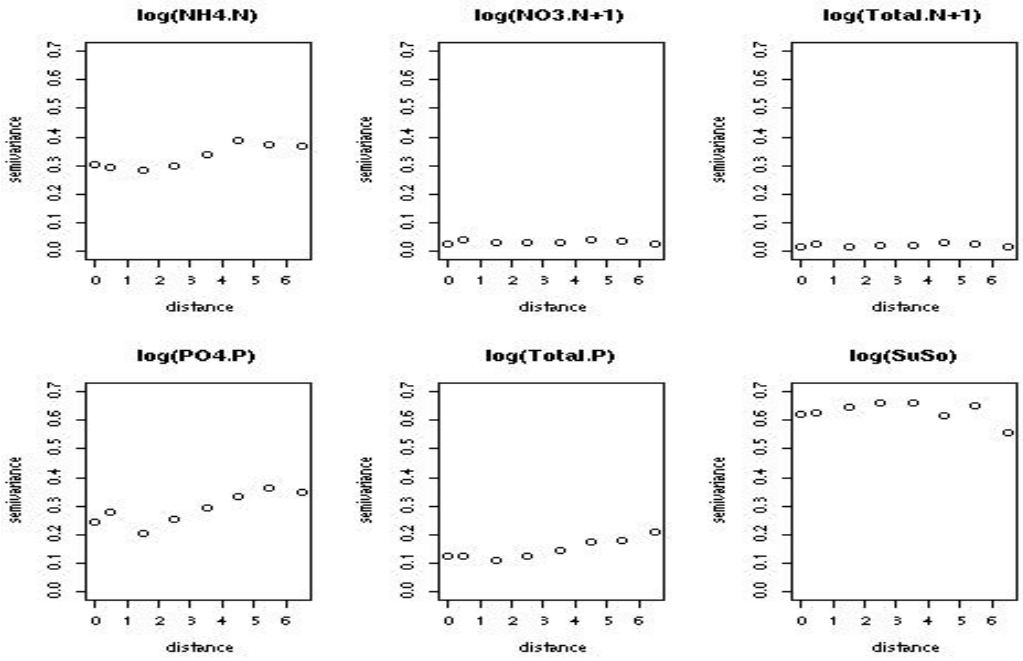

Figure 3.38. Cressie and Hawkins variogram for residuals under an additive model including flow

Spatial model over residuals model 3.3				
Model	Nugget	Sill	Range	Sum of Squares
Pure Nugget	0.1494	0	0	0.0008
Exponential	0.1074	150.93	10998	0.0001
Spherical	0.1075	17.294	1890.71	0.0001
Gaussian	0.1141	0.1284	6.012	0.0006

Table 3.8. Spatial model for residuals under an additive model including flow

Given that the assumption made here corresponds to a specific case, it is important to highlight that the use of additive models for correlated data is possible, regarding that the principal effects of correlation are in the calculation of standard errors and in the implementation of model comparison [Giannitrapani et al. (2005)]. In the case of correlated data, a modification of the RSS through the generalised least squares criterion allows us to include the correlation structure as

$$RSS = y^t(I - P)^t V^{-1}(I - P)y,$$

where V corresponds to the estimate of the correlation matrix, providing a solution for correlated data.

3.6 Summary

The ability of additive models to fit a smooth function, unrestricted with respect to shape, improves the representation of trend over time and space simultaneously. This helpfully extends the statistical tools available for environmental data, when a linear approach does not offer a suitable description.

In the analysis made in this chapter, a sequence of steps was presented to cover the following issues:

- assessment of time and/or space correlation
- fit of additive models
- model comparison (between models and linear versus nonparametric effect)
- sensitivity analysis

The comparison between a linear and a nonparametric effect allows us to confirm whether a nonparametric effect is required. Although the test suggested by Hastie and Tibshirani (1990) does not follow all the properties of an F distribution, it provides good guidance that allows comparisons between different models to be

made.

The inclusion of flow as a covariate provided useful information that significantly improves the fitted models.

For $log(NH4.N)$, $log(PO4.P)$ $log(TotalP)$ and $log(SuSo)$ there is a upward trend with a peak when flow values are between 1.2 and 2.0. For $log(NO3.N+1)$ and $log(TotalN+1)$ the values fluctuate showing a peak close to 1.3 and two trough in 0.5 and 1.7.

The comparison between a linear effect and a nonparametric effect for $log(NO3.N+1)$ and $log(TotalN+1)$ allows us to identify that a semi-parametric model is the best approach.

Regarding the scientific question mentioned at the beginning of this chapter, it can be observed how each variable shows a fluctuation over time and a clear trend over space. The following conclusions correspond to the results observed in model (3.2), given that there are more data available and the interest of the Macaulay Institute corresponds to this period of time. This corresponds to the information collected regularly at 17 sites from April 2004 to November 2008.

- For $log(NH4.N)$ there is fluctuation over the years showing a slightly downward trend with two peaks, the first one in the middle of 2005 and the second one in the middle of 2008, while a trough is observed between 2007 and 2008. Over space, a trend in the direction of site 1 and 31 is observed where the highest values are located.

- $log(NO3.N+1)$ showed a fluctuation over the years with a slightly upward trend, with two peaks in the middle of 2005 and the beginning of 2008 and two troughs at the end of 2004 and in the middle of 2007. With respect to

space, the highest concentration are observed in the site 2, 5, 6, 13, 15 and 27, with 6 and 13 the sites with the highest values.

- $log(TotalN+1)$ showed a fluctuation over the years, with a slightly upward trend with two peaks in the middle of 2005 and the beginning of 2008 and two troughs at the end of 2004 and in the middle of 2007. Over space, the variable followed the same pattern as $log(NO3.N+1)$.

- $log(PO4.P)$ showed a downward trend over time with two main drops in the middle of 2006 and at the end of 2008. Over space, the highest concentrations were observed in the sites 1, 13,14, 27 and 31.

- $log(TotalP)$ showed a clear downward trend until 2007, with a subsequent shift in the trend, reaching a peak in the middle of 2008. Over space, the conclusions follow the same pattern observed for $log(PO4.P)$.

- $log(SuSo)$ showed a clear downward trend until the beginning of 2005 with a diminution in the slope but still downward until the end of 2007, reaching a peak in the middle of 2008. Over space, the highest values were observed in the sites 20, 30 and 33, with 30 the site with the highest value .

Chapter 4

Statistical Models for River Networks

In the previous chapter, the main aim was to fit an additive model to capture the trend over time and space simultaneously, assuming that all the 17 sites were connected over space and a Euclidean distance model was appropriate. In this chapter, the main aim is to fit an additive model using distance measures which reflect the fact that not all sites are flow-connected and that distance should be measured along the river network.

The first part of this chapter provides an introduction to the uses of spatial modelling over a river network. The second part which is the main aim of this chapter, discusses how this idea can be used in fitting an additive model.

The analysis of a river network comes with two main questions: 1) what is the proper distance measure to be used to capture the behaviour of the river over space and 2) are the current models based on Euclidean distance suitable to capture the behaviour over space.

The use of Euclidean distance over the river seems to be a good first approach. However, the river distance discussed by Ver Hoef et al (2005) might be more

appropriate. Indeed, it would also be feasible to use a mixture of both, river distance and Euclidean distance [Cressie et al. (2006)], adding the continuity of land over space too.

The use of river distance, as discussed by Ver Hoef et al. (2005), has the problem that the spatial autocovariance models based on Euclidean distance might not be positive-definite, resulting in an invalid model. To tackle this problem the use of moving averages, or kernel convolutions, provides suitable models for the spatial covariance structure. The integration of

$$Z(s) = \int_{-\infty}^{\infty} g(x - s|\theta)W(x)dx, \tag{4.1}$$

where $W(x)$ is white noise and $g(x|\theta)$ is called the moving average function, allows a valid autocovariance function defined as

$$C(h|\theta) = \begin{cases} \int_{-\infty}^{\infty}(g(x|\theta))^2dx + \nu^2, & \text{if h=0,} \\ \int_{-\infty}^{\infty} g(x|\theta)g(x-h|\theta)dx, & \text{if h} > 0, \end{cases} \tag{4.2}$$

where h corresponds to Euclidean distance and v_j^2 corresponds to the nugget effect at $h = 0$.

It is necessary to include in this expression a proper weighting to compensate for the effect in the variance caused by splits in some part of the river [Ver Hoef et al. (2005)]. The idea is to provide a weight to those cases where there are splits upriver in such a way that the sum of all of them is equal to 1.

Different options have been mentioned as possible weights: flow, area of each basin or river order. This modifies expression (4.2) by adding the proper weightings as

$$C(s_i, t_j|\theta) = \begin{cases} 0 & \text{if } s \text{ and } t \text{ are not flow-connected,} \\ C_1(0) + \nu^2 & \text{if } s{=}t, \\ \prod_{j \in B_{s_i,t_j}} \sqrt{w_j} C_1(d(s_i, t_j)) & \text{otherwise.} \end{cases} \tag{4.3}$$

where $C_1(h) = \int_{-\infty}^{\infty} g(x|\theta)g(x-h|\theta)dx$ and $d(s_i, t_j)$ is the river distance.

Following the same idea, Cressie et al (2006) suggested a mixture of river distance and Euclidean distance, including a parameter $\lambda \in [0, 1]$ which controls the contribution of spatial dependence provided by river distance and Euclidean distance. In the particular case of the exercise developed by Cressie et al (2006), the kernel chosen was $(1-\frac{d}{r})$ I$(0 \leq d \leq r)$. Following the same notation of Cressie et al (2006) the covariance function can be written as

$$\begin{aligned} cov(Y(s), Y(t)) = \lambda\sigma^2 \left[\Big(\frac{\Omega(t)}{\Omega(s)}\Big)^{\frac{1}{2}} \Big(1 - \frac{3}{2}\frac{|s-t|}{r_1}\Big) + \frac{1}{2}\Big(\frac{|s-t|}{r_1}\Big)^3 \right] \\ + (1-\lambda)\sigma^2 \left[\Big(1 - \frac{3}{2}\frac{||s-t||}{r_2}\Big) + \frac{1}{2}\Big(\frac{||s-t||}{r_2}\Big)^3 \right] \end{aligned} \tag{4.4}$$

where $\Omega(t)$ and $\Omega(s)$ provide the weighting based on the basin order and $|s-t| = d \leq r_1$ and $||s-t|| = d \leq r_2$ correspond to the river distance and Euclidean distance respectively.

This brief explanation introduces the problem of working with river network information and the use of different distances, Euclidean and/or river distance, with a proper weighting to ensure a proper autocovariance function.

4.1 River Network Modelling using Nonparametric Regression

The main aim of this section is to present the use of nonparametric regression using river distance, including the fact that all sites are not flow-connected.

The use of a directed acyclic graph (DAG) [Whittaker (1990)], helps us to understand the meaning of flow-connected or connectedness. Each of the circles represent a site measured over the river network, while the arrows allow the flow direction and the sites which are flow-connected to be identified.

According to Figure 4.1, S1, S2, S4 and S5 are flow-connected as well as S3, S4 and S5, while S1 and S2 are not flow-connected with S3.

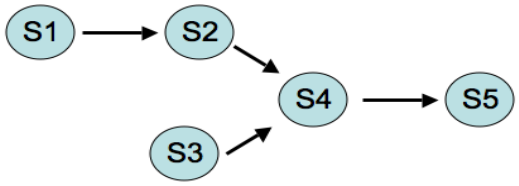

Figure 4.1. Directed acyclic graph (DAG) to explain flow connectedness using 5 sites measured over a river network.

The connectedness between sites can be expressed as an n x n symmetric matrix, where n corresponds to the number of sites. The connectedness matrix corresponds to a matrix with 1's in the diagonal, while off the diagonal the matrix has the value 1 if both sites are connected and otherwise takes the value 0. In this case the connectedness matrix corresponds to a 17 x 17 matrix which was defined based on Figure 3.1 with a water flow direction to site 1.

The distance corresponds to river distance between each of the sites expressed in a distance matrix. Alternatively, a vector of distances from the river mouth to each observation can also being used. In this case a vector d with the distance from each site to site 1 was calculated along the river network.

Having defined a distance along the river and indicated which sites are flow-connected through a connectedness matrix, the modelling of a river network is carried out using a nonparametric model of the form

$$y_i = m(x_i) + \varepsilon_i \qquad i = 1, ..., n \tag{4.5}$$

where $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$. An estimate for $\hat{m}(x)$ can be obtained by a local mean estimator as

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} w(x_i - x; h) y_i}{\sum_{i=1}^{n} w(x_i - x; h)}, \tag{4.6}$$

where $w(x_i - x; h)$, the weight function chosen, corresponds to a normal density centred on zero with standard deviation equal to h, with h the smoothing parameter.

Given that we are using river distance, $(x_i - x)$ is replaced by (d_i) in (4.6) as,

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} w(d_i; h) \delta_{ij} y_i}{\sum_{i=1}^{n} w(d_i; h) \delta_{ij}}, \tag{4.7}$$

where

$$\delta_{ij} = \begin{cases} 1 & \text{if i and j are flow-connected,} \\ 0 & \text{otherwise.} \end{cases}$$

allowing us to obtain an estimate $\hat{m}(x)$, using only flow-connected points.

If the interest is to obtain an estimate for a new location along the river network, the same idea can be used over new points. Expression (4.7) can be defined as

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} w(x_i - x; h)\delta_i y_i}{\sum_{i=1}^{n} w(x_i - x; h)\delta_i}, \tag{4.8}$$

where x corresponds to the new point to be estimated. In the same way, δ_i allows us to identify if x is flow-connected to x_i, where .

$$\delta_i = \begin{cases} 1 & \text{if } x \text{ is flow-connected to } x_i, \\ 0 & \text{otherwise.} \end{cases}$$

At this point we have the weighting, the river distance and the connectedness but it still is necessary to select a suitable value for h. The idea is to evaluate how different values of h capture the trend observed over the 17 sites.

Under a particular value of h, it is possible to assess if the pattern in the observed values is reflected in the estimate, however in an attempt to provide a better representation, an alternative is to obtain an estimate over new points along the river network. The idea of this section is to provide the estimate of the observed values, the estimate over new observations along the river network and to assess how different values of h change the estimates.

For simplicity this example was carried out for one specific date (12 April 2004), using the observed values of $log(NH4.N)$, while the new points correspond to 136 points generated to reproduce the pattern on Figure 3.1. To obtain an estimate

for the new points, the river distance between each of the new points and site 1 was calculated, including how the new points are flow-connected to each of the 17 sites.

Figure 4.2 depicts the estimates and the new estimates under different values of h. In addition, it is possible to observe how the estimates tend to be more similar as the h value is increased. This provides a graphical approach to choose particular values of h. For this specific example a value of $h = 1.5$ seems to provide a suitable result.

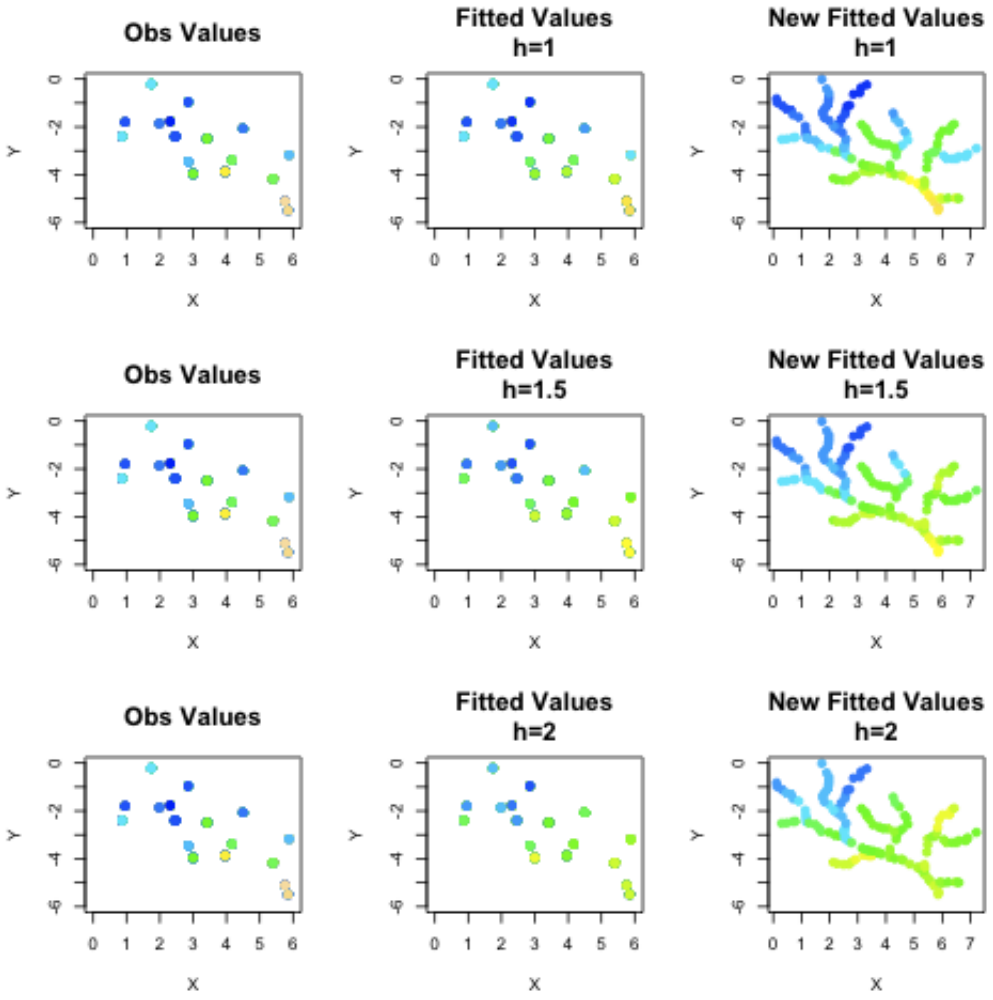

Figure 4.2. Choosing a smoothing parameter for $log(NH4.N)$ on 12-April 2004

4.2 Additive Model Including River Distance

Following this idea, the model to be fitted corresponds to model (4.9) under the assumption that ε_i are independent with mean 0 and constant variance σ^2, where d corresponds to the upstream distance from each site to site 1 in kilometres.

$$y = \beta_0 + m_1(year) + m_2(days) + m_3(d) + \varepsilon_i \qquad i = 1, ..., n \tag{4.9}$$

Figures 4.3 to 4.8 depict the additive model fitted for each of the variables, showing the smooth function fitted for year, day and upstream distance. For year and day, the partial residuals can be observed as well as $\pm$ 2 standard error bands.

To assess if the term $m_3(d)$ was capturing the trend over space suitably, the strategy was to obtain the partial residuals as $r_i = y_i - \bar{y} - \hat{m}_1(year) - \hat{m}_2(day)$, calculate the average partial residuals by site and calculate the estimate over the 136 points created, using expression (4.8). The average partial residuals provides a guide to the pattern. These values correspond to the bigger circles for each of the 17 sites.

For $log(NH4.N)$, $log(NO3.N+1)$, $log(TotalN+1)$, $log(PO4.P)$ and $log(TotalP)$ a similar value of degrees of freedom was used for all the variables. For $log(SuSo)$ the value for the smoothing parameter h for the upstream distance was lower, as a smaller value was required to capture the trend over space in the catchment.

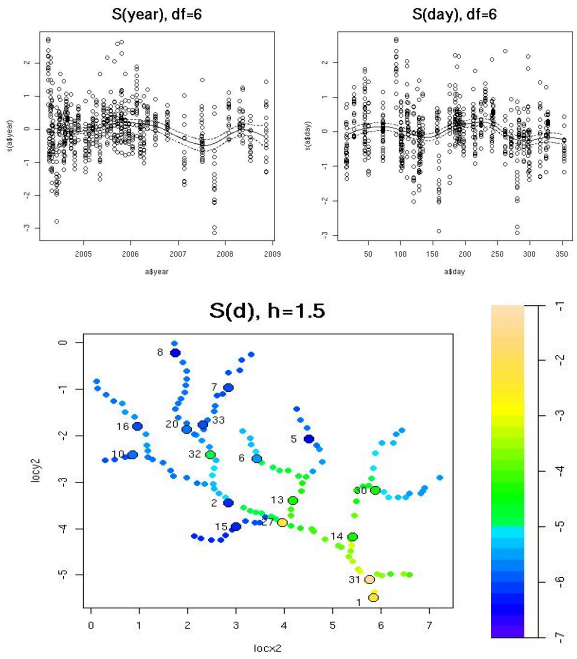

Figure 4.3. Plot of the components additive model for $log(NH4.N)$ river network structure

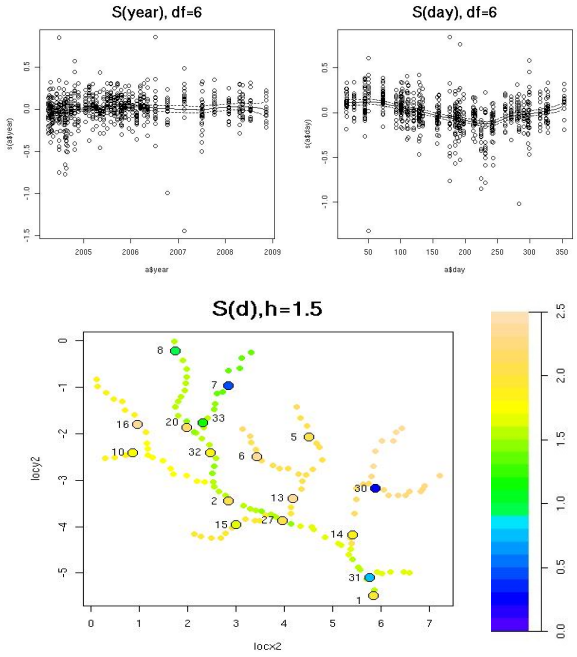

Figure 4.4. Plot of the components additive model for $log(NO3.N + 1)$ river network structure

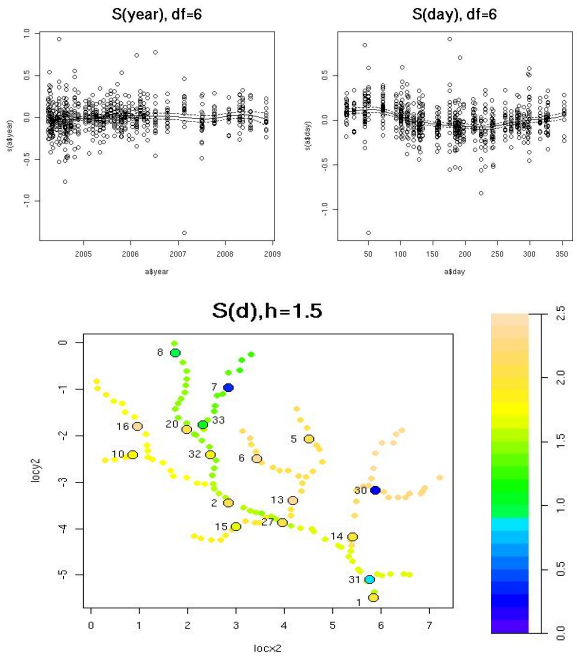

Figure 4.5. Plot of the components additive model for $log(TotalN + 1)$ river network structure

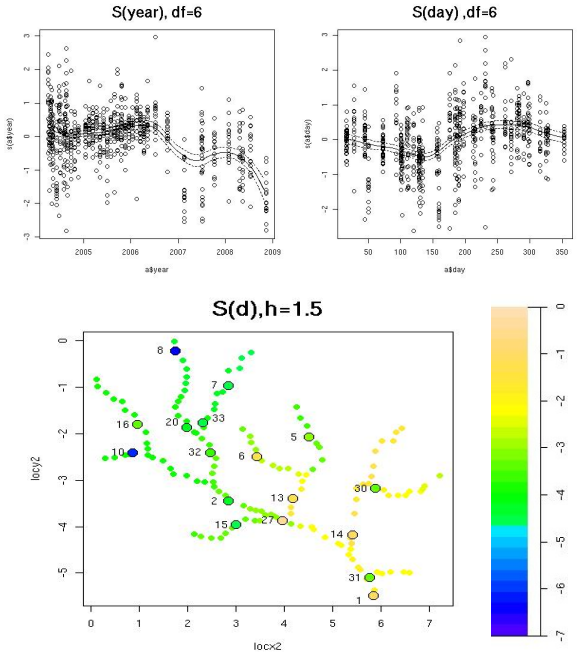

Figure 4.6. Plot of the components additive model for $log(PO4.P)$ river network structure

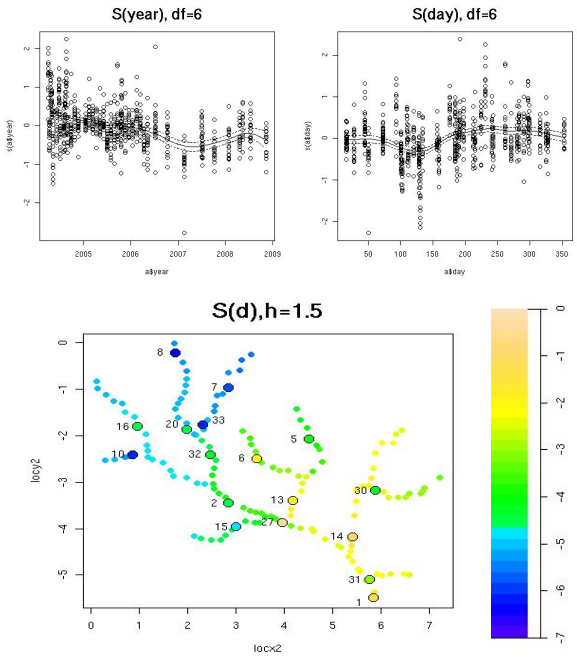

Figure 4.7. Plot of the components additive model for $log(TotalP)$ river network structure

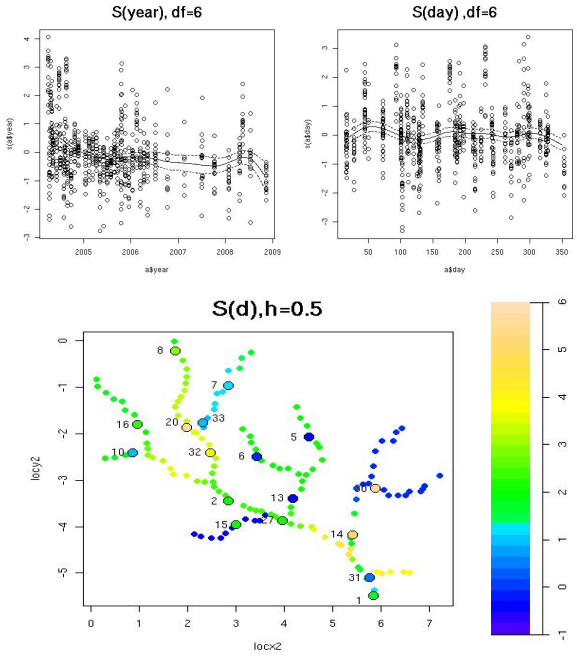

Figure 4.8. Plot of the components additive model for $log(SuSo)$ river network structure

4.3 Testing for No Effect and Sensitivity Analysis using the River Network

The results presented in this section were obtained under the assumption of independent data. The evidence for independence over time and space observed in earlier chapters, allows us to make that assumption, ensuring that the results for no effect and the sensitivity analysis are valid.

Having fitted the additive models the next step is to assess the need for a nonparametric effect versus a linear effect. Following the same idea as in the previous chapters, Table 4.1 shows the results indicating that a nonparametric effect is required for year, day and d.

The sensitivity analysis allows us to assess the sensitivity of the test under different values of degrees of freedom. Table 4.2 confirms the need for a nonparametric effect for year, day and d.

$log(NH4.N)$		$log(NO3.N+1)$		$log(TotalN+1)$	
Parameter	p-value	Parameter	p-value	Parameter	p-value
year df(6)	<0.001	year df(6)	<0.001	year df(6)	<0.001
day df(6)	<0.001	day df(6)	<0.001	day df(6)	<0.001
(d) h(1.5)	<0.001	(d) h(1.5)	<0.001	(d) h(1.5)	<0.001
$log(PO4.P)$		$log(TotalP)$		$log(SuSo)$	
Parameter	p-value	Parameter	p-value	Parameter	p-value
year df(6)	<0.001	year df(6)	<0.001	year df(6)	<0.001
day df(6)	<0.001	day df(6)	<0.001	day df(6)	<0.001
(d) h(1.5)	<0.001	(d) h(1.5)	<0.001	(d) h(0.5)	<0.001

Table 4.1. p-values for test of the need for a nonparametric effect opposed to a linear effect, River Network structure

p values $log(NH4.N)$							
	df=4	df=6	df=8		h=2	h=1.5	h=1
year	<0.001	<0.001	<0.001	(d)	<0.001	<0.001	<0.001
day	<0.001	<0.001	<0.001				
p values $log(NO3.N + 1)$							
	df=4	df=6	df=8		h=2	h=1.5	h=1
year	<0.001	<0.001	<0.001	(d)	<0.001	<0.001	<0.001
day	<0.001	<0.001	<0.001				
p values $log(TotalN + 1)$							
	df=4	df=6	df=8		h=2	h=1.5	h=1
year	<0.001	<0.001	<0.001	(d)	<0.001	<0.001	<0.001
day	<0.001	<0.001	<0.001				
p values $log(PO4.P)$							
	df=4	df=6	df=8		h=2	h=1.5	h=1
year	<0.001	<0.001	<0.001	(d)	<0.001	<0.001	<0.001
day	<0.001	<0.001	<0.001				
p values $log(TotalP)$							
	df=4	df=6	df=8		h=2	h=1.5	h=1
year	<0.001	<0.001	<0.001	(d)	0.005	<0.001	<0.001
day	<0.001	<0.001	<0.001				
p values $log(SuSo)$							
	df=4	df=6	df=8		h=2	h=1.5	h=1
year	<0.001	<0.001	<0.001	(d)	<0.001	<0.001	<0.001
day	<0.001	<0.001	<0.001				

Table 4.2. p-values sensitivity Analysis under different degrees of freedom, River Network structure

The residuals versus fitted values (Figure 4.9) show no evidence that the models do not fit well, confirming that the models perform suitably over all the six variables. For $log(PO4.P)$ the linear pattern corresponds to limit of detection values.

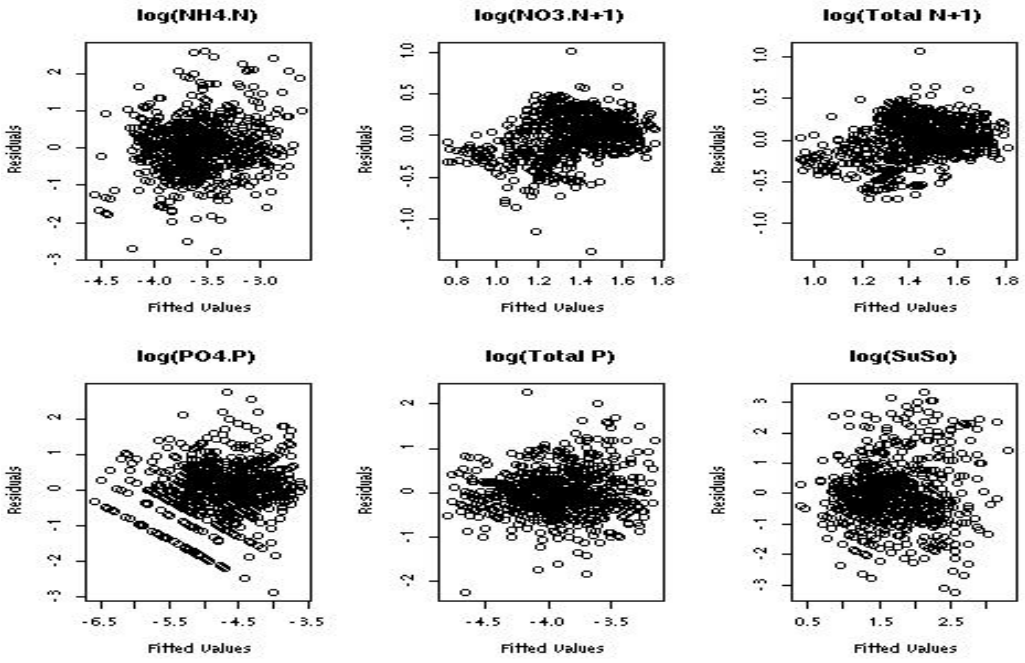

Figure 4.9. Residuals versus Fitted Values River Network

4.4 Comparison of Euclidean and upstream distance

As part of this chapter, one of the main aims is to assess if there is an improvement when the trend over space is captured by river distance rather than Euclidean distance. The comparison for all variables was carried out, assessing the performance of the additive models using Euclidean and river distance.

Figures 4.10 to 4.15 depict the smooth function for each variable using both distances as well as the residuals versus the fitted values.

According to these results for $log(NH4.N)$, $log(NO3.N+1)$, $log(TotalN+1)$, $log(PO4.P)$ and $log(TotalP)$, both results are similar, while for $log(SuSo)$ the smooth function using Euclidean distance, captured better the trend over space.

The residuals versus fitted values indicated that the models fitted using river distance and Euclidean distance, fit well for $log(NH4.N)$, $log(PO4.P)$, $log(TotalP)$ and $log(SuSo)$.

For $log(NO3.N+1)$ and $log(TotalN+1)$ both models showed a good performance, although a slight trend is observed in the residuals when the additive model is fitted using river distance.

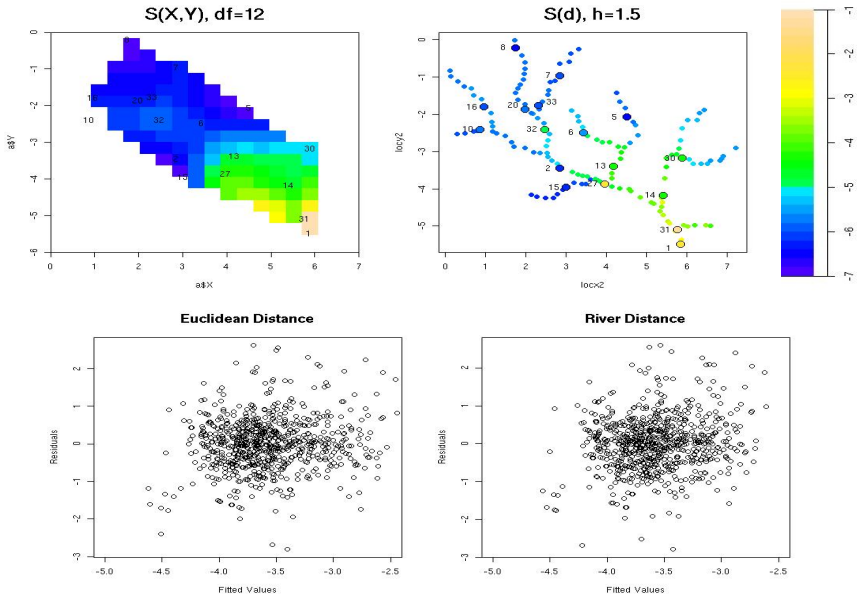

Figure 4.10. $log(NH4.N)$ comparison of the smooth function fitted to capture the trend over space and the residuals using Euclidean and river distance

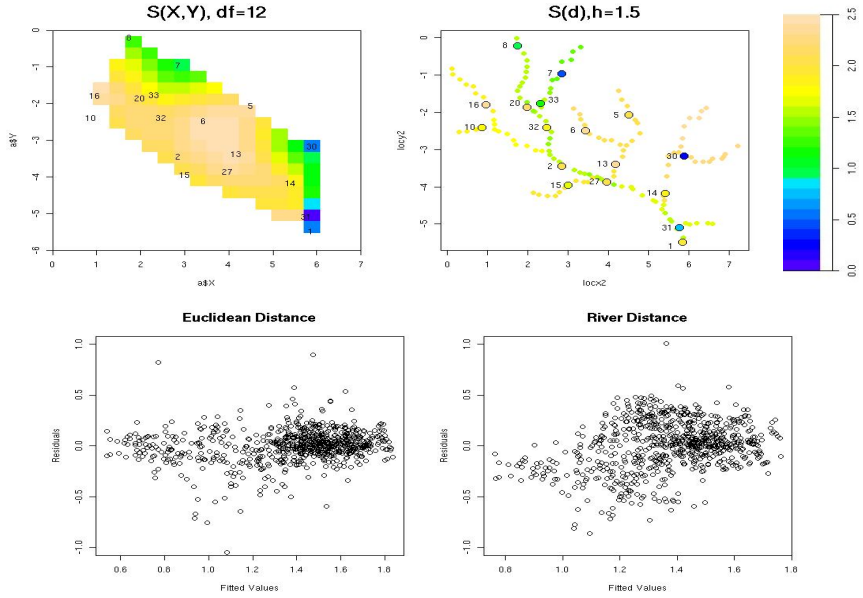

Figure 4.11. $log(NO3.N + 1)$ comparison of the smooth function fitted to capture the trend over space and the residuals using Euclidean and river distance

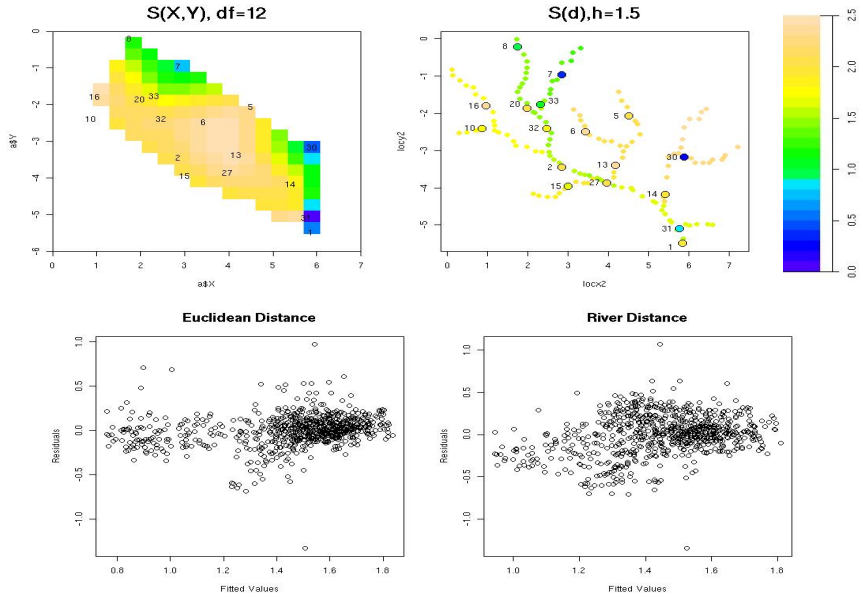

Figure 4.12. $log(TotalN + 1)$ comparison of the smooth function fitted to capture the trend over space and the residuals using Euclidean and river distance

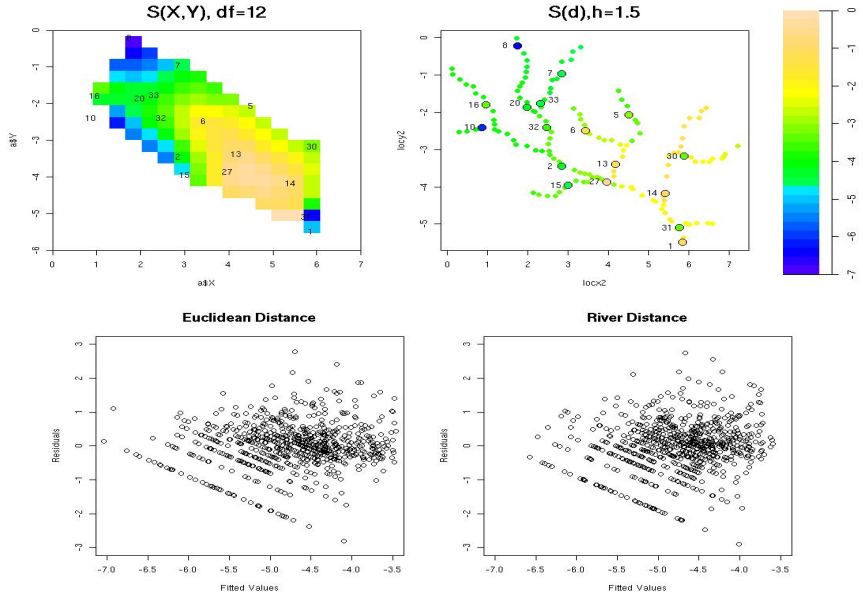

Figure 4.13. $log(PO4.P)$ comparison of the smooth function fitted to capture the trend over space and the residuals using Euclidean and river distance

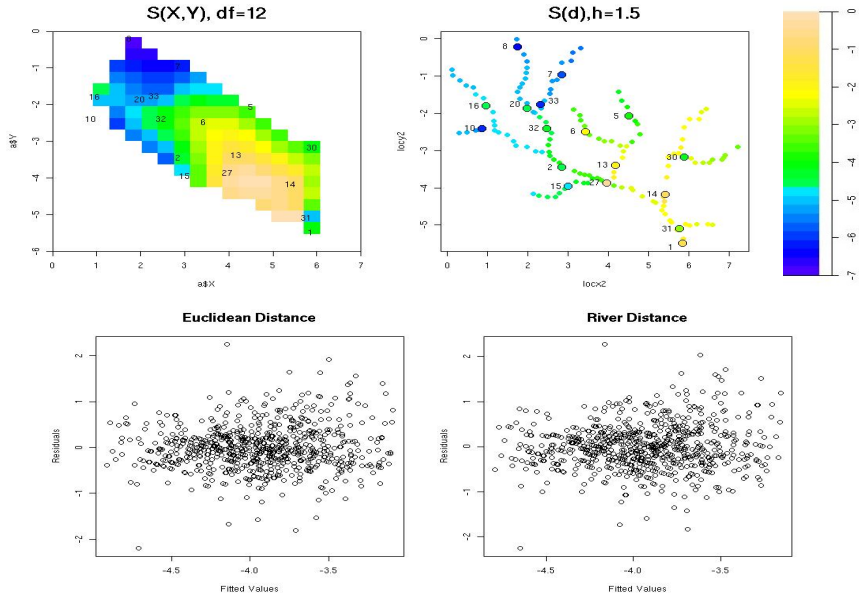

Figure 4.14. $log(TotalP)$ comparison of the smooth function fitted to capture the trend over space and the residuals using Euclidean and river distance

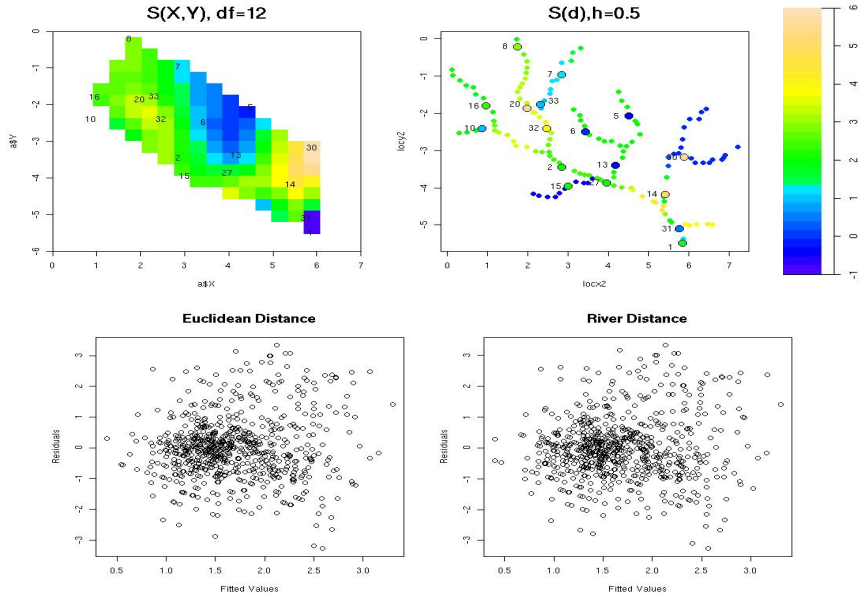

Figure 4.15. $log(SuSo)$ comparison of the smooth function fitted to capture the trend over space and the residuals using Euclidean and river distance

4.5 Summary

The main aim in this chapter was to test a new methodology for modelling trend over space for river networks, adding the structure of the catchment into an additive model indicating which stations are connected and which are not and measuring distance along the river. This provides an approach that allows us to be more accurate, providing a closer representation of the catchment.

Based on the scientific questions established by the Macaulay Institute and the use of river distances to capture the trend over space rather than Euclidean distances, the main findings are:

- For $log(NH4.N)$ there is a fluctuation over the years showing a slightly downward trend with two peaks, the first one in the middle of 2005 and the second one in the middle of 2008, while a trough is observed between 2007 and 2008.

- $log(NO3.N+1)$ showed a fluctuation over the years with a slightly upward trend, with two peaks in the middle of 2005 and the beginning of 2008 and two troughs at the end of 2004 and in the middle of 2007.

- $log(TotalN+1)$ showed a fluctuation over the years, with a slightly upward trend with two peaks in the middle of 2005 and the beginning of 2008 and two troughs at the end of 2004 and in the middle of 2007.

- $log(TotalP)$ showed a clear downward trend until 2007, with a subsequent shift in the trend, reaching a peak in the middle of 2008.

- $log(SuSo)$ showed a clear downward trend until the beginning of 2005 with a diminution in the slope but still downward until the end of 2007, reaching a peak in the middle of 2008.

- The term for trend over space for $log(NH4.N)$, $log(PO4.P)$, $log(TotalP)$, $log(NO3.N+1)$ and $log(TotalN+1)$ capture the variability of the catchment suitably, according to the graphs of the partial residuals and the estimates.

- The comparison between the smooth function for trend, using upstream and Euclidean distances, indicates that for $log(NH4.N)$, $log(PO4.P)$ and $log(TotalP)$ the conclusion is the same indicating a trend in the direction of site 1 and 31 where the highest values are observed. This indicates an increment in the level of these variables in the direction south-east (SE).

- For $log(NO3.N+1)$ and $log(TotalN+1)$ the trend over space is the same indicating higher values in sites 2, 5, 6, 13, 15 and 27. Over the 136 points created, only site 30 shows a high value while the observed value is low. However to be sure about the performance of the model, the value obtained for this site was verified checking the fitted values, where a low value for this site was observed confirming that the model performs well.

- For $log(SuSo)$ the comparison of the average partial residuals values in

respect to the 136 points created, indicates that for this variable the result is not as good as the other variables. Despite that the residuals versus fitted values shows that the model performs well, this model does not allows us to answer properly one of the scientific questions in respect to the trend over space.

- The sensitivity analysis confirms the need for a nonparametric effect to explain the trend over time and space for all the variables.

- The comparison between models using Euclidean and river distance, indicated that for $log(NH4.N)$, $log(PO4.P)$ and $log(TotalP)$, the additive models using river distances are better. For these three variables the models captured the trend over space, the residuals versus fitted value showed that the models fitted well and allowed to provide a better description of the structure of the catchment, adding the fact that not all the points are flow-connected.

- For $log(NO3.N + 1)$ and $log(TotalN + 1)$, the models using Euclidean distance perform better, according to the residuals versus fitted values. These models perform well, capturing the trend over space.

- For $log(SuSo)$ despite that the residuals versus fitted values showed that both models perform well, the model using river distance did not capture the variability of the catchment over space.

Chapter 5

Conclusions and Discussion

5.1 Statistical Methodologies

Throughout this thesis, different statistical methodologies have been presented to analyse environmental data, providing a framework for modelling seasonal patterns and to capture trends over time and space.

One of the main objectives was to provide results for data analysed over time and space simultaneously, rather than take a marginal approach for time and space. The use of time and space as covariates in a single model, allows us to obtain a better understanding of environmental changes, ensuring that the influence of both sources of variability are included.

Each of the methodologies used here have strengths and weaknesses which were evaluated to obtain the best model, looking to reach a closer representation of the variables analysed and providing an adequate answer to the scientific questions.

As a first approach, a linear model is useful to capture trends and seasonal patterns [Esterby (1993)] over time, when a single plane suitably captures the variability of the covariates.

The inclusion of additive models provides an opportunity to relax the linear assumptions [Hastie & Tibshirani (1990)], allowing the fitting of models where the covariates exhibit patterns beyond a linear trend.

Since both methodologies assume that ε_i are independent with $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$, an independence test for time and space [Diblasi & Bowman (2001)] was included as part of the diagnostic check on the residuals. This allows an assessment of the need to include a proper covariance structure over time and space and the possible resulting changes in the conclusions [McMullan et al. (2007)].

The comparison between different models is an important step in the modelling procedure developed in this thesis, allowing a choice to be made between different models. The use of an approximate F-test [Hastie & Tibshirani (1990)] allows the identification of the most suitable model and also the assessment of the need for a linear effect versus a nonparametric effect, ensuring this evidence through a sensitivity analysis.

The use of nonparametric regression provides an alternative to modelling river networks, allowing a closer representation of the variability of the river, using river distance and a connectedness matrix.

5.2 ECN and AWMN

The analysis carried out covers a descriptive analysis and a linear approach to identify differences between variables, assessing the presence of linear trend and seasonal components in both sources of information. The outputs presented throughout this thesis, have identified that there are some differences between the two data sources of information and therefore also the characteristics of the 11 variables analysed.

Throughout this section, limits of detection values (LOD) were included in the analysis, where a value equal to zero was assigned for those observation classified as LOD, according to the method adopted by the ECN and the AWMN. However, it is important to highlight that there are different methods such as survival and replacement analyses [Eastoe et al. (2006)] to cope with this problem, when the main interest is to evaluate trends.

In this particular case would have been better to be able to treat these observation as censored data through a survival analysis rather than assign a value equal to zero, mainly for variables such as $log(Fe + 0.5)$, $log(NO_3 + 0.5)$, and $log(Al + 0.5)$ in the AWMN data and $log(Cl + 3)$ in the ECN data. The reason to suggest the survival analysis rather than the replacement analysis, obey to the fact that under the presence of high number of LOD, the survival analysis has shown a better performance [Eastoe et al. (2006)].

The comparison between both sources of information through the descriptive analyses, indicated higher variability in the ECN data. This difference may be explained by the collection process with only one observation per month for the AWMN, while the ECN data collected more observations per month, allowing us to include the variability over the years and also the variability within the month.

The fact that the information was collected at different locations might help to explain the higher variability, indicating different physical characteristics for the 11 variables. In the specific case of $log(Fe+0.5)$, $log(NO_3+0.5)$ and $log(Al+0.5)$ in the AWMN data, the lower variability can perhaps be explained by the presence of limit of detection values.

The Bland-Altman plots provided evidence that only pH, $log(DOC)$ and $log(SO_4(S)+1)$, show a good level of agreement, while for the rest of the variables the ECN data showed higher values than the AWMN.

A linear approach provided a suitable result to identify trends over time along with a sine/cosine term to capture seasonal patterns. The analysis of the parameters for trend and seasonality allowed us to confirm the difference between both data sets, indicating that only $log(Ca + 2)$ and $log(Al + 0.5)$, showed the same conclusion in respect to the parameters for trend and seasonal components. However, only for $log(Al + 0.5)$ the parameters for trend are statistically equal.

The analysis of the parameter for trend allowed us to establish that despite being significantly different from zero, the trend observed in the variables over time is weak.

5.3 Tarland Catchment

The use of time and space simultaneously rather than a marginal approach separating time and space, allowed a closer representation for the catchment, including both sources of variability.

The first approach through a linear model gave a good performance. However for variables such as $log(NO3.N + 1)$, $log(TotalN + 1)$ and $log(SuSo)$, the trend over space was not properly captured, indicating the need for a different approach.

The use of additive models to capture trends over time and space simultaneously, allows the assumptions of linear models to be relaxed. The flexibility of additive models allowed non-linear patterns over time and space to be captured, providing a closer representation of the variability of the catchment using year, day and the coordinates of each site as covariates.

The question of the best value for the degrees of freedom is still a matter of discussion and research nowadays. The existing automatic methods tend to be

less reliable and far more computationally expensive to implement, mainly when several degrees of freedom must be selected simultaneously [Hastie & Tibshirani (1990)]. In addition, these methods are not suitable for information collected over time and/or space with a temporal [Hart (1991)] or spatial correlation structure.

The selection of the degrees of freedom throughout this thesis was carried out looking for flexibility from a linear shape and to avoid overfitting the data, relying on a graphical approach where the assessment of the partial residuals, allows identification of a suitable value for each covariate. The models fitted showed a good performance capturing the trend over time and space for all six variables, confirming the need for a nonparametric effect through the sensitivity analysis.

Regarding the assumption of independence of the residuals of the additive models, the need to incorporate correlated errors over time and/or space was assessed, and there was no evidence that this was required. Only $log(TotalP)$ showed weak evidence indicating the need of a covariance structure over space, although according to the results the best model to explain the variability of the residuals was a pure nugget effect.

The independence test applied [Diblasi & Bowman (2001)], was developed to assess correlation over space for a single sample, although the test also works as a diagnostic check for models assuming independence for the residuals. In the particular case of the Tarland catchment, the test was used to assess independence over space, based on the conclusion of no autocorrelation over time.

As part of the modelling process the river flow information was included to assess the effect of this variable. These models were fitted for a shorter period of time according to the data available for the flow river information obtained for site 1. The results showed an improvement in those models which include flow information, although the sensitivity analysis has identified that a semi-parametric

model provides a better approach for $log(NO3.N + 1)$ and $log(TotalN + 1)$.

The overall conclusions in respect to trend over time and space were obtained from the additive models using year, day and the coordinates of each site. This corresponds to information collected regularly at 17 sites from April 2004 to November 2008.

The analysis of the smooth functions confirms a nonlinear trend fluctuating over time exhibiting peaks and troughs. Variables such as $log(NH4.N)$, $log(PO4.P)$, $log(TotalP)$ and $log(SuSo)$ showed a downward trend while $log(NO3.N+1)$ and $log(TotalN + 1)$ showed a slightly upward trend.

The analysis of the trend over space indicated that $log(NH4.N)$, $log(PO4.P)$ and $log(TotalP)$, presented a clear trend in direction of site 1 and 31 where the highest values are located. The increment in the concentration level for these three variables is progressive, indicating a clear pattern in direction south-east (SE).

Variables such as $log(NO3.N + 1)$ and $log(TotalN + 1)$ showed a higher concentration in the large majority of the catchment, showing higher values in sites 2, 5, 6, 13, 15 and 27 and lower values in sites 7, 8 and 33. For $log(SuSo)$ the higher concentrations are in sites 20, 30 and 33 with lower values in sites 5, 6 and 13.

This reveals similar patterns between variables allowing us to gather them in two groups in respect to the trend over time. Group one corresponds to $log(NH4.N)$, $log(PO4.P)$, $log(TotalP)$ and $log(SuSo)$ showing a downward trend, although the peaks and troughs are not located at the same dates. A second group corresponds to $log(NO3.N + 1)$ and $log(TotalN + 1)$ showing a similar behaviour with a slightly upward trend.

In the same way but over space three groups can be observed. The first group corresponds to $log(NH4.N)$, $log(PO4.P)$ and $log(TotalP)$, showing a similar trend in direction south-east (SE). A second group corresponds to $log(NO3.N+1)$ and $log(TotalN+1)$ with similar behaviour and finally $log(SuSo)$ exhibit a different trend in respect to the other 5 variables with two peaks (sites 20,30 and 33) and a trough (sites 5, 6 and 13).

The advantages to use additive models to capture the trend over time and space is clear making it easier to fit a model to variables with non-linear patterns. Despite that the linear model approach indicated a similar conclusion in respect to a positive or negative trend, the additive model allowed to obtain the same conclusion for the trend, reproducing better the peaks and troughs observed over time.

The improvement to capture the trend over space is also clear, mainly for $log(NO3.N+1)$, $log(TotalN+1)$ and $log(SuSo)$, where the use of a smooth function fitted as a surface is able to capture the trend suitably, allowing us to get a closer representation.

5.4 Modelling of River Networks

The modelling of river networks comes with a lot of questions about the best way to include the structure of the river into a spatial model. The existing methodologies tackle the problem through the design of new spatial models, using Euclidean, river distance or both, adding a weighting to ensure a valid covariance structure [Ver Hoef et al. (2005)] , [Cressie et al. (2006)].

The approach used in this thesis adopted this idea using a nonparametric regression (local mean estimator), allowing the structure of the river to be included

into an additive model using river distance rather than Euclidean distance. The inclusion of a connectedness matrix n x n, with n the number of sampling points in the river network, allowed us to get a closer representation of the catchment, by including which sites are flow-connected.

The choice of h for the smooth function to capture the trend over space, was made with a view to maintain a balance between smoothness and goodness-of-fit, assessing the performance with new points to reproduce the shape of the catchment and providing a graphical result of the partial residuals.

This approach to modelling river networks has several advantages, since that the smooth function for the upstream distance is fitted under the framework of an additive model. This allows us to assess the effect of a linear effect versus a nonparametric effect and to assess the sensitivity of the test under different values of degrees of freedom.

The comparison between Euclidean distance and river distance, indicated that for $log(NH4.N)$, $log(PO4.P)$ and $log(TotalP)$, the additive models using river distance are better based on how these models capture the trend over space, an absence of evidence that indicates that the models do not fit well and the fact that they provide a closer representation of the catchment, by including which sites are flow-connected.

For $log(NO3.N + 1)$, $log(TotalN + 1)$ and $log(SuSo)$, the additive models using Euclidean distance perform better. For $log(SuSo)$, the trend over space is captured better using Euclidean distance rather than river distance.

This type of model offers a new tool to tackle the analysis of river networks to obtain a closer representation, by accounting for the fact that the variability of river networks requires a different approach, since that not all the points are

flow-connected and therefore the effect of each sampling point does not affect the variability of the river in the same way.

5.5 Suggestions

The main aim of this section is to provide recommendations based on the results obtained throughout this thesis, suggesting better practices or simply some ideas that could lead to the implementation of strategies, to improve the quality of the environmental variables analysed.

For the ECN and AWMN data, one of the main reasons to support the differences between both sources of information was the differences in the procedure to collect the data. This problem is common and it is clear that is impossible to minimise the variability of external factors. However, it is important to indicate that if it is possible to reduce the variability generated by the sampling process, collection frequency or issues related to how the outcome of interest was measured, the results obtained could lead to better understanding of the information analysed.

Based on the results presented throughout this analysis for the Tarland catchment, it is important to define a strategy to evaluate how the levels of NO3.N and Total N can be reduced in the catchment, given the slightly upward trend observed over time and the presences of high values in the large majority of the sampling stations. For NH4.N, PO4.P and Total P, the results indicate a downward trend over time with a progressive trend over space in direction south-east (SE). The implementation of agreements to protect locations in direction north-west (NW) could lead to an improvement of the water quality of the catchment and a reduction in the levels observed nowadays.

5.6 Further Work

The work presented through this thesis provided a set of tools to identify trends over time and space simultaneously, allowing changes in environmental data to be assessed. Looking to improve the modelling process to understand better the complex systems exhibited in the environment, there are two main ideas to be considered as the next step in the modelling of river networks.

According to the ideas developed by Ver Hoef et al. (2005) and Cressie et al. (2006), the distance to be used is an important factor to ensure a proper model, although the weight assigned is the key part to ensure a validate covariance structure. The use of different weights added to the additive model could provide a closer representation of river networks, where river order, basin area or flow are some of the possibilities suggested.

The assumption of independence over the residuals in the particular case of the Tarland catchment was assessed over all the variables, indicating there was no need to include a covariance structure for time and space. The analysis of bigger and more complex river networks could lead to the need for a covariance structure over space, where the traditional spatial models do not work properly. This represents an opportunity to use different spatial models [Ver Hoef et al. (2005), Cressie et al. (2006)] to include correlation over space and at the same time to fit models to capture the trend, taking advantage of the flexibility of additive models.

Bibliography

AWMN(Webpage) (n.d.), Electronic Resource [Accessed 30/04/2009].
URL: *http://www.ukawmn.ucl.ac.uk/*

Azzalini, A. & Bowman, A. (1993), 'On the uses of nonparametric regression for checking linear relationships', *Journal Royal Statistics Society Ser B 55, 549-557* .

Bates, B., Kundzewicz, Z., Wu, S. & Palutikof, J. (2008), 'Climate change and water', *Technical Paper of the Intergovernmental Panel on Climate Change, IPCC Secretariat, Geneva, 210 pp.* .

Bland, J. & Altman, D. (1986), 'Statistical methods for assessing agreement between two methods of clinical measurement', *Lancet 1, 307-312* .

Bowman, A. & Azzalini, A. (1997), *Applied Smoothing Techniques for Data Analyis: the kernel approach with S-Plus illustrations*, Oxford University Press.

Bowman, A., Giannitrapani, M. & Scott, E. (2009), 'Spatialtemporal smoothing and sulphur dioxide trends over europe', *Royal Statistical Society Series C 58, Part 5, pp 737-752* .

Bowman, A. & Young, S. (1996), 'Graphical comparison of nonparametric curves.', *Applied Statistics 45, 83-98* .

Clement, L., Thas, O., Vanrolleghem, P. & Ottoy, J. (2006), 'Spatial-temporal statistical models for river monitoring networks', *Water Science and Technology Vol 53 No 1 pp 9-15* .

Cothern, C. & Ross, N. P. (1994), *Environmental Statistics, Assessment and Forecasting*, CRC, Press.

Cressie, N. (1993), *Statistics for Spatial Data, revised Edition*, John Wiley and Sons.

Cressie, N., Frey, J., Harch, B. & Smith, M. (2006), Spatial prediction on a river distance, Technical report, The Ohio State University.

Cressie, N. & Hawkins, D. (1980), 'Robust estimation of the variogram: I', *Mathematical Geology, 12 (2), 115-125* .

Diblasi, A. & Bowman, A. (2001), 'On the use of the variogram in checking for independence in spatial data', *Biometrics 57, 211-218* .

Eastoe, E. F., Halsall, C. J., Heffernan, J. E. & Hung, H. (2006), 'A statistical comparison of survival and replacement analyses for the use of censored data in a contaminant air database: A case of study from the canadian arctic', *Atmospheric Environment 40, 6528-6540* .

ECN(Webpage) (n.d.), Electronic Resource [Accessed 22/04/2009].
URL: *http://www.ecn.ac.uk/*

Esterby, S. (1993), 'Trend analysis methods for environmental data', *Environmetrics, 4(4), 459-481* .

Esterby, S., Shaarawi, A., Keeler, L. & Block, H. (1991), 'Testing for trend in water quality monitoring data', *National Water Research Institute Contribution No. 91-02.* .

Eubank, R. (1999), *Nonparametric Regression and Spline Smoothing*, Marcel Dekker, New York.

European Commission (1967), 'Council directive 67/548/eec of 27 june 1967 on the approximation of laws, regulations and administrative provisions relating

to the classification, packaging and labelling of dangerous substances.', *Official Journal L 129 , 18/05/1976 P. 0023 - 0029* .

European Commission (1991*a*), 'Council directive 91/271/eec of 21 may 1991 concerning urban waste-water treatment', *Official Journal L 135, 30.5.1991, p. 40–52* .

European Commission (1991*b*), 'Council directive 91/676/eec of 12 december 1991 concerning the protection of waters against pollution caused by nitrates from agricultural sources', *Official Journal L 375 , 31/12/1991 P. 0001 - 0008* .

European Commission (2000), 'Directive 2000/60/ec of the european parliament and of the council of 23rd october 2000 establishing a framework for community action in the field of water policy', *Official Journal 22 December 2000 L 327/1, Brussels: European Commission.* .

Fan, J. & Gijbels, I. (1996), *Local polynomial modelling and its applications*, Chapman and Hall: London.

Giannitrapani, M., Bowman, A. & Scott, E. (2005), Additives models for correlated data with applications to air pollution monitoring, Technical report, University of Glasgow.

Graham, B. & McBride (1948), *Using Statistical Methods for Water Quality Management*, John Wiley and Sons.

Green, P. & Silverman, B. (1994), *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman and Hall: London.

Hart, J. (1991), 'Kernel regression estimation with time series errors', *Journal Royal Statistics Society Ser B 53, 173-87* .

Hastie, T. & Tibshirani, R. (1990), *Generalized Additive Models*, Monographs on Statistics and Applied Probability 43, Chapman and Hall.

Hawkins, D. & Cressie, N. (1984), 'Robust kriging- a proposal', *Journal of the International Association of Mathematical Geology 16, 13-18* .

Hirst, D. (1998), 'Estimating trends in stream water quality with a time-varying flow relationship', *Austrian Journal of Statistics 27: 39-48* .

McMullan, A., Bowman, A. & Scott, E. (2007), 'Water quality in the river clyde: A case study of additive and interaction models', *Environmetrics* .

M.I.(Webpage) (n.d.), Electronic Resource [Accessed 30/04/2009].
URL: *http://www.macaulay.ac.uk/*

Nadaraya, E. (1964*a*), 'On estimating regression', *Theory Probab. Appl. 10, 186-90* .

Nadaraya, E. (1964*b*), 'Some new estimates for distribution functions', *Theory Probab. Appl. 9, 497-500* .

Pinheiro, J. & Bates, D. (2000), *Mixed-Effects Models in S and S-PLUS*, New York; London:Springer.

Pinheiro, J. & Diggle, P. (2009), *Analysis of geostatistical data*, 1.6-26 edn.

Ruppert, D., Wand, M. & Carroll, R. (2003), *Semiparametric Regression*, CUP, London.

Venables, V. & Smith, D. (2009), *An Introduction to R*, R Development Core Team.

Ver Hoef, Peterson, E. & Theobald, D. (2005), 'Spatial statistical models that uses flow and stream distance', *Environ Ecol Stat 13:449-464* .

Watson, G. (1964), 'Smooth regression analysis', *Sankhya, Ser. A, 26, 359-72* .

Webster, R. & Oliver, M. A. (2007), *Geostatistics for Environmetal Scientists*, John Wiley and Sons.

Whittaker, J. (1990), *Graphical models in applied multivariate statistics*, Wiley, Chichester, UK.

Wood, S. (2006), *Generalized Additive Models: An Introduction with R*, Chapman and Hall/CRC: London.