ir. Lieven Clement

Statistical validation and spatio-temporal modelling of river monitoring networks

Thesis submitted in fulfilment of the requirements for the degree of
Doctor (Ph.D.) in Applied Biological Sciences

to Fien, Wiebe, Pepijn and Jacob

Examination Committee
ir. Michiel Blind (RIZA, the Netherlands)
Prof. Dr. Adrian Bowman (University of Glasgow, UK)
Prof. Dr. Anders Grimvall (Linköping University, Sweden)
Prof. Dr. ir. Marc Van Meirvenne (Ghent University)
Prof. Dr. Stijn Vansteelandt (Ghent University)
Prof. Dr. ir. Nico Verhoest (Ghent University)
Prof. Dr. ir. Walter Steurbaut (Ghent University, chair)

Promotors
Prof. Dr. ir. O. Thas & Prof. Dr. ir. P. A. Vanrolleghem
Department of Applied Mathematics, Biometrics and Process Control, Ghent University

Dean
Prof. Dr. ir. H. Van Langenhove

Rector
Prof. Dr. P. Van Cauwenberge

ir. Lieven Clement

Statistical validation and spatio-temporal modelling of river monitoring networks

Thesis submitted in fulfilment of the requirements for the degree of
Doctor (Ph.D.) in Applied Biological Sciences

Dutch translation of the title:
Statistische validatie en ruimtelijk-temporele modellering van riviermonitoringnetwerken

Please refer to this work as follows:
Lieven Clement, 2007. Statistical validation and spatio-temporal modelling of river monitoring networks. Ph.D. thesis, Ghent University, Gent, Belgium.

ISBN-number: 978-90-5989-179-1

Dankwoord

Over hoe de aanloop naar dit proefschrift een verhaal werd van heel veel spelers. Over een verhaal dat begon als een sprong van de hak op de tak om uiteindelijk opnieuw overtuigend te eindigen op de hak. En zoals het verhaal, zo ook de structuur van dit dankwoord. Het ultieme bewijs dat je alweer gelijk hebt, Olivier. Ik denk aan te veel tergelijkertijd en wil dan al die spinsels wanhopig in één zin neerschrijven.

De aanloop naar de hak
Onderzoekertje spelen zat er bij mij steeds in.
Tot groot jolijt, en af en toe tot grote ergernis,
heb ik mijn omgeving
al van in mijn vroege jeugd
met vragen gebombardeerd.

Deze drang naar kennis
werd het eerst gekanaliseerd
door Antoon Verelst.
In het secundair legde hij mijn fundamenten
in de wetenschappen.

Op de unief
vond ik in LabMET mijn tweede mentor.
Bij prof. Verstraete en prof. Top
leerde ik hoe intrigerend en creatief
wetenschappelijk onderzoek kan zijn.

2000: De hak
Een verhaal dat startte met de VMM.
Bedankt voor het aanbrengen van de problematiek,
de financiering en de data.
Statistiek werd zo mijn doel
en daar belandde ik dan als groentje
op BIOMATH. Al snel maakte ik kennis
met het enthousiasme en de mens
achter prof. Peter Vanrolleghem.

Mijn privé was toen één grote puinhoop
en dat had Peter al snel in het oog.
Hij gaf me in die eerste weken de ruimte
om mijn leven ook buiten de universiteit
weer op de rails te zetten.
Het resultaat van die turbulente weken mag er wezen.
Fien en ik vonden toen elkaar.
Twee werd drie dus samen vijf.

2000-2001: Tussen hak en tak
In mijn eerste maanden
had ik af en toe heimwee
naar het werken in een labo.
Samen met Peters onstuitbare enthousiasme
vertaalde dat zich in de aanvraag van een beurs.

Prof. Olivier Thas,
toen nog mijn inspirerende bureaugenoot,
wijdde me ondertussen in
in de beginselen van de statistiek.
Hoe verder het jaar vorderde,
hoe meer plezier ik kreeg
in het statistische spel van analyses.

En die besmetting van de statistiekmicrobe
veranderde ook stilaan mijn wereldbeeld.
Gedaan met het denken in zwart-wit,
bij elke uitspraak komt sindsdien
die onvermijdelijke nuance.

2001: Op de tak
Met spijt in het hart
nam ik in oktober 2001
afscheid van de statistiek.

Een nieuwe uitdaging lag te wachten:
Een BOF-mandaat bij prof. Vanrolleghem en prof. Sorgeloos.
Maar de rotiferen hadden het niet op mij begrepen.
Ze gingen dood nog voor ik me vertoonde.

2003: En nu voorgoed op de hak
In 2003 werd ik assistent.
En daar lonkte het statistische beestje weer.

Opnieuw werd het een oefening tussen twee promotoren.
Prof. Peter Vanrolleghem van BIOMATH
en Prof. Olivier Thas van BIOSTAT.
Bedankt voor de kennis, de vrijheid, het plezier en alle kansen.
Als ik terugblik op mijn doctoraatsperiode,
is het eigenlijk een uit de hand gelopen hobby geworden.

Peter,
terug kon ik bij jou terecht,
nu voor de link met de praktijk.

Olivier,
mijn mentor in de statistiek,
bedankt voor je talloze ideeën,
de vele boeiende en verhelderende gesprekken,
je gevoel voor humor,
en af toe de nodige portie hilariteit.
Onze zoektocht in Southampton
en onze dolle vijfdaagse in Pamplona
zal ik niet zo snel vergeten.
Ik kijk alvast uit naar onze nieuwe episode.

De sociale ‘glue’ op de vakgroep
De eerste jaren was ik nogal dikwijls op de ‘move’.
Heel wat bureaus heb ik toen versleten.
Memorabel waren de momenten
tijdens mijn verblijf bij Ellen, Heidi en Olivier.

Daarna stockeerde Peter me maandenlang
met Veronique, een hub en tussen stapels dozen
in de ‘cartonage’ van Mie.
Dat werd werken, lachen en af en toe
op ontdekkingstocht tussen al die relikwieën.

De beruchte ‘incubator room’,
daar heb ik ook gezeten.
Gehersenspoeld door Guru Gazza
kon ik weer op doortocht,
maar nu als volleerde linux-nerd.

Een tussendoortje met de ‘Klis’
‘Yu’, ‘Yu’, ‘Yu’
Chinese vis, dat is niet mis.

Ten gepaste tijden,
gezever en gezwets,
het spastische ontstressen.
Bram en mijn nieuwe buren Peter P. en Petra,
die moesten het nu ontgelden.

Het P-team met zijn fuiven, zijn weekends, cocktails, ...
Ester die me redde met de TEX-templates.
Ellen en Heidi met hun kunst- en vliegwerk op diverse vlakken.
En zo kan ik uren doorgaan.

Kortom iedereen
van secretariaat,
KERMIT,
BIOMATH
tot BIOSTAT: dankjewel voor de schitterende tijd!

'De achterban'
Papa, mama,
Katrijn en Mieke,
de kansen,
de warmte en de inspirerende omgeving
hebben zeker de kiem gelegd
voor het voltooien van dit werk.

Frie en Dirk,
Pieter en Melanie,
jullie gaven Fien en mij de kans
om er in de drukke tijden
zo nu en dan eens
tussenuit te knijpen.

Wiebe, Pepijn en Jacob,
het geluid van jullie
kleine trippelende kindervoetjes
deden me niet vergeten
dat er naast het werk
een leven is.

En Fien,
wat moet ik zeggen?
Je combineerde
zowat alles tegelijk
en alles bleef maar draaien.
Als mama, vrouw en luisterend oor,
mijn wandelende agenda,
styliste in tijden dat ik niets meer zag,
minnares en beste moàtje, ...
ik heb je lief.

Lieven Clement
Gent, juni 2007

Contents

List of Abbreviations & Symbols

AIC	Akaike Information Criterion
AICi	improved Akaike Information Criterion
AM	Additive Model
aPI	analytical Prediction Interval
AR(1)	AutoRegressive process of order 1
ARIMA	AutoRegressive Integrated Moving Average
ARMA	AutoRegressive Moving Average
ARX	AutoRegressive Exogenous
BC_a	Acceleration and Bias Corrected
BLUP	Best Linear Unbiased Predictor
%bPI	percentile based bootstrap Prediction Interval
CI	Confidence Interval
CV	Cross Validation
COD	Chemical Oxygen Demand
df	degrees of freedom
df^{err}	degrees of freedom of the errors
DAG	Directed Acyclic Graph
DO	Dissolved Oxygen
EC	European Commission
EM	Expectation Maximisation
ECM	Expectation Conditional Maximisation
EU	European Union
EWMA	Exponentially Weighted Moving Average
FIM	Fisher Information Matrix
FGLS	Feasible Generalised Least Squares
GCV	Generalised Cross Validation

GLM	Generalised Linear Model
GLMM	Generalised Linear Mixed Model
GLS	Generalised Least Squares
i.i.d.	Independent & Identically Distributed
ICT	Information and Communications Technology
LCL	Lower Confidence Limit
MA	Moving Average
MLE	Maximum Likelihood Estimator
MSE	Mean Squared Error
MISE	Mean Integrated Squared Error
MVN	Multivariate Normal Distribution
MQL	Marginalised Quasi Likelihood
NLT	Nonlinear Trend
NO_2^-	Nitrite
NO_3^-	Nitrate
OLS	Ordinary Least Squares
PCA	Principle Component Analysis
PI	Prediction Interval
PQL	Penalised Quasi Likelihood
RSS	Residual Sum of Squares
sbPI	studentized prediction error based Prediction Interval
SEM	Structural Equation Model
T	Temperature
$\mathrm{tr}(.)$	trace of a matrix
UCL	Upper Confidence Limit
WFD	Water Framework Directive
w.r.t	with respect to
$\perp\!\!\!\perp$	Independent

Chapter 1

Introduction

1.1 Setting

The European Water Framework Directive (WFD)(EC, 2000) is one of the driving forces in environmental policy in the European Union (EU). The WFD's overall environmental objective is the achievement of a 'good status' for all of Europe's surface- and ground waters within a 15-year period. Its implementation is a big challenge for the European environmental managers. The WFD triggered the water authorities to design monitoring programmes. Thus, large amounts of environmental data are being collected, processed and stored throughout Europe. They are for instance needed for a coherent and comprehensive overview of the water status, to identify pressures on water systems, as a warning system for detecting negative changes in the water quality and to detect trends. Like other environmental data, water quality data have a complex nature. They contain a considerable amount of

noise, due to their natural variability and the measurement error. They often contain missing values, are often censored due to the detection limits of the measuring methods, and are commonly gathered on irregular time intervals. They also may be mutually dependent, non-normally distributed, possess cyclic variations and show nonlinear trends (e.g. Hirsch et al., 1982, Van Belle and Hughes, 1984, Cai and Tiwari, 2000 and McMullan, 2004).

Due to the large amount of data and their complex nature, modelling has become an essential tool to extract information from these observations. Within the research community, monitoring and modelling have now become generally accepted to be interlinked activities (e.g. Parr et al., 2003; Højberg et al., 2007). From this perspective, models can be used for a number of different purposes. For instance, they can be useful to assure data quality, for inter- and extrapolation in time and space, to increase the conceptual understanding of the underlying processes, to evaluate the impact of (future) management strategies, to assess the effect of anthropogenic activities and to design monitoring programmes (Højberg et al., 2007).

High quality data is essential for an adequate management of the water resources. Therefore, quality assurance is specifically mentioned as an important activity in the WFD guidance document on monitoring (EC, 2003; Højberg et al., 2007). Thus before the data can be used in an assessment, they have to be validated. Errors might be introduced during the analysis in the laboratory, wrong calibration of the equipment or while entering the data. It is, however, also possible that there is a change in the system that causes changes in the water quality. The purpose of the validation procedure is thus twofold: it should act as a tool to provide a quality check and as a warning system to detect negative changes. The large amount of water quality data and its complex nature, however, make it difficult for the environmental agencies to validate all incoming data. An ICT tool could be of great help to assist experts with the maintenance of monitoring databases compelled by the WFD. Such a tool should be able to deal with the complex nature of the water quality data and it also should be adaptive because the environmental system is likely to change, e.g. due to more stringent environmental legislation.

Once the environmental agencies have a consistent database at their disposal, the data should be used to assess the evolution of the water status and to evaluate the impact of their management strategies. Such an assessment should be possible at the level of individual sampling locations as well as on a more regional scale. Many classical statistical techniques cannot be used for these purposes because data originating from environmental monitoring networks are clearly not independent. They

are sampled from a dynamic process that evolves over space and time. Therefore, the methodology should incorporate this spatio-temporal dependence structure in order to provide valid statistical inference. Until recently, researchers mainly focussed on the assessment of water quality at the level of the individual sampling locations. There have been some attempts in the past to provide techniques to perform an analysis on a spatial scale, but they used rather ad hoc methods to account for the spatial dependence. Only the last couple of years spatio-temporal models have been developed to take the specific spatio-temporal dependence structure of river networks explicitly into account (Gardner et al., 2003; Monestiez et al., 2005; Cressie et al., 2006; Ver Hoef et al., 2006). But they are all related to spatial prediction in river networks. Our aim, however, is to enable an assessment on the data that is observed at the sampling locations. Therefore the observations of the monitoring network at a certain time instant can be considered as the realisation of a finite-dimensional multivariate random variable with each dimension corresponding to each of the p sampling locations. Here, the spatio-temporal dependence also has to be taken into account to provide valid statistical inference.

Both the data validation problem and the development of spatio-temporal models for river networks have become the major themes of this dissertation. Before we give the outline of this dissertation, we will introduce the data that were used throughout the work to test and illustrate the developed methodology.

1.2 Introduction to the data used in this study

In the region of Flanders (Belgium), the Flemish Environmental Agency (VMM) established several monitoring networks. Their physico-chemical monitoring network was established in 1989 and now covers 1425 sampling locations distributed over the different catchments of Flanders. Each sampling location is evaluated 12 to 26 times a year on a basic spectrum of physico-chemical variables: water temperature, dissolved oxygen (DO), pH, chemical oxygen demand (COD), nitrogen compounds, phosphorus, chloride and conductivity. All these data are stored in a database, which is also managed by the VMM. The data can be classified according to the catchment it belongs to. One of the catchment area's is the Yzer basin. The data of this catchment is considered in this dissertation.

The Yzer is a typical lowland river, located in a polder area. A map of the Yzer catchment indicating the sampling locations maintained by the VMM is given in

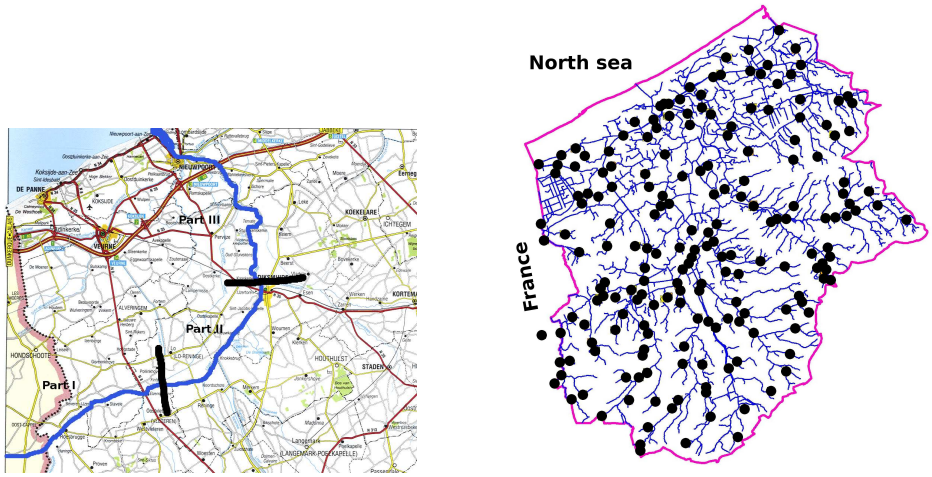

Figure 1.1: The Yzer catchment. In the left panel the main river is shown and the three parts are indicated. In the right panel the entire catchment is given along with the sampling locations of the VMM (indicated with black circles)

Figure 1.1. The total area of the catchment is 1101 km^2. Its spring and one third of the catchment is located in France, two thirds are located in Belgium. The stream length is 76 km and 44 km of it is located in Belgium. At the French border the river is relatively narrow, between 8 to 10 m. The river gets gradually wider to reach a width of 20 to 25 m near to its mouth at Nieuwpoort, Belgium. The river enters the North Sea by a complex of sluices. In Belgium, the river can be subdivided in 3 major parts. Part I is an area where the river is more or less in its original state. In part II, the river is straightened and has marshes to its right side. In part III, the river has artificial dammed banks (De Rycke et al., 2001). The Yzer is used for the production of potable water, so that the water should meet the standards for this production. However, the river is subject to eutrophication due to the high nitrate and phosphate concentrations originating from intensive agricultural activity. Besides the agricultural pollution, other sources are from an industrial origin and from untreated sewage discharged by households.

In most chapters we will illustrate our methods on five sampling locations where a considerable amount of data is available. The five sampling locations are located along two joining river reaches.

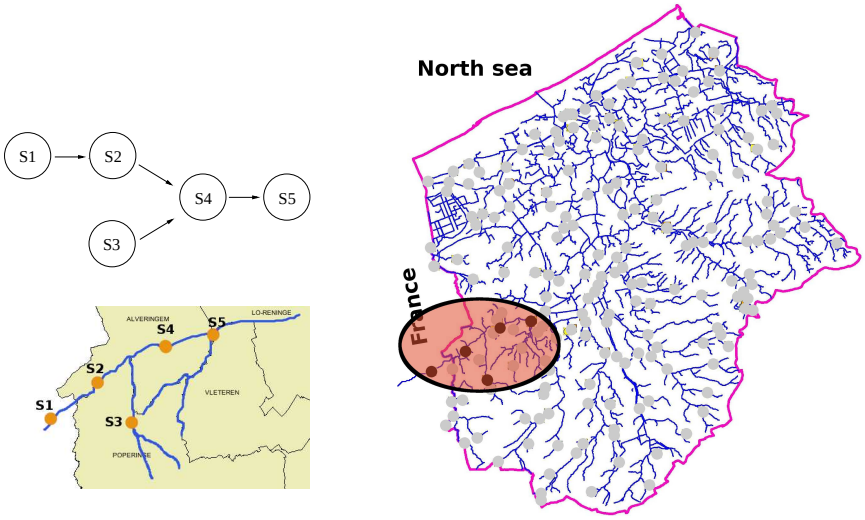

Figure 1.2: Top Left: Network topology of the sampling locations. Bottom Left: Map of the river reaches considered in this case study. Locations S1, S2, S4 and S5 are located on the Yzer river while location S3 is located on a joining creek. Right: Map of the part of the Yzer catchment located in Flanders, Belgium. The area considered in this study is indicated with the ellipse and the five sampling locations are indicated with black dots

Sampling locations S1, S2, S4 and S5 are located on the Yzer while sampling location S3 is located on a joining creek. Their river network topology and locations in the catchment are shown in Figure 1.2. Monthly observations are available between January 1990 and August 2004.

1.2.1 Data exploration

The nitrate series at each sampling location is presented in Figure 1.3. From the measurements it can be seen that a number of observations are missing. For example all observations of 1995 are missing at the sampling locations of the main river, and the observations between January 1996 and November 1997 are missing at sampling location S3. To enable the use of methods that cannot handle missing data, the nitrate series was augmented with simulated data. To simulate the missing

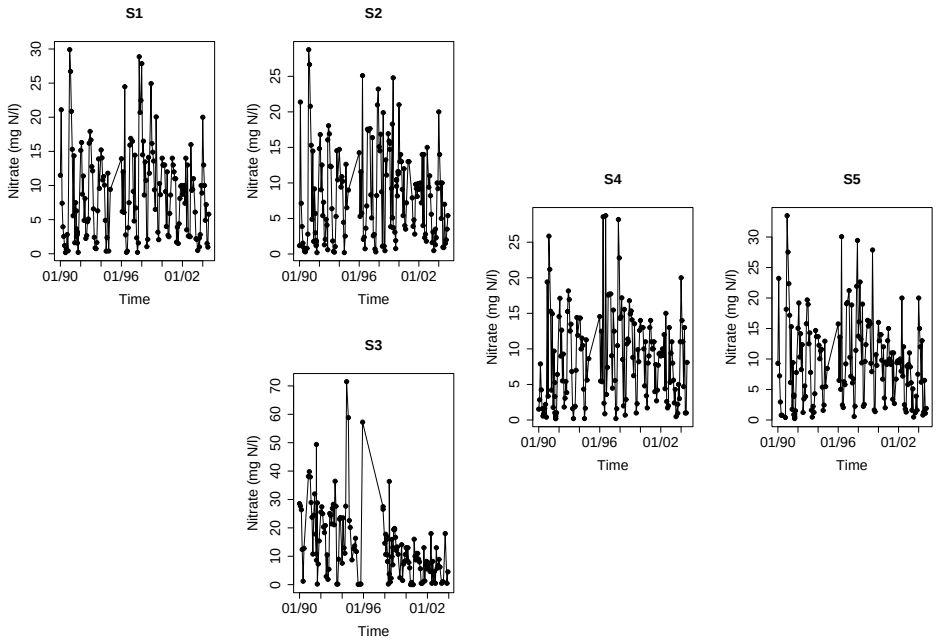

Figure 1.3: Nitrate observations at five sampling locations of the river Yzer. Sampling locations S1, S2, S4, S5 are located on the Yzer river while sampling location S3 is located on a tributary

data an additive model that consists of a trend component and a seasonal effect is used. An introduction to additive models can be found in Chapter 2. Simulated observations y_t^s are generated by using the prediction of the additive model $\hat{y}_t$ and by adding a random residual $\hat{e}_t^*$ to it, $y_t^s = \hat{y}_t + \hat{e}_t^*$. The augmented dataset is presented in Figure 1.4. When this augmented dataset is used, we ignore that missing data was present and we act as if all the data from the augmented dataset was observed.

An interesting method for a first examination of the water quality is the use of the loess scatterplot smoother (Cleveland and Grosse, 1991; McMullan, 2004). The loess smoother is based on local polynomial regression and a more detailed description can be found in Section 2.2.3. Cleveland et al. (1990) developed a loess based method to decompose the data into a seasonal (S), a trend (T) and a residual (R) component. They referred to it as the STL procedure. Suppose that the re-

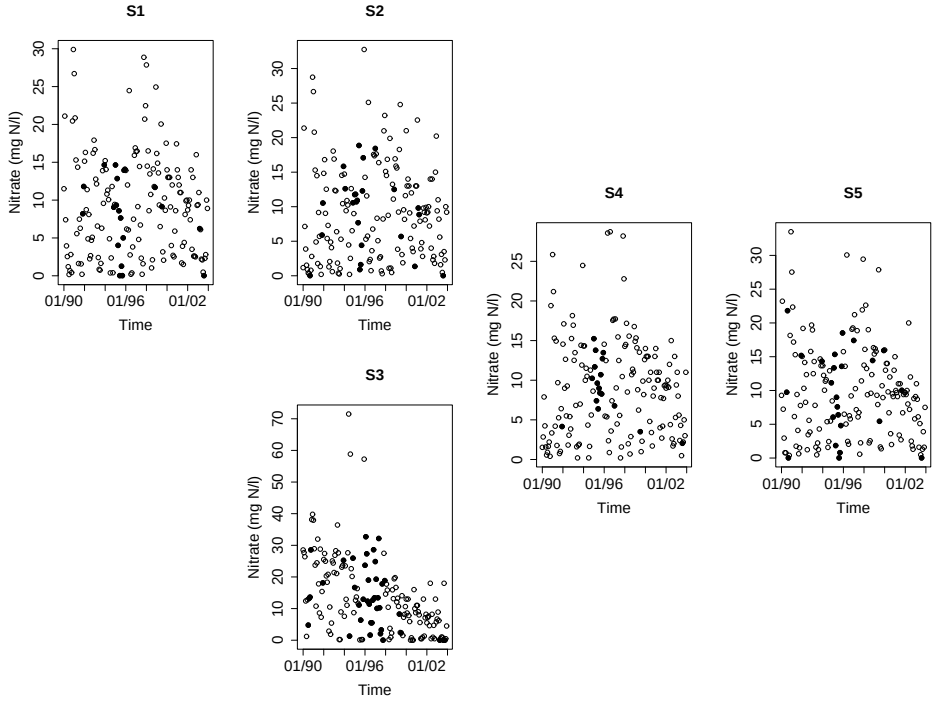

Figure 1.4: Nitrate observations at five sampling locations of the river Yzer. Sampling locations S1, S2, S4, S5 are located on the Yzer river while sampling location S3 is located on a tributary. Open circles: Observed data, Dots: Augmented data

sponse $\boldsymbol{y} = (y_1, \ldots, y_n)$ consists of n observations measured at time $t = 1, \ldots, n$ then the STL procedure decomposes $\boldsymbol{y}$ into

$$\boldsymbol{y} = \boldsymbol{S} + \boldsymbol{T} + \boldsymbol{R}.$$

This method is implemented in the STL-routine of the tseries package for R (Trapletti, 2004). The STL-method is applied to the data of sampling location S1. As many standard techniques for time series analysis, the method however cannot handle missing data. Hence the augmented dataset is used. An STL plot is shown in Figure 1.5. The seasonal pattern is very obvious, it has an amplitude of 13.7 mg N/l. The contribution of the seasonal effect is low in summer and high in winter. This could be expected, during summer the fertilised nitrogen is still in the form of insoluble ammonium which is converted in the soil by micro-organisms to soluble

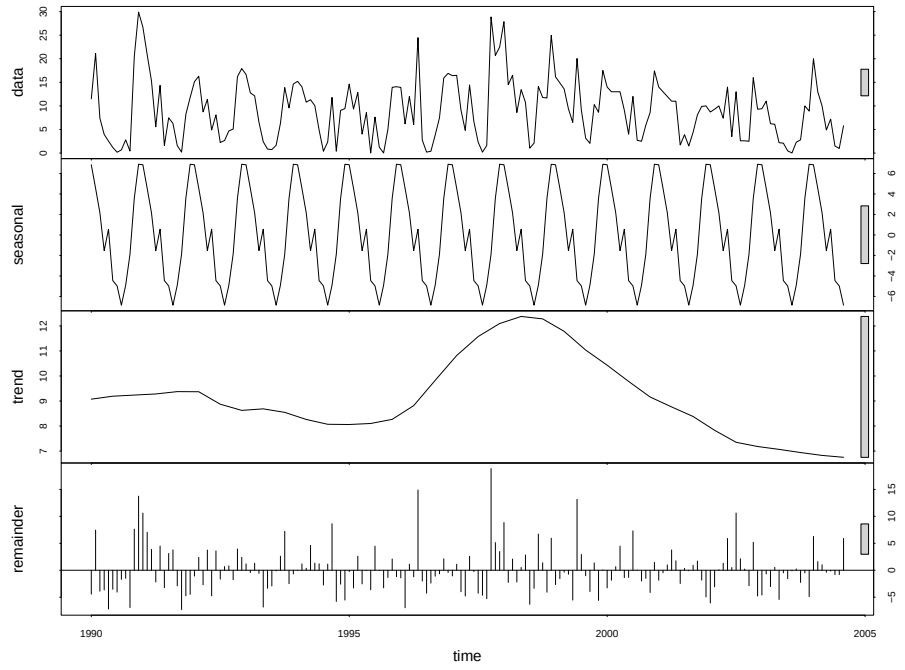

Figure 1.5: STL analysis of the augmented data at sampling location S1

nitrate. The accumulated nitrate is then washed out in the colder and wet winter period. The trend component indicates an increasing trend in the beginning of the series, the trend reaches its maximum in 1998 and from there on a decreasing trend is established that seems to level off at the end of the series. Note that this method is only explorative. It does not provide formal tests and/or variance estimates on the estimated seasonal effect and the trend which are needed for inference purposes.

The seasonal effect is also obvious when data from the raw nitrate series are plotted against the day of year (d, which has support $[1, 365]$). A common approach to model the variation is to include sinusoidal functions of fixed periods to describe the seasonal cycle within a year (e.g. Hirst, 1998; Cai and Tiwari, 2000; McMullan et al., 2003; McMullan, 2004). Another possibility is to use the loess smoother to let the data drive the functional relationship between the nitrate measurements and the day. A plot of nitrate in function of d is shown in Figure 1.6. From the raw data the seasonal effect is estimated with a smoother and with the Fourier basis

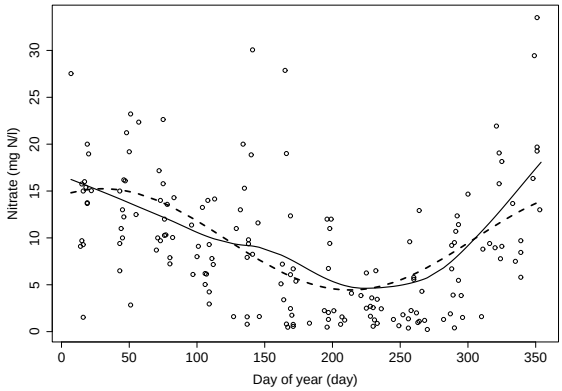

Figure 1.6: Plot of the nitrate data collected at sampling location S1 against the day of the year (all years are included in the plot). The dashed line represents the fit with the Fourier basis and the full line represents the fit with a local linear regression smoother

$(\gamma_1 \sin(2\pi\ d/365) + \gamma_2 \cos(2\pi\ d/365))$, where d is the day of year. The resulting estimated functions are added to the plot. Both fits clearly indicate the presence of the seasonal pattern. Again the contribution is high in winter and low in summer. The smoother and the Fourier basis are the two methods that will be used in this dissertation to model the seasonal effect.

Now that the reader is familiar with the data, we will give the outline of this work.

1.3 Objectives and outline

In this dissertation we have two major objectives. On the one hand, we aim to develop of a semi-automatic data validation procedure for water quality data. On the other hand, we want to develop spatio-temporal models to assess the observations at the sampling locations of a river network.

A validation tool could be of great help for experts in environmental agencies since it would enable them to focus on potential suspicious observations instead of having to validate all the data. An important feature of a validation tool that is used on

a day-to-day basis is that it requires a minimum amount of user interaction. The method should also be able to handle data acquired at irregular time intervals, and it has to deal with nonlinear relationships and trends present in water quality data. Preferable, the method should also be able to detect observations when their value is not in agreement with the measured values of the other water quality variables.

Spatio-temporal models for river networks should enable a valid statistical assessment of the water quality. The method should enable statistical inference at the level of individual sampling locations and on a more regional scale. We will first develop a fully parametric model. Then we will relax the assumptions to allow the estimation of nonlinear trends. Finally, we aim to generalise the spatio-temporal model in order to handle non-normal data that is distributed according to another member of the exponential family. Remark that the methods in this dissertation are not designed to perform interpolation at intermediate locations. Thus in this work, an analysis on a 'regional' scale should be interpreted as a simultaneous analysis at a finite number of sampling locations that are located within the study region.

An schematic overview of our objectives is given below,

1. The development of a statistical data validation procedure for water quality data that
   (a) can handle the nonlinear relationships and trends in water quality data,
   (b) can handle data acquired at irregular time intervals,
   (c) restricts the amount of user interaction,
   (d) enables experts in environmental agencies to focus mainly on potential suspicious observations,
   (e) and detects suspicious observations when their relationship with other water quality variables is unusual.
2. The development of spatio-temporal models for river monitoring networks that
   (a) enable valid statistical inference at individual sampling locations as well as at a larger spatial scale,
   (b) can be used for the estimation of nonlinear trends,
   (c) and can deal with non-Gaussian observations.

The first issue that will be addressed in this dissertation is the complex nature of the water quality data that makes the assumption of linearity often too rigid. The assumption of full parametrical models is easily relaxed by using nonparametric smoothing techniques. These techniques are much more flexible and can capture local trends and complex relationships between environmental variables. When multiple predictor variables are available, these techniques can be easily extended to surface smoothing (e.g. Cleveland and Devlin, 1988). Buja et al. (1989), however, showed that there were a number of disadvantages related to multivariate smoothers. Therefore, they introduced additive models as a nonparametric tool to model a multivariate regression surface. Instead of combining all predictors in one multivariate smoother, they have proposed an additive model structure where each component is a one-dimensional smoother which models the contribution of a particular predictor. In time-series studies of air pollution and mortality, the use of nonparametric models is widespread, since they allow for adjustments for nonlinear confounding effects of seasonality, trends, and other environmental conditions such as meteorological conditions (e.g. Dominici et al., 2002 and Giannitrapani et al., 2005). Recently, nonparametric modelling has also been considered for modelling water quality (e.g. Qian et al., 2000, Cai and Tiwari, 2000, Stålnacke et al., 1999, McMullan et al., 2003, McMullan, 2004). Within this context, we will explore additive models in **Chapter 2** and **Chapter 3**. In both chapters inference is performed at the level of the individual sampling locations and missing data was present in all examples. **Chapter 2** gives a general introduction to additive models and in **Chapter 3** we make use of additive models to design a semi-automatic data validation procedure. In our approach, additive models are used to extract information from the historical data, and the bootstrap is used to incorporate the sampling variability.

The second part of this dissertation deals with the development of spatio-temporal models that enable the analysis of water quality data at the level of individual sampling locations as well as on a more regional scale. To our knowledge only a few references are available on spatio-temporal models for river networks (Gardner et al., 2003; Monestiez et al., 2005; Cressie et al., 2006; Ver Hoef et al., 2006) and they are all related to spatial prediction. The focus in this dissertation, however, is on the assessment of the data that is observed at the sampling locations themselves. Therefore the observations of the monitoring network at a certain time instant can be considered as the realisation of a finite-dimensional multivariate random variable with each dimension corresponding to each of the p sampling locations. This enables us to write the model as a p-dimensional state-space model. After we have given a general introduction to state-space models in **Chapter 4**, we will construct

a model for the spatio-temporal correlation structure of river monitoring networks in **Chapter 5**. The river topology is used to derive the spatial dependence structure and an autoregressive process is proposed for the temporal dependence. To assess the evolution in water quality, a parametric mean model is used. In **Chapter 5** we will also provide an efficient algorithm to estimate the model parameters. However, many environmental processes are characterised by a nonlinear trend. In **Chapter 6** we therefore combine the spatio-temporal correlation structure developed in **Chapter 5** with a semi-parametric mean model. The evaluation of the local trend can be done by testing whether the first derivative of the nonlinear trend is different form zero. Because these tests have to be performed at each time instant, multiplicity is another problem which has to be addressed in this chapter. In **Chapters 5** and **6** the observations are assumed to be Gaussian. To deal with non-Gaussian observations a generalisation of our spatio-temporal model is presented in **Chapter 7**. In particular a Bernoulli response is considered. Environmental compliance is often based on threshold levels, providing a binary response to the decision maker. We will make use of generalised linear mixed models so that binary responses can be used to assess trends in the violation frequency of water quality standards.

With **Chapter 8** we conclude this dissertation by a discussion of the presented methodologies, we bring the main conclusion and provide an outlook to future research perspectives.

Part I

Statistical data validation by the use of additive models

A selection of the presented work will appear in

Clement, L., Thas, O., Ottoy, J.P. and Vanrolleghem, P.A. (2007). Data management of river water quality data - a semi-automatic procedure for data validation. *Water Resources Research*, accepted.

Chapter 2

Review of additive models (AM's)

2.1 Introduction

In Europe, the design of water quality monitoring networks is one of the key actions of the Water Framework Directive (WFD)(EC, 2000). This results in a high amount of water data, that is collected, stored and processed in Europe. Like other environmental data, water quality data have a complex nature. They contain a considerable amount of noise, due to their natural variability and the measurement error. They may contain missing values, may be censored due to the detection limits of the measuring methods and are commonly gathered on irregular time instants. They also may be mutually dependent, non-normally distributed, possess cyclic variations and contain nonlinear trends (e.g. Hirsch et al., 1982; McMullan, 2004).

The large amount of the data together with its complex nature have triggered modelling as an additional tool to extract information for those observations. Within the research community, monitoring and modelling have become generally accepted to be related activities (e.g. Parr et al., 2003; Højberg et al., 2007). In this dissertation, we will focus on modelling to validate new observations, to identify trends and to evaluate the impact of actions which were taken to improve the water quality. With respect to these aims, we will use models from three different perspectives. Firstly, the models are used to describe the dependence of a water quality variable of interest, the response (or dependent) variable (Y), on several predictor (or independent) variables ($X_1, \ldots, X_q$). Possible predictors are for instance a trend term, a seasonal component, a temperature effect and other water quality variables which are measured simultaneously. This use typically involves estimation of parameters and/or regression functions. Secondly, the relative contribution of each of the predictors in explaining Y can be studied and this gives us insight in the evolution/trend of the response and its relationships with other water quality variables. A third purpose is prediction, where we want to predict the mean response given a certain set of values $X_1, \ldots, X_q$.

We now introduce the modelling framework which we will use for these purposes. Suppose we have n observations of the response Y sampled at different times $t = 1, \ldots, n$. They are denoted by an $n \times 1$ vector $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ and they are measured simultaneously with the q predictor vectors $\boldsymbol{x}_j = (x_{j1}, \ldots, x_{jn})^T$, $j = 1, \ldots, q$. Then a typical water quality dataset $\boldsymbol{D}$ is represented by an $n \times (q+1)$ matrix $\boldsymbol{D} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_q, \boldsymbol{y})$. A general framework to model the relationships between the mean of Y and its predictors X can be written in the following form,

$$Y = m(X_1, \ldots, X_q) + \epsilon, \tag{2.1}$$

where m is the unknown regression function and ϵ is a zero mean random term. The data-analyst now has to choose a certain structural form to model the conditional mean $m(X_1, \ldots, X_q)$. This can be done in a parametric, nonparametric or semi-parametric way. When a parametric model is used, it is assumed that the functional form is known and can be completely parameterised. In a nonparametric regression analysis, however, no functional form is assumed and the regression is completely data-driven. In a semiparameteric approach, the functional form is not fully specified and some components are modelled parametrically while others are modelled in a nonparametric way. A well known example of a fully parametric model is the standard multiple linear regression model. Because the relationships between the response and the predictors are assumed to be linear, Equation (2.1)

can be written as

$$Y = \alpha + \beta_1 X_1 + \ldots + \beta_q X_q + \epsilon = \alpha + \sum_{j=1}^{q} \beta_j X_j + \epsilon, \tag{2.2}$$

with the parameter α and a $q \times 1$ parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^T$. To fit the model to the data, the parameters have to be tuned so that the fitted values $\hat{\boldsymbol{y}} = \hat{\boldsymbol{\alpha}} + \sum_{j=1}^{q} \hat{\beta}_j \boldsymbol{x}_j$ are in some sense as close as possible to the observed values $\boldsymbol{y}$ (e.g. by using least squares). The popularity of linear models is largely due to their simplicity and easy interpretation. However, the model depends on a strong assumption of linearity between the predictors and the response. Unfortunately, like other environmental data, trends and relations between water quality data typically are nonlinear (e.g. Cai and Tiwari, 2000; Dominici et al., 2002; Wood and Augustin, 2002; McMullan et al., 2003; McMullan, 2004). Therefore it would be better to let the data drive the specification of the functional relation between the predictor variables and the response. This is exactly what scatterplot smoothers do for the two-dimensional case (Y, X_1). They model Y as $Y = f_1(X_1) + \epsilon$, where $f_1(X_1)$ is a smooth function used to approximate the underlying function $m(X_1)$ without imposing a rigid parametric relationship such as linearity. A principle used by many smoothers is to estimate the regression surface locally instead of globally. The fit at a certain predictor value x_i is only based on the data that lays in a certain neighbourhood of x_i. This adds much more flexibility to the estimation of the underlying function. An example is the loess smoother (Cleveland and Devlin, 1988), which is illustrated in Figure 2.1. The resulting smooth indicates an increase in the nitrate level between January 1990 and December 1997, and from there on a steady decrease in the average nitrate concentration is established. The linear regression fit remains more or less constant over the entire temporal domain because it cannot handle slope changes. The idea of scatterplot smoothing can be easily extended to the q-dimensional case (e.g. Cleveland and Devlin, 1988; Cleveland and Grosse, 1991; Loader, 1999b) where $m(X_1, \ldots, X_q)$ is approximated by a q-dimensional smoother $f_{1\ldots q}(X_1, \ldots, X_q)$. Note that the number of dimensions equals the number of regressors. There are unfortunately some problems related to multidimensional smoothers. In particular,

1. Buja et al. (1989) showed that most multidimensional extensions of univariate smoothers are not attractive from a computational point of view.

2. Due to their multivariate nature, they also suffer from "the curse of dimensionality". These problems are mainly triggered by the multidimensional neighbourhoods which have to be defined. Hastie et al. (2001) illustrated that

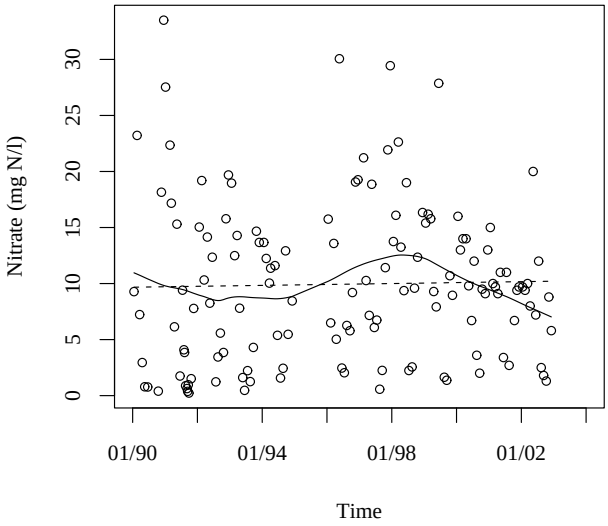

Figure 2.1: Scatterplot of nitrate concentration in function of time, a least squares regression line (dashed line) and a loess smoother (solid line)

the neighbourhoods are less local when the number of predictors increases. Another issue is also related to the data sparseness in a high dimensional setting. Therefore, more data ends up in the boundary region. Since smoother estimates are known to be more biased in the boundary regions, the boundary problem is more dominant in a multidimensional setting.

3. It is difficult to define a sensible metric for the multidimensional neighbourhoods, because the predictors are often measured in different units.

4. The visualisation of multivariate smoothers is less obvious. Especially when the number of predictors gets beyond two. In order to study the effects of the individual predictors, projections from the hyper-surface can be made on a lower dimensional space, but this projection depends on the fixed values of the remaining predictors and thus they are rather noisy.

To overcome the above mentioned problems, Buja et al. (1989) came up with an alternative approach. They suggested to use one-dimensional smoothers as additive

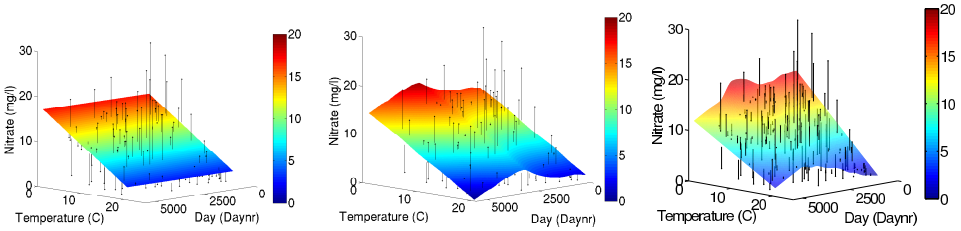

Figure 2.2: Nitrate concentration in function of the time (day number) and temperature (°C). Left panel: linear model, Middle panel: additive model using two univariate local linear regression smoothers and Right panel: a two-dimensional local linear regression smoother

building blocks of the model. This results in a more restricted class of nonparametric regression models, also referred to as additive models. Additive models extend standard linear models and model this response variable as

$$Y = \alpha + \sum_{j=1}^{q} f_j(X_j) + \epsilon, \tag{2.3}$$

where f_j can be any function, however in most cases smoothers are used. Similar to linear models, additive models are additive in the covariates but not necessarily in a linear way. Due to this additivity, the effect of a predictor on the fitted response surface does not depend on the values of the other predictors. Thus, the contribution of each predictor can still be studied individually. This enables the user to decompose the model in each of its smooth functions, which can be graphically depicted. Figure 2.2 shows the differences between a linear model $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, an additive model $Y = \alpha + f_1(X_1) + f_2(X_2) + \epsilon$ and a multivariate regression smoother $Y = f_{12}(X_1, X_2) + \epsilon$, where Y is the nitrate concentration, X_1 represents time and X_2 temperature. Due to the additivity assumption, the bump at low temperatures and at intermediate dates is less high for the additive model than in the multivariate smoother model . However the bump is situated in a data sparse region and might be a boundary effect from the multivariate smoother. Apart from this feature, the fits by the additive model and the multivariate smoother look similar. This can also be seen in Figure 2.3 where the fitted models are plotted as a function of each predictor separately. Similar to what was observed in Figure 2.1, both fitted models show higher fits around the end of 1997 and the beginning of 1998 and seem to decline afterwards. The trend is obscured by the large oscillations on a smaller time scale. They originate from the temperature effect which is modelled simultaneously. When the modelled surface

is projected onto the nitrate-temperature plane, the overall trend seems to indicate an inverse relationship between nitrate levels and temperature. This can be easily explained. During summer the fertilised nitrogen is still in the form of insoluble ammonium which is converted in the soil by micro-organisms to soluble nitrate. The accumulated nitrate is then washed out in the colder and wet winter period. The oscillations observed in the temperature effect are due to similar temperatures which are measured on different dates.

The additive model, however, enables the analyst to look at the contribution of each of the predictors separately. This is illustrated in Figure 2.4. At the end of 1992 the contribution of the long term trend shows a steep incline and reaches a maximum at the beginning of 1998. From this point on, it seems to decline up to the present. This decline is possibly due to the introduction of two manure action plans (MAP's) introduced in 1996 and 2000 (Vlaams Parlement, 1995, 1999), respectively. The aim of these MAP's was the reduction of the nutrient pollution originating from agriculture. In Figure 2.4 the interpretation of the contributions of each of the predictors is much easier. The inverse relation between temperature and nitrate is also more obvious.

Because smoothers are used as the basic building blocks of the additive models, a brief review on smoothing is needed before we can move on to model fitting and selection.

2.2 Smoothing

Hart (1997) mentioned that the aim of smoothing is to remove data variability that has no assignable cause and to make systematic features of the data more apparent. Smoothing however has become synonymous with a variety of nonparametric methods used in the estimation of functions. In this dissertation, the term smoothing is used in the latter sense. Smoothing resorts under the class of nonparametric tools for regression analysis since they generally do not assume a rigid form for the dependence between the mean response and the predictor variables. Hastie and Tibshirani (1990) defined a smoother as a tool for summarising the trend of a response Y as a function of one or more predictors $X_1, \ldots, X_q$, and it produces an estimator which is less variable than Y itself. According to Cleveland and Devlin (1988) and Hastie and Tibshirani (1990), they have three major uses. The first is to provide an exploratory graphical tool which gives the user insight into the be-

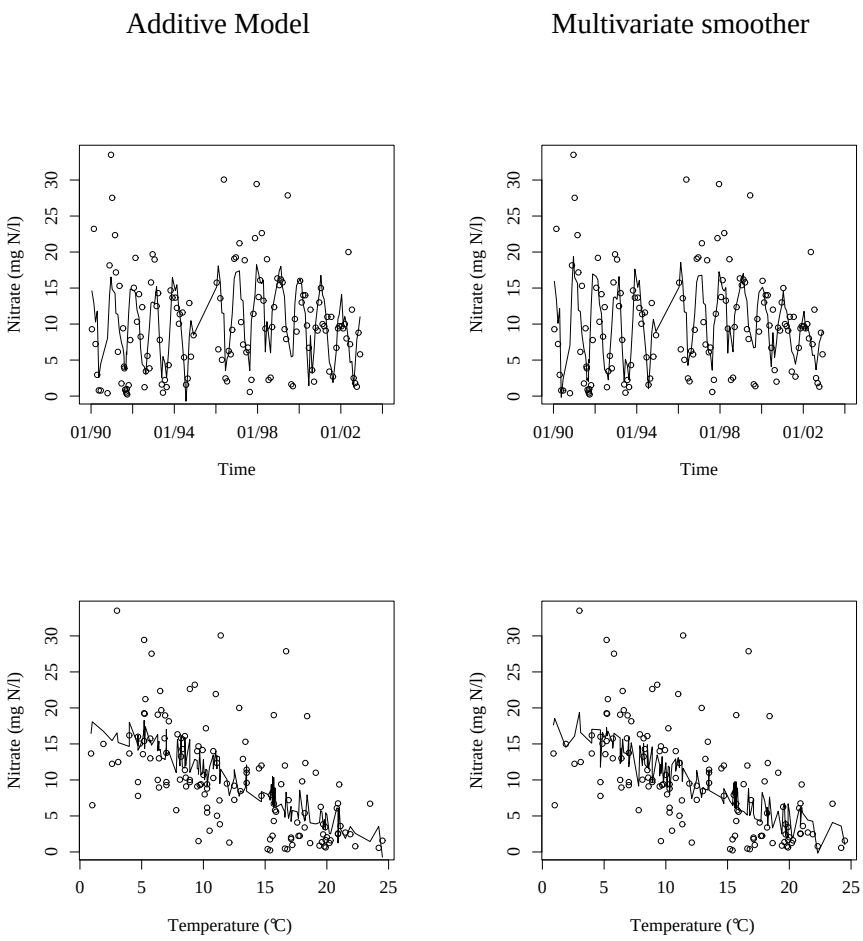

Figure 2.3: Nitrate concentration in function of the date (top panels) and temperature (bottom panels). In the left panels nitrate data are represented along with the fit of the additive model, in the right panels the nitrate data was modelled using a two-dimensional local linear regression smoother

haviour of the data and which helps him/her to choose an appropriate parametric model. Secondly, they are used as a regression diagnostic to check the adequacy of the fitted parametric models. Their third use is to estimate a regression surface, without resorting to a parametric class of functions. This can be done directly, using a multivariate smoother or by the use of additive models, which use univariate

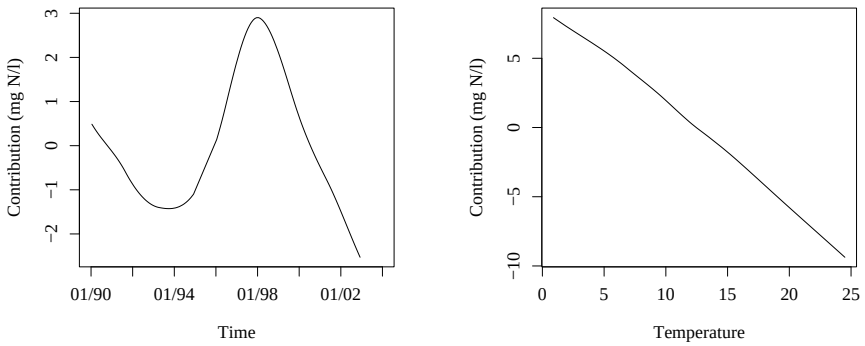

Figure 2.4: Contribution of long term trend (left) and temperature (right) to the nitrate concentration predicted by an additive model

smoothers as basic building blocks. We will mainly focus on smoothers from that third point of view. In this section, for notational comfort, only one predictor is taken into account. Hence, the model in Equation (2.1) reduces to

$$Y = m(X) + \epsilon. \tag{2.4}$$

When a smoother is used to estimate the function m, it is basically an approximation of the true regression function, and it generally contains a certain amount of bias. To stress that the smoother is only an approximation of the true function m, we will use the notation f to represent the smooth function. Our brief overview of smoothing is restricted to linear smoothers since their statistical properties are well studied in literature. A linear smoother f can be estimated as a linear combinations of the response. Linear smoothers can thus always be written as

$$\hat{\boldsymbol{f}} = \boldsymbol{S}\boldsymbol{y}, \tag{2.5}$$

where $\hat{\boldsymbol{f}}$ is the $n \times 1$ vector of the estimations of f at each of the n observations $\hat{\boldsymbol{f}} = (\hat{f}_1, \ldots, \hat{f}_n)^T$ and $\boldsymbol{S}$ is the $n \times n$ smoother matrix which consists of a set of unique weights S_{ij} for each x_i. The specific value of the weights depends on the type of smoother that is used. For local polynomial smoothing they are defined in Section 2.2.3. When similar assumptions are made as in the parametric regression framework, linear smoothers can inherit a whole set of inference procedures known from the classical parametric regression context, e.g. the construction

of confidence intervals (Cleveland and Devlin, 1988). Two important examples of such assumptions are Gaussian residuals and an unbiased estimation of m by the function $\hat{f}(x)$. In this section, splines, kernel smoothing and local polynomial regression smoothers are covered.

2.2.1 Splines

We start with a brief introduction to univariate splines, where the mean response is modelled as a function of the one-dimensional predictor variable X. To provide a flexible tool for approximating the underlying process of Y, piecewise polynomials can be used to represent the function $f(X)$. This transforms the global nature of polynomial regression into a fitting procedure which has a more local nature (Hastie and Tibshirani, 1990). The piecewise polynomials are obtained by dividing the domain on X into intervals using a number of breakpoints, also known as knots. In each interval, f is then locally represented by a separate polynomial (Hastie et al., 2001). In most applications, one typically wants the function to join smoothly at these knots. By allowing more knots, the function becomes more flexible. Our eye is apparently skilled to pick up second order and lower order discontinuities, but not the higher order discontinuities. Therefore the polynomials are generally forced to be continuous up to the second order derivatives at the knots. Cubic splines are the lowest order splines that fulfil these conditions. Unless one is interested in smooth derivatives, there is generally no real reason to use splines of higher orders (Hastie et al., 2001).

When splines are used for prediction, they are known to suffer from large extrapolation errors. Polynomial fits are known to be erratic at the boundaries, and thus extrapolation can be dangerous. This behaviour is even more explicit when using splines, because the fit at the endpoints is based on less data. To reduce these errors, natural cubic splines can be used. These add an additional constraint at the endpoints of the regression, the second (f'') and third derivative (f'') of f are equal to zero, $f''' = f'' = 0$. In this way they impose the spline to behave in a linear way beyond the boundary knots and thus stabilise the variance of the spline near the endpoints (Hastie and Tibshirani, 1990).

Computationally, cubic splines can be calculated by using a set of basis functions, say $\boldsymbol{s}_1(\boldsymbol{x}), \ldots, \boldsymbol{s}_m(\boldsymbol{x})$. The regression function is then estimated by simply regressing the response $\boldsymbol{y}$ against the $\boldsymbol{s}_j(\boldsymbol{x})$'s. Basis functions which are commonly used are for instance the truncated power basis or the numerically superior B-spline

basis (Hastie et al., 2001; Hastie and Tibshirani, 1990; Hart, 1997; Eilers and Marx, 1996). Splines are computationally attractive when the number and the location of the knots are known, because this reduces the fitting procedure to a linear regression problem. The difficulty however in fitting splines is choosing the number and the location of the knots. When a small number of knots is used, the spline often shows some undesired non-local behaviour. However, using more knots is often limited by the available number of degrees of freedom. Controlling the desired amount of smoothing by restricting the number of knots is not straightforward. For smoothing purposes, it is easier to use a third type of splines, smoothing splines. In contrast to the previous splines, they originate from the following optimisation problem in which

$$\sum_{t=1}^{n}(y_t - f(x_t))^2 + \gamma \int_a^b f''(x)^2 dx \tag{2.6}$$

is to be minimised. Here γ is a fixed constant, and $a \leq x_1 \leq \ldots \leq x_n \leq b$. The first term is the sum of squared errors, which measures the closeness of the fitted regression function to the data, while the second term penalises curvature in the function. Remarkably, it can be shown that Equation (2.6) has a unique minimiser which is a natural cubic spline with knots at the each x_t, $t = 1, \ldots, n$. At first sight, the family seems overparameterised: fitting such a spline requires n parameters. However, the coefficients are constrained as well due to the penalisation term. This brings down the effective dimension drastically (Hastie and Tibshirani, 1990). The parameter γ controls the amount of smoothness. Large values of γ produce smoother curves, while smaller values produce more wiggly curves.

We now continue with a totally different concept of smoothing, which is kernel smoothing.

2.2.2 Kernel smoothing

Local averaging is a very intuitive and appealing method for the approximation of a regression function m. It is easy to understand that points close to x contain more information on $m(x)$ than points which are more remote from x. The principle can be improved by computing a locally weighted average. This is exactly what is done by kernel smoothers. Let K_h be a real-valued weight function which depends on the bandwidth h. The function K_h is assumed to be a symmetric probability density function and is also referred to as the *kernel function*. With this notation,

the Nadaraya-Watson kernel estimator is given by

$$\hat{f}_h(x) = \frac{\sum_{t=1}^{n} K_h(x_t - x) y_t}{\sum_{t=1}^{n} K_h(x_t - x)}, \tag{2.7}$$

which was independently derived by Nadaraya (1964) and Watson (1964).

The Gasser-Müller estimator is another common kernel estimator,

$$\hat{f}_h(x) = \sum_{t=1}^{n} y_t \int_{s_{t-1}}^{s_t} K_h(u - x) du, \tag{2.8}$$

with $s_t = (x_t + x_{t+1})/2$, $x_0 = -\infty$ and $x_{n+1} = +\infty$ (Gasser and Müller, 1979). Both kernel estimators are zero order approximations of the regression function m. Fan (1992) and Fan and Gijbels (1996) have shown that the Nadaraya-Watson kernel gives a more biased estimator in an interior point than the Gasser-Müller kernel, but the latter corrects the bias at the expense of the variance.

Both methods are also seriously biased at the boundaries. Although boundary correction methods are possible, they are complicated and not as effective and intuitive as the automatic boundary correction of the local polynomial regression smoother (Fan, 1992; Fan and Gijbels, 1996; Hasti and Loader, 1993), which we introduce in the next section.

2.2.3 Local polynomial regression

The idea of local polynomial regression can easily be motived by approximating the regression function m in a neighbourhood of x_0 by a Taylor expansion,

$$m(x) \approx m(x_0) + \sum_{k=1}^{p} \frac{m^{(k)}(x_0)}{k!} (x - x_0)^k, \tag{2.9}$$

where $m^{(k)}(x_0) = \frac{\partial^k m}{\partial x^k}|_{x_0}$. Local weighted least squares can be used to fit this polynomial by minimising

$$\sum_{t=1}^{n} [y_t - \sum_{k=0}^{p} \boldsymbol{\beta}_0 (x_t - x_0)^k]^2 K\left(\frac{x_t - x_0}{h}\right). \tag{2.10}$$

where $K(.)$ is a kernel function which will be introduced later on and h is the bandwidth which defines the size of the neighbourhood $(x_0 - h, x_0 + h)$. The kernel function assigns weights to each observation.

The solution to this local weighted least squares problem is

$$\hat{\boldsymbol{\beta}}_0 = (\boldsymbol{x}_c^T \boldsymbol{W}_0 \boldsymbol{x}_c)^{-1} \boldsymbol{x}_c^T \boldsymbol{W}_0 \boldsymbol{y}, \tag{2.11}$$

where $\mathbf{x}_c$ is an $n \times (p+1)$ matrix $\mathbf{x}_c = (\mathbf{1}, \mathbf{x}_{vc}, \ldots, \mathbf{x}_{vc}^p)$, $\mathbf{1} = (1, \ldots, 1)^T$ is an $n \times 1$ vector of ones, $\mathbf{x}_{vc} = (x_1 - x_0, \ldots, x_n - x_0)^T$ is an $n \times 1$ vector and $\mathbf{W}_0$ is an $n \times n$ diagonal matrix build up by the kernel weights (Fan and Gijbels, 1996). The response y_0 corresponding to x_0, is then estimated by

$$\hat{y}_0 = \begin{bmatrix} 1 & 0 & \ldots & 0 \end{bmatrix} \hat{\boldsymbol{\beta}}_0 = \begin{bmatrix} 1 & 0 & \ldots & 0 \end{bmatrix} (\mathbf{x}_c^T \boldsymbol{W}_0 \boldsymbol{x}_c)^{-1} \mathbf{x}_c^T \boldsymbol{W}_0 \boldsymbol{y} = \mathbf{S}_0 \boldsymbol{y}, \tag{2.12}$$

where the centered vector of x_0 is $\begin{bmatrix} 1 & (x_0 - x_0) & \ldots & (x_0 - x_0)^p \end{bmatrix} = \begin{bmatrix} 1 & 0 & \ldots & 0 \end{bmatrix}$. Hence the fit of local polynomial smoothers is a linear combination of the responses. If this procedure is performed for all n observations (x_t, y_t), $t = 1, \ldots, n$, the fit $\hat{\mathbf{y}}$ can be written as

$$\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}, \tag{2.13}$$

where $\mathbf{S}$ is an $n \times n$ matrix and is also referred to as the smoother matrix. Since this predictor is of the same form as Equation (2.5), this estimator is also a linear smoother.

Several important choices have to be made before local polynomial regression can be used. The size of the bandwidth has to be selected, but descriptions of practical procedures for bandwidth selection are kept for the next section. The degree of the polynomial has to be set. Because the bias is mainly controlled by the bandwidth, the choice of the degree of the local polynomial is less important. However, for a fixed bandwidth, increasing the degree reduces the bias, but this is at the expense of an increasing variance of the fit and of a higher computational cost. A very important issue was pointed out by Fan (1992, 1993). He showed that the variability remains unchanged by going from a local constant to a local linear fit. He also showed that the local constant fit suffers from low asymptotic efficiency as compared to the local linear fit. Fan and Gijbels (1996) generalised these results and proved that the variability does not increase by going from an even order polynomial fit to an odd order polynomial fit. The extra parameter can however reduce the bias significantly. They also argued that even order fits suffer from serious boundary effects, in contrast to odd order fits which have nice adaptive boundary properties. From this point of view Fan and Gijbels (1996) recommended to use the

lowest possible odd order for the polynomial fit. Hence, throughout this dissertation the degree is set to 1, unless derivatives have to be estimated. Another question which has to be addressed is the choice of the kernel function K. The choice of the kernel is not that important from a practical point of view. However, Fan and Gijbels (1996) have shown that the Epanechnikov kernel, $K(u) = 3/4(1 - u^2)$ for $-1 < u < 1$ and zero for u outside that range, is asymptotically optimal for the interior of the domain. When the Epanechnikov kernel is to be used in Equation (2.10), u has to be replaced by $u = (x_i - x_0)/h$. This kernel is used in the remainder of this dissertation.

There is a vast amount of literature on the attractiveness and advantages of local linear regression smoothers (e.g. Cleveland, 1979; Cleveland and Devlin, 1988; Fan, 1992, 1993; Hasti and Loader, 1993; Fan and Gijbels, 1996; Loader, 1999b). Fan (1992) showed that the local linear regression smoother is the best among linear smoothers. Fan and Gijbels (1996) studied the linear minimax risk of polynomial regression in order to compare with other linear smoothers. For an appropriate choice of the bandwidth and the kernel, estimating the mean $m(x_0)$ by local linear regression and the first derivative $m^{(1)}(x_0)$ by a local polynomial of second order is efficient both in the interior of the design and at the boundaries. Fan (1992), Fan and Gijbels (1996), and Hasti and Loader (1993) also showed that local polynomial regression adjusts automatically for bias at the boundaries and is design adaptive in the sense that they also adjust for bias in regions where the predictors are nonuniform. As another advantage, the weighted least squares approach also enables straightforward generalisations of classical statistical inference procedures (Cleveland and Devlin, 1988; Fan and Gijbels, 1996; Loader, 1999b). Finally, Fan and Gijbels (1996) and Loader (1999b) also reviewed some fast computing algorithms for local polynomial regression, which enable them to compete with other numerical smoothing techniques from a computational point of view. Their computational aspects, simplicity and attractive properties are a strong plea in favour of the use of local polynomial smoothing. Therefore the local linear smoother is used as the basic smoothing procedure throughout this dissertation.

2.2.4 Tuning the smoothing parameters of local polynomial regression

The bandwidth and the kernel function K control the size of the local neighbourhood. Therefore the choice of the bandwidth in local polynomial regression is a crucial one. When taking the bandwidth close to zero, the data are interpolated, leading to an overparameterised model. A bandwidth taken arbitrarily large, re-

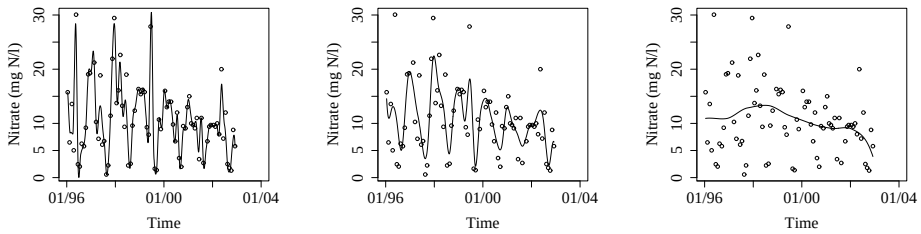

Figure 2.5: Fit of the nitrate data with local linear regression with bandwidth equal to 2 months (left), 4 months (middle), 2 years (right panel)

sults in a polynomial of degree p which is fitted globally. Hence, the bandwidth is a key element in controlling the complexity of the smoother. The smaller the bandwidth, the more degrees of freedom that can be used for controlling the bias. But this reduction in bias does not come for free. Smaller bandwidths will also lead to an increase of the variance associated with the estimators. Too small bandwidths typically result in more wiggly curves and this can conceal the main features which are present in the data. This problem is addressed in literature as undersmoothing. Too large bandwidths on the other hand tend to oversmooth the data and this can lead to the introduction of a substantial bias. This is illustrated in Figure 2.5 where nitrate data are modelled using a local linear smoother and 3 different bandwidths. When a bandwidth of two months is taken, the obtained curve is too wiggly and highlights features which may be inherent to the sampling variability. A bandwidth of 4 months still highlights the cyclic pattern in the nitrate concentration but produces a smoother fit. A large bandwidth is sensitive to oversmoothing, leading to an estimate which can miss certain features of the curve. A bandwidth of 2 years, for example, looses the ability to pick up the cyclic behaviour of the nitrate concentration. The size of the bandwidth can be chosen to be constant over the domain of X, or can be variable. For variable bandwidths a further distinction can be made between local variable bandwidths, $h(x_0)$, varying with the location x_0, and global variable bandwidths, $h(x_i)$, changing with the observations x_i. An example of a variable bandwidth with a very simple nature is the nearest neighbour bandwidth. The selector requires that a fixed fraction of the data is included in the neighbourhood. This fraction is referred to as the span s. It adapts automatically the amount of smoothing to the local situation, using small bandwidths in a dense design region, and large bandwidths in sparse regions (Altman, 1992; Fan and Gijbels, 1996; Loader, 1999b). This method thus prevents that the regression

in sparse data regions is based on only a limited number of points. For the trade off between the amount of bias and the associated variance of the estimator, a criterion is needed which takes both terms into account.

Two criteria are commonly used for this purpose: the Mean Squared Error (MSE) and the Mean Integrated Squared Error (MISE). The MSE is defined as

$$MSE(x_0) = \mathrm{E}\left((\hat{f}_h(x_0) - m(x_0))^2\right) = \mathrm{var}\left(\hat{f}_h(x_0)\right) + \left[\mathrm{E}\left(\hat{f}_h(x_0) - m(x_0)\right)\right]^2. \tag{2.14}$$

It is the sum of the variance and the squared bias of the estimator. Minimising this criterion will give the theoretical optimal local bandwidth. However, this choice depends on the true underlying function m, which is unknown. An optimal bandwidth could be defined as h which minimises $MSE(x_0)$. An asymptotical approximation of this optimal bandwidth is given by Fan and Gijbels (1996),

$$h_{opt}(x_0) = C_{v,p}(K)\left[\frac{\sigma^2(x_0)}{\{m^{(p+1)}(x_0)\}^2 P(x_0)}\right]^{1/(2p+3)} n^{-1/(2+3p)}, \tag{2.15}$$

where v is the order of the derivative of m of interest (it is zero when we are interested in the mean function and 1 if the prime interest is the first derivative), p is the order of the polynomial, which is usually equal to $v+1$, $P(x_0)$ is the design density function evaluated in x_0 and $C_{v,p}(K)$ is a constant depending on the kernel (e. g. for the Epanechnikov kernel it is 1.719 when $v = 0$ and $p = 1$ and 2.275 when $v = 1$ and $p = 2$). The mathematical definition of $C_{v,p}(K)$ and values for other kernels and/or other values of v and p are reported in Fan and Gijbels (1996).

The second criterion is the MISE. It is defined as a weighted integration of the MSE over the domain of x, Θ_x,

$$\int_{\Theta_x} MSE(x)w(x)dx, \tag{2.16}$$

where w is a non-negative weight function. Using the MISE leads to the specification of a global optimal bandwidth, which stays fixed over the entire domain of x. The following solution is provided by Fan and Gijbels (1996),

$$h_{opt} = C_{v,p}(K)\left[\frac{\int_{\Theta_x} \sigma^2(x)w(x)/P(x)dx}{\int_{\Theta_x} \{m^{(p+1)}(x)\}^2 w(x)dx}\right]^{1/(2p+3)} n^{-1/(2+3p)}. \tag{2.17}$$

These asymptotical results can not be used directly to find the optimal bandwidth, because they rely on some unknown quantities such as the design density $P(.)$ at

the design points x, the conditional variance $\sigma^2(.)$ and the conditional mean $m(.)$. Therefore alternative methods are needed to estimate the bandwidth in practice. Two main approaches are described in the literature: classical methods and plug-in methods.

Classical methods consist in the minimisation of certain criteria as the leave-one-out cross validation (CV), generalised cross validation (GCV) or the Akaike information criterion (AIC). The CV criterion can be written as

$$CV(h) = \frac{1}{n}\sum_{t=1}^{n}[y_t - \hat{f}_h^{-t}(x_t)]^2, \tag{2.18}$$

where $\hat{f}_h^{-t}(x_t)$ indicates the estimation at x_t obtained by fitting the smoother f to the reduced dataset which does not contain the data point at time t. For linear smoothers the CV criterion can be rewritten so that the explicit recalculation of the smoother is not needed.

$$CV(h) = \frac{1}{n}\sum_{t=1}^{n}\{\frac{y_t - \hat{f}_h(x_t)}{1 - S_{tt}(h)}\}^2, \tag{2.19}$$

where S_{tt} indicates the t^{th} diagonal element of the smoother matrix $\boldsymbol{S}$. The GCV criterion replaces the S_{ii} by their average tr$(\boldsymbol{S}/n)$. The GCV is thus defined as

$$GCV(h) = \frac{1}{n}\sum_{t=1}^{n}\{\frac{y_t - \hat{f}_h(x_t)}{1 - \text{tr}(S)/n}\}^2. \tag{2.20}$$

The use of the CV and GCV criteria can be justified because they are both consistent estimators for the MISE (Fan and Gijbels, 1996).

Another popular criterion is the Akaike information criterion (AIC),

$$AIC = -2l + 2df, \tag{2.21}$$

where l is the log-likelihood and df are the degrees of freedom used by the smoother, calculated as tr$(\boldsymbol{S})$. When the errors are Gaussian and when the variance has to be estimated, the first term reduces to $-2l = n\log(2\pi) + n\log(\sum_{t=1}^{n}(\hat{e}_t^2)/n) + n$. Note that in this case df has to be increased by 1 due to the estimation of the unknown variance. In particular, $df = \text{tr}(\boldsymbol{S}) + 1$.

The concept of plug-in bandwidth selection is based on the replacement of the unknown quantities needed in Equation (2.17) by their estimates. These estimates

are mainly based on local polynomial regression of higher order polynomials using pilot bandwidths. By doing so the problem of bandwidth selection is shifted to the selection of appropriate pilot bandwidths. A plug-in estimator for local least squares regression can be found in Ruppert et al. (1995). An often repeated criticism on the classical approach is that the resulting bandwidths are often too variable and frequently undersmooth (e.g. Loader, 1999a). When in a simulation study repeated samples are drawn from a model, cross validation can select bandwidths that are very different from sample to sample. However, Loader (1999a) argued that this can be expected when bandwidth selection is applied to problems with features which are difficult to detect since the selector has to decide which features in the dataset are real. He also showed that less variable bandwidth selectors display this difficulty in another way: by consistently oversmoothing. Therefore he claimed that the variability of the bandwidth estimates by classical methods is rather a symptom than a problem of the difficulty in estimating the bandwidth. Loader (1999a) also showed that the plug-in based estimates are asymptotically beaten by their pilot estimates and prone to oversmooth when they are presented to difficult smoothing problems.

Bandwidth selection in this section has to be seen in the framework of univariate smoothers used in additive models, where q bandwidths have to be selected which are not mutually independent. Apart from the date, the other water quality covariates behave as a random design, and therefore a variable bandwidth selector is more appropriate. For additive models, only a few references on fixed plug-in bandwidth estimators exist to our knowledge (Opsomer and Ruppert, 1998; Opsomer, 2000; Mammen and Park, 2005). The definition of variable plug-in bandwidths is even more complex. The use of nearest neighbourhood bandwidths in an AM context is relatively simple, and guarantees that each local regression uses an appropriate number of observations. The spans $s_1, \ldots, s_q$ are typically estimated by classical methods. In this respect, and given Loader's (1999a) comments, we will not go deeper into the problem of plug-in based bandwidth estimators and we use nearest neighbourhood bandwidths which are defined by the span in the remainder of this dissertation. We now explore the fitting procedure of additive models.

2.3 Fitting additive models

In practice, the backfitting algorithm proposed by Buja et al. (1989) is the most widely used method to estimate the additive components. From Equation (2.3) it

is obvious that each function can be written as

$$f_j(X_j) = Y - \alpha - \sum_{k \neq j} f_k(X_k) - \epsilon. \tag{2.22}$$

When f_j is a linear smoother with smoother matrix $\boldsymbol{S}_j$ and in the hypothetical case that the other predictor terms are known, f_j can be estimated as

$$\hat{\boldsymbol{f}}_j = \boldsymbol{S}_j\{\boldsymbol{y} - \boldsymbol{\alpha} - \sum_{k \neq j} \boldsymbol{f}_k\}, \tag{2.23}$$

where $\boldsymbol{f}_k$ is the $n \times 1$ vector $(f_k(x_{k1}), \ldots, f_k(x_{kn}))^T$ and $\boldsymbol{\alpha}$ is an $n \times 1$ vector $(\alpha, \ldots, \alpha)^T$. When only linear smoothers are used in the model, a similar expression can be used for each smoother. By combining all these expressions, the following set of equations has to be solved,

$$\begin{bmatrix} \boldsymbol{I} & \boldsymbol{S}_1 & \boldsymbol{S}_1 & \ldots & \boldsymbol{S}_1 & \mathbf{1} \\ \boldsymbol{S}_2 & \boldsymbol{I} & \boldsymbol{S}_2 & \ldots & \boldsymbol{S}_2 & \mathbf{1} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \boldsymbol{S}_q & \boldsymbol{S}_q & \boldsymbol{S}_q & \ldots & \boldsymbol{I} & \mathbf{1} \\ \mathbf{1}/n & \mathbf{1}/n & \mathbf{1}/n & \ldots & \mathbf{1}/n & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{f}_1 \\ \boldsymbol{f}_2 \\ . \\ . \\ . \\ \boldsymbol{f}_q \\ \alpha \end{bmatrix} = \begin{bmatrix} \boldsymbol{S}_1 \\ \boldsymbol{S}_2 \\ . \\ . \\ . \\ \boldsymbol{S}_q \\ \mathbf{1}/n \end{bmatrix} \boldsymbol{y} \tag{2.24}$$

where $\boldsymbol{I}$ is the $n \times n$ identity matrix and $\mathbf{1}$ is the $n \times 1$ vector $(1, \ldots, 1)^T$. The backfitting algorithm solves this set of equations iteratively. In the l^{th} iteration, $\boldsymbol{f}_j^{(l-1)}$ is updated by

$$\boldsymbol{f}_j^{(l)} = \boldsymbol{S}_j(\boldsymbol{y} - \boldsymbol{\alpha} - \sum_{k<j} \boldsymbol{f}_k^{(l)} - \sum_{k>j} \boldsymbol{f}_k^{(l-1)}). \tag{2.25}$$

In order to make each function identifiable, an additional constraint has to be introduced, $\sum_{t=1}^{n} f_j(x_{jt}) = 0$. This is simply done by replacing each $\boldsymbol{S_j}$ in Equations (2.23)-(2.25) by the centered smoother matrix $\boldsymbol{S_j^*} = (\boldsymbol{I} - \mathbf{1}\mathbf{1}^{\boldsymbol{T}}/n)\boldsymbol{S}_j$ (Kauermann and Opsomer, 2004). This also forces α to be estimated by the sample mean $\overline{y}$.

For a semi-parametric model, say $\boldsymbol{y} = \boldsymbol{f}_1 + \boldsymbol{X_m}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, which contains only one linear smoother, Hastie and Tibshirani (1990) showed that an explicit solution exists,

$$\hat{\boldsymbol{f}}_1 = \boldsymbol{S}_1(\boldsymbol{y} - \boldsymbol{X}_m\hat{\boldsymbol{\beta}}) \tag{2.26}$$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}_m^T(\boldsymbol{I} - \boldsymbol{S}_1)\boldsymbol{X}_m)^{-1}\boldsymbol{X}_m^T(\boldsymbol{I} - \boldsymbol{S}_1)\boldsymbol{y}, \tag{2.27}$$

provided that $(\boldsymbol{X}_m^T(\boldsymbol{I}-\boldsymbol{S}_1)\boldsymbol{X}_m)^{-1}$ exists. They also showed that for the bivariate additive model $Y = f_1(X_1) + f_2(X_2) + \epsilon$, the backfitting estimators converge to

$$\boldsymbol{f}_1^{(\infty)} = (\boldsymbol{I} - (\boldsymbol{I} - \boldsymbol{S}_1\boldsymbol{S}_2)^{-1}(\boldsymbol{I} - \boldsymbol{S}_1))\boldsymbol{y} \tag{2.28}$$

$$\boldsymbol{f}_2^{(\infty)} = (\boldsymbol{I} - (\boldsymbol{I} - \boldsymbol{S}_2\boldsymbol{S}_1)^{-1}(\boldsymbol{I} - \boldsymbol{S}_2))\boldsymbol{y}, \tag{2.29}$$

as the number of backfitting iterations approaches infinity and given that the norm $\|\boldsymbol{S}_1\boldsymbol{S}_2\| < 1$. The fit is then given by

$$\hat{\boldsymbol{y}} = \boldsymbol{f}_1^{(\infty)} + \boldsymbol{f}_2^{(\infty)} = (\boldsymbol{I} - (\boldsymbol{I} - \boldsymbol{S}_2)(\boldsymbol{I} - \boldsymbol{S}_2\boldsymbol{S}_1)^{-1}(\boldsymbol{I} - \boldsymbol{S}_1))\boldsymbol{y}, \tag{2.30}$$

which shows that $\hat{\boldsymbol{y}}$ is a linear combination of $\boldsymbol{y}$ with the $n \times n$ projection matrix $\boldsymbol{H} = (\boldsymbol{I} - (\boldsymbol{I} - \boldsymbol{S}_2)(\boldsymbol{I} - \boldsymbol{S}_2\boldsymbol{S}_1)^{-1}(\boldsymbol{I} - \boldsymbol{S}_1))$.

After fitting the model, predictions and point estimates of the smooth functions are obtained at each predictor combination. To assess their uncertainty, methods to derive variance estimators and confidence intervals are introduced in the next section.

2.4 Confidence intervals for additive models

Since we use linear smoothers, we can rely on techniques from classical linear regression to derive variance estimates and pointwise confidence intervals. Pointwise intervals have a local nature. They reflect the uncertainty associated with a particular predictor location. This will be done in Section 2.4.1. In Section 2.4.2 we will consider the bootstrap as a nonparametric procedure to derive the pointwise confidence bands and in Section 2.4.3 global confidence sets will be derived for additive models.

2.4.1 Variance estimator and pointwise confidence intervals

In classical parametric statistics, a variance estimate is the key element for statistical inference. Similar to linear regression, the residual sum of squares (RSS) can be used for variance estimation. The RSS is defined as usual, $RSS = (\boldsymbol{y} - \hat{\boldsymbol{y}})^T(\boldsymbol{y} - \hat{\boldsymbol{y}})$. In linear regression, the variance estimate then becomes $\hat{\sigma}^2 = RSS/df$, where its degrees of freedom (df) equals $n-(q+1)$, with $q+1$ the number of parameters

that have been estimated and n the number of observations. In the context of linear regression smoothers we already have used the trace of the smoother matrix, tr($\boldsymbol{S}$), as a definition of the degrees of freedom when we defined the GCV and the AIC criteria. Hastie and Tibshirani (1990) showed that it is better to use another definition for the degrees of freedom of the RSS. In the next paragraph their definition is explained in detail.

When all components of the AM are linear or linear smoothers, an $n \times n$ projection matrix $\boldsymbol{H}$ can be derived such that $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$. This is already illustrated in Equation (2.30) for the case of two linear smoothers. For nonparametric AM's using linear smoothers, the additive component functions can be solved by a set of normal equations presented in Equation (2.24). Equation (2.24) can thus also be written as

$$\hat{\boldsymbol{P}}\hat{\boldsymbol{f}} = \hat{\boldsymbol{Q}}\boldsymbol{y}, \tag{2.31}$$

where $\hat{\boldsymbol{f}}$ is the $nq + 1$ vector $\hat{\boldsymbol{f}} = (\boldsymbol{f}_1^T, \ldots, \boldsymbol{f}_q^T, \alpha)^T$. In general, a solution for Equation (2.24) is found by applying a backfitting algorithm. However, as Opsomer (2000) mentioned, it is possible, at least conceptually, to write the estimators directly as

$$\hat{\boldsymbol{f}} = \hat{\boldsymbol{P}}^{\boldsymbol{-1}}\hat{\boldsymbol{Q}}\boldsymbol{y}, \tag{2.32}$$

and after obtaining $\hat{\boldsymbol{P}}^{\boldsymbol{-1}}$, $\hat{\boldsymbol{y}}$ can be written as

$$\begin{aligned} \hat{\boldsymbol{y}} &= \begin{bmatrix} \boldsymbol{I} & \boldsymbol{I} & \boldsymbol{I} & \ldots & \boldsymbol{I} & \mathbf{1} \end{bmatrix} \hat{\boldsymbol{f}} \\ &= \begin{bmatrix} \boldsymbol{I} & \boldsymbol{I} & \boldsymbol{I} & \ldots & \boldsymbol{I} & \mathbf{1} \end{bmatrix} \hat{\boldsymbol{P}}^{\boldsymbol{-1}}\hat{\boldsymbol{Q}}\boldsymbol{y} \\ &= \boldsymbol{H}\boldsymbol{y}. \end{aligned} \tag{2.33}$$

Recall that $\boldsymbol{I}$ is the $n \times n$ identity vector and $\mathbf{1}$ is an $n \times 1$ vector of ones, and so $[\boldsymbol{I} \ \ldots \ \boldsymbol{I} \ \mathbf{1}]$ is an $n \times (qn + 1)$ matrix. From this derivation, it is clear that an additive model using linear smoothers is a linear smoother itself with an $n \times n$ projection matrix $\boldsymbol{H}$. Further, for linear smoothers, it can be shown that the RSS has the expectation $\mathrm{E}\,(RSS) = \{n - \mathrm{tr}(2\boldsymbol{H} - \boldsymbol{H}\boldsymbol{H}^{\boldsymbol{T}})\}\sigma^2 + \boldsymbol{b}^{\boldsymbol{T}}\boldsymbol{b}$, where $\boldsymbol{b}$ is the bias (Hastie and Tibshirani, 1990). The bias $\boldsymbol{b}$ is defined as $\boldsymbol{b} = \boldsymbol{m} - \mathrm{E}\,(\boldsymbol{H}\boldsymbol{y}) = \boldsymbol{m} - \boldsymbol{H}\boldsymbol{m}$. Thus, when the bias is negligible, the variance can be estimated by

$$\hat{\sigma}^2 = \frac{RSS}{n - \mathrm{tr}(2\boldsymbol{H} - \boldsymbol{H}\boldsymbol{H}^{\boldsymbol{T}})}, \tag{2.34}$$

where, in analogy with linear regression, the *degrees of freedom of the errors* can be defined as $df^{err} = n - \mathrm{tr}(2\boldsymbol{H} - \boldsymbol{H}\boldsymbol{H}^{\boldsymbol{T}})$.

When the residuals are i.i.d, the estimate of variance-covariance matrix of $\hat{\boldsymbol{y}}$ can be calculated as

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{y}}} = \boldsymbol{H}\boldsymbol{H}^{\boldsymbol{T}}\hat{\sigma}^2. \tag{2.35}$$

Similar to $\hat{\boldsymbol{y}}$, a projection matrix $\boldsymbol{H}_j$ can be defined for each component $\hat{\boldsymbol{f}}_j = \boldsymbol{H}_j\boldsymbol{y}$. The variance-covariance matrix of each component is simply obtained by replacing $\boldsymbol{H}$ in Equation (2.35) by $\boldsymbol{H}_j$.

The calculation of $\hat{\boldsymbol{P}}^{-\mathbf{1}}$ is not interesting from computational point of view since it involve inverting an $(nq+1)\times(nq+1)$ matrix. Moreover, the inverse of $\hat{\boldsymbol{P}}$ does not always exist. Recently, Giannitrapani et al. (2005) provided a very simple method to keep track of the important projection matrices while the backfitting algorithm proceeds. When linear smoothers are used, the estimate of each component $\boldsymbol{f}_j^{(l)}$ in the l^{th} iteration step can be written as $\boldsymbol{f}_j^{(l)} = \boldsymbol{H}_j^{(l)}\boldsymbol{y}$. Hence, the backfitting scheme can be expressed as

$$\boldsymbol{H}_j^{(l)} = \boldsymbol{S}_j^*(\boldsymbol{I} - \sum_{k<j}\boldsymbol{H}_k^{(l)} - \sum_{k>j}\boldsymbol{H}_k^{(l-1)}). \tag{2.36}$$

At each stage, the updated projection matrix $\boldsymbol{H}_j^{(l)}$ remains independent of $\boldsymbol{y}$. When the backfitting algorithm has converged, a set of projection matrices $\{\boldsymbol{H}_j, j = 1,\ldots,q\}$ is obtained. They can be used to estimate the individual components and the fitted values $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$, where $\boldsymbol{H} = \mathbf{1}\mathbf{1}^{\boldsymbol{T}}/n + \sum_{j=1}^{q}\boldsymbol{H}_j$. The variance estimators can now be used for construction of approximate $(1-\alpha)$ confidence intervals. The term approximate confidence interval is used because it only holds when the bias is negligible. Here the interval is only given explicitly for the estimator $\hat{y}_t$,

$$[\hat{y}_t - z_{(1-\frac{\alpha}{2})}\hat{\sigma}_{y_t}, \hat{y}_t + z_{(1-\frac{\alpha}{2})}\hat{\sigma}_{y_t}], \tag{2.37}$$

where $z_{(1-\frac{\alpha}{2})}$ is the $(1-\frac{\alpha}{2})$ percentile of the standard normal distribution and $\hat{\sigma}_{y_t}$ the square root of the t^{th} diagonal element of $\hat{\boldsymbol{\Sigma}}_{\hat{y}}$. The formulation of confidence bands for the component functions $f_j(x_{jt})$ is trivial. We still have to keep in mind that the intervals are only correct when the bias is negligible. When this is not the case, the additive model fit $\hat{\boldsymbol{y}}$ is a fit for $\boldsymbol{H}\boldsymbol{m}$ rather than for the true underlying surface $\boldsymbol{m}$ evaluated at the design points (Hastie and Tibshirani, 1990).

Pointwise confidence intervals are illustrated in Figure 2.6. Note that, the intervals are centred about the estimates.

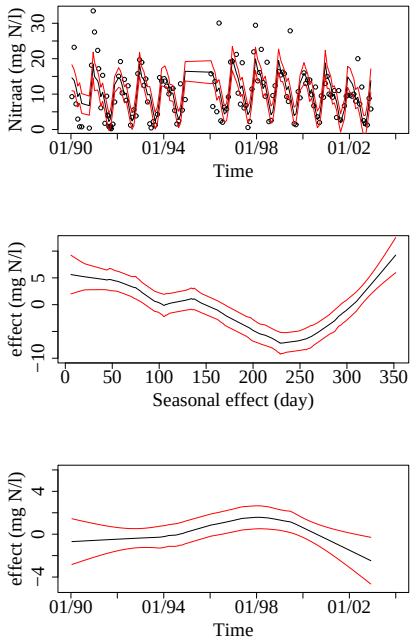

Figure 2.6: 95% pointwise confidence bands for an additive model with a seasonal component and a trend. Top panel: data and the model fit, Middle panel: contributions of the seasonal component, Bottom panel: contribution of the trend (fit in black and 95% pointwise confidence bands in grey)

2.4.2 Pointwise bootstrap confidence intervals

The bootstrap is a statistical inference technique that relies on only some weak distributional assumptions. Bootstrapping consists of resampling from a sample $\mathbf{D} = (D_1, \ldots, D_n)$, with replacement, to generate bootstrap replicates $\mathbf{D}^*(b)$, $b = 1, \ldots, B$, of the same size n. The bootstrap replicates are then used to simulate B estimates of a given statistic, resulting in an empirical probability distribution of the statistic. Suppose one wishes to estimate the empirical cumulative distribution function G of a statistic $\theta = t(\mathbf{D})$ which is estimated from a given sample $\mathbf{D} = (\mathbf{x_1}, \ldots, \mathbf{x_p}, \mathbf{y})$. Each observation $\mathbf{D_t} = (x_{1t}, \ldots, x_{qt}, y_t)$ is sampled with replacement, and with an equal probability of $1/n$. The sample $\mathbf{D}$ is resampled with replacement B times, until B bootstrap replicates $\mathbf{D}^*(b)$, $b = 1, \ldots, B$, are

generated. With each bootstrap replicate $\mathbf{D}^*(b)$, the statistic θ can be evaluated, yielding B bootstrap estimates $\hat{\theta}^*(b)$. The acquired empirical distribution $\hat{G}^*$ can also be used to calculate for instance the variance or confidence intervals on $\hat{\theta}$.

When applying the bootstrap in a regression context, there are two common approaches for generating bootstrap samples: (1) by resampling the cases $\mathbf{D_t} = (x_{1t}, \ldots, x_{qt}, y_t)$ or (2) by resampling the errors ($\hat{e}_t$). The method of resampling cases is not really an option, since it changes the sample design. Water quality data are gathered over time, and so the time covariate is not sampled at random. Environmental agencies commonly sample water quality data at intervals larger than two weeks. For such a sampling frequency, a large portion of the temporal dependencies are related to seasonality and trend (Van Belle and Hughes, 1984). These considerations provide a strong argument in favour of resampling residuals. In this case, bootstrap samples are generated by resampling from the empirical distribution of the residuals, say $\hat{F}$, and creating the bootstrapped responses

$$\mathbf{y}^*(b) = \hat{\mathbf{y}} + \mathbf{e}^*(b), \tag{2.38}$$

where $\mathbf{e}^*(b)$ is a bootstrap replicate of the errors. A bootstrap dataset is then constructed as $\mathbf{D}^*(b) = (\mathbf{x_1}, \ldots, \mathbf{x_q}, \mathbf{y}^*(b))$. The most straightforward method to obtain $\mathbf{e}^*(b)$ is to resample the crude errors $\hat{e}_t$. When a projection matrix $\mathbf{H}$ exists for the models, Davison and Hinkley (1997), however, suggested to sample the errors from the distribution of the centred adjusted residuals $r_t - \bar{r}$, where r_t is defined as

$$r_t = \frac{\hat{\epsilon}_t}{\sqrt{1 - h_{tt}}}, \tag{2.39}$$

where h_{tt} is the t^{th} diagonal element of the projection matrix $\mathbf{H}$ and $\bar{r}$ is the average of the r_t. For linear smoothers it can be shown that the variance of the estimated residuals $\hat{\epsilon}_t$ is equal to $\sigma^2(1 - h_{tt})$. Hence, resampling from the distribution of the centred adjusted residuals is preferred because they have the same variance as the true errors ϵ_t. Now that the bootstrap is introduced in the regression context, it can be applied for inference purposes. Suppose the aim is to construct a confidence interval for the fitted mean corresponding to a certain predictor combination $(x_{1t}, \ldots, x_{qt})$. Then the point estimate, $\hat{\theta} = t(\mathbf{D})$, is a prediction with the additive model $\hat{\theta} = t(\mathbf{D}) = \hat{m}(x_{1t}, \ldots, x_{qt})$.

A very natural way to calculate $1 - \alpha$ bootstrap confidence intervals, is to take the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ percentiles of the bootstrap distribution $\hat{G}^*$. This can be easily done by first ranking the $\hat{\theta}^*$'s into $\hat{\theta}^*_{(1)} \leq \ldots \leq \hat{\theta}^*_{(B)}$ and then take the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$

percentiles, giving the interval

$$[\hat{\theta}^*_{(\lfloor B(\frac{\alpha}{2})\rfloor)}, \hat{\theta}^*_{(\lfloor B(1-\frac{\alpha}{2})\rfloor+1)}]. \tag{2.40}$$

This interval is known as the bootstrap percentile interval. However, the coverages of these intervals are known to be problematic (Efron and Tibshirani, 1993; Davison and Hinkley, 1997). To improve the coverages, corrections have been proposed such as bias-corrected and accelerated bootstrap confidence intervals, which are referred to as the BC$_a$ intervals. Instead of taking the $(\alpha/2)^{th}$ and $(1-\alpha/2)^{th}$ percentile of the bootstrap distribution $\hat{G}^*$, the BC$_a$ interval is given by (Efron and Tibshirani, 1993)

$$[\hat{\theta}^*_{(\alpha_1)}, \hat{\theta}^*_{(\alpha_2)}], \tag{2.41}$$

where

$$\begin{aligned}\alpha_1 =& \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{(\frac{\alpha}{2})}}{1-\hat{a}(\hat{z}_0 + z_{(\frac{\alpha}{2})})}\right)\\ \alpha_2 =& \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{(1-\frac{\alpha}{2})}}{1-\hat{a}(\hat{z}_0 + z_{(1-\frac{\alpha}{2})})}\right).\end{aligned} \tag{2.42}$$

$$\tag{2.43}$$

Here $\Phi(.)$ indicates the standard normal cumulative distribution function and $z_{(\alpha/2)}$ is its 100 α^{th} percentile point. We still have to define $\hat{a}$ and $\hat{z}_0$. The bias correction can be easily calculated from the fraction of the bootstrap replications that is less than the plug-in estimate $\hat{\theta}$,

$$\hat{z}_0 = \Phi^{-1}(\#\{\hat{\theta}^*(b) < \hat{\theta}\}/B), \tag{2.44}$$

where $\Phi^{-1}(.)$ indicates the inverse of the standard normal cumulative distribution (Efron and Tibshirani, 1993; Davison and Hinkley, 1997). An easy way to compute the acceleration $\hat{a}$ is provided by Efron and Tibshirani (1993), using the jackknife values of the statistic $\hat{\theta} = t(\boldsymbol{D})$. A jackknife value for the t^{th} observation is obtained when the statistic is calculated on the original sample without the observation at time t. Let $\boldsymbol{D}_{(t)}$ represent the original sample without the t^{th} observation $\boldsymbol{D}_t$, $\hat{\theta}_{(t)} = t(\boldsymbol{D}_{(t)})$ and $\hat{\theta}_{(.)} = \sum_{t=1}^{n} \hat{\theta}_{(t)}/n$, then the acceleration is calculated as

$$\hat{a} = \frac{\sum_{t=1}^{n}(\hat{\theta}_{(.)} - \hat{\theta}_{(t)})^3}{6[\sum_{t=1}^{n}(\hat{\theta}_{(.)} - \hat{\theta}_{(t)})^2]^{3/2}}. \tag{2.45}$$

A third type of intervals which we will consider, are studentised bootstrap intervals. These intervals are acquired by computing for each bootstrap replicate $\boldsymbol{D}^*(b)$,

$$z^*(b) = \frac{\hat{\theta}^*(b) - \hat{\theta}}{\hat{\sigma}^*(b)}, \tag{2.46}$$

where $\hat{\sigma}^*(b)$ is the estimated standard error of $\hat{\theta}^*(b)$. The studentised bootstrap interval, after ordering the z^*'s to $z^*_{(1)} \leq \ldots \leq z^*_{(B)}$ is then given by

$$[\hat{\theta} - z^*_{((\lfloor 1-\frac{\alpha}{2})B\rfloor+1)}\hat{\sigma}, \hat{\theta} - z^*_{(\lfloor \frac{\alpha}{2}B\rfloor)}\hat{\sigma}]. \tag{2.47}$$

Davison and Hinkley (1997) showed that the studentised bootstrap confidence intervals and BC$_a$ intervals are preferred over bootstrap percentile intervals, both on the basis of empirical as well as theoretical arguments. Efron and Tibshirani (1993), however, argued that the studentised bootstrap confidence intervals are particularly applicable to location statistics, like the mean, median or percentiles. We will use the bootstrap for inference on location statistics such as the mean $m(X_1, \ldots, X_q)$ and for the location of the contributions of the components $\boldsymbol{f}_j$, so we do not expect problems related to the use of the studentised bootstrap.

Pointwise intervals for $\hat{\boldsymbol{y}}$ or $\hat{\boldsymbol{f}}_j$ can be obtained by applying the above bootstrap methods to each $\hat{y}_t$ or $\hat{f}_j(x_{jt})$. Pointwise bootstrap intervals are illustrated in Figure 2.7. Two different intervals are shown, percentile based bootstrap intervals and studentised bootstrap intervals. The intervals are fairly similar to each other and to the analytical intervals for the model fit and the seasonal component. For the trend, differences are observed at the peak of the curve. Here the percentile based confidence interval is shifted downwards and the studentised confidence interval is shifted upwards in comparison with the analytical confidence interval and the estimated curve. In this region a number of very high nitrate levels are observed, and most of the bootstrap replicates result in a trend effect, which is systematically lower than the estimated trend from the original dataset. The studentised interval corrects for this because it calculates the bootstrap replicates $z^*(b) = \frac{\hat{\theta}^*(b)-\hat{\theta}}{\hat{\sigma}^*(b)}$ and $\hat{\theta}^*(b)$ will be systematically lower than $\hat{\theta}$. This reflects the high nitrate levels which are observed. From Equation (2.47) it can be seen that this leads to an upward shift of the intervals in this region.

Sometimes it is useful to make simultaneous inference on more than one point in the covariate space (e.g. when checking whether a predictor function is significantly different from a least squares fit). Pointwise intervals are not appropriate for

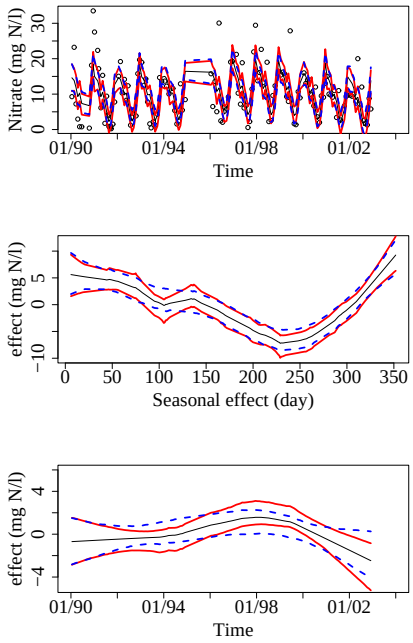

Figure 2.7: 95% bootstrap pointwise confidence intervals for an additive model with a seasonal component and a trend. Top panel: The data and the model fit, Middle panel: Contribution of the seasonal component and Bottom panel: Contribution of the trend. Dashed line: percentile based bootstrap intervals, Solid line: studentised bootstrap intervals

this purpose. Therefore we will now introduce global confidence sets for simultaneous inference about an entire regression curve or surface.

2.4.3 Global confidence sets

When we wish to infer on more than one function value at the same time, pointwise confidence intervals may be misleading. Suppose we would like to check if a straight line fits in a confidence band of one of the predictor functions, we need a kind of global confidence band. A common approach to go from pointwise confidence bands to global confidence bands is to make the pointwise bands wider to implicitly correct for multiple comparisons (Eubank and Speckman, 1993). How-

ever, Hastie and Tibshirani (1990) disagree with this approach. They motivate that a confidence set for the n values of the true underlying function is a set in an n-dimensional space and a global confidence band is a projection or a "shadow" of such a set onto each direction. Therefore the information of a confidence band is limited. Moreover, it gives no information on the functional shape of the members of the n-dimensional set. Hastie and Tibshirani (1990) also argued that a projection of an n-dimensional global confidence set into a confidence band does not have to be larger than the pointwise confidence bands. To construct such a confidence band, we should be able to construct curves that belong to the global confidence set. Those curves can then be used to construct the global confidence band.

We here discuss the approach of Hastie and Tibshirani (1990) to construct a confidence set for $\boldsymbol{g} = \boldsymbol{H}\boldsymbol{m}$. When the errors are assumed to be Gaussian, the likelihood ratio method for constructing such a set uses the approximate studentised pivotal $(\hat{\boldsymbol{y}} - \boldsymbol{g})^T(\boldsymbol{H}\boldsymbol{H}^{\boldsymbol{T}}\sigma^2)^{-1}(\hat{\boldsymbol{y}} - \boldsymbol{g})$ which is asymptotically χ^2-distributed. Since σ^2 is unknown, it has to be estimated. Hence the approximate pivotal

$$\nu = (\hat{\boldsymbol{y}} - \boldsymbol{g})^T(\boldsymbol{H}\boldsymbol{H}^{\boldsymbol{T}}\hat{\sigma}^2)^{-1}(\hat{\boldsymbol{y}} - \boldsymbol{g}) \tag{2.48}$$

should be used. Let G denote the distribution of ν. Hastie and Tibshirani (1990) showed that this distribution could be approximated based on the F-distribution or by using the bootstrap. According to their results, the bootstrapped approximation works better than the one based on the F-distribution. Hence, we restrict ourselves to the bootstrap approximation of G. The bootstrap is used to generate $\boldsymbol{y}^*(b)$ as described in the previous section and to calculate bootstrapped statistics $\hat{\boldsymbol{y}}^*(b) = \boldsymbol{H}\boldsymbol{y}^*(b)$, $\hat{\sigma}^{*2}(b) = RSS^*(b)/df^{err}$, and

$$\nu^*(b) = (\hat{y}^*(b) - \hat{y})^T(\boldsymbol{H}\boldsymbol{H}^{\boldsymbol{T}}\hat{\sigma}^{*2}(b))^{-1}(\hat{y}^*(b) - \hat{y}). \tag{2.49}$$

After ranking the $\nu^*(b)$'s so that $\nu^*_{(1)} \leq \ldots \leq \nu^*_{(B)}$, the $(1-\alpha)$ confidence set can be derived from the interval $[\nu^*_{(\lfloor B\alpha/2 \rfloor)}, \nu^*_{(\lfloor (1-\alpha/2) \rfloor + 1)}]$. All bootstrap replicates $\hat{\boldsymbol{y}}^*$'s which resulted in $\nu^*(b)$'s within this interval belong to the confidence set. A simultaneous confidence band can be displayed by creating an envelope which is containing these curves. Once an envelope is established, a projection of the envelope onto each direction can be made.

The simultaneous interval obtained after projecting the envelopes of the global bootstrap confidence sets is illustrated in Figure 2.8. The intervals are indeed fairly similar to the intervals shown in Figure 2.7 and support the findings of Hastie and Tibshirani (1990).

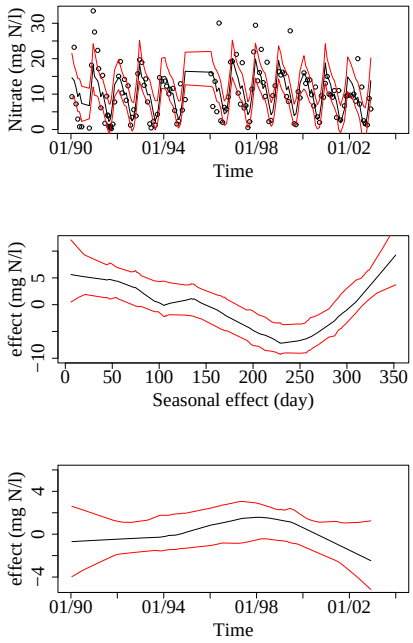

Figure 2.8: Envelopes for 95% global bootstrap confidence bands for an additive model with a seasonal component and a trend. Top panel: data and the model fit, Middle panel: the contribution of the seasonal component and Bottom panel: the contribution of the trend

2.5 Model selection

Although model selection is a fundamental part in the building process of statistical models, we will only give a very brief overview on model selection and restrict our attention to the procedure used in this dissertation.

Model selection for additive models is often performed in two stages: (1) bandwidth or span selection of the smoothing parameters $(s_1, \ldots, s_q)$, and (2) selection of predictor variables in the model. As mentioned in Section 2.2.4, nearest neighbourhood bandwidths are used for the local polynomial smoothers in the model. A neighbourhood contains a fixed fraction of the total number of observations n, this fraction is referred to as the span s_j. The spans $s_1, \ldots, s_q$ are typically tuned by

using criteria as the AIC, CV and GCV. In principle, numerical optimisers could be used for this purpose, but generally a grid search is used to determine the optimal spans. When the number of smoothers q increases, this leads to an exponential increase in the number of AM's to be evaluated. For variable selection, this procedure has to be further embedded in a model selection procedure, such as classical forward and backward stepwise selection techniques (e.g. Hastie and Tibshirani, 1990). In the forward approach, one starts with a one-dimensional model which contains the predictor that results in the best evaluation of the selection criterion. In each cycle, the model is extended with the predictor which results in the largest improvement of the criterion. The procedure stops when the criterion is not further improved by the addition of a predictor or when all predictors are entered in the model. The backward procedure starts with the most complex model and then leaves out, in each step, the predictor which results in the model with the best evaluation of the criterion. The algorithm proceeds as long as the criterion improves by the reduction of the model complexity. To ensure that the appropriate smoothing parameters are used in each step of both procedures, the smoothing parameters of each of the candidate models should be determined. When q gets large and when a dense grid is used for the selection of the smoothing parameters, this approach gets quickly computationally demanding. When the contributions of the predictors are orthogonal, the computational burden can be reduced. The multidimensional grid search can then be replaced by an iterative procedure where each iteration is a one-dimensional grid search to find the optimal s_j by keeping the other smoothing parameters (s_k, $k \neq j$) fixed.

Hastie and Tibshirani (1990) introduced the BRUTO algorithm as a pragmatic solution to keep the computational burden limited. The algorithm is an adaptation of the backfitting algorithm so that it combines model fitting, smoothing parameter selection and model selection. To avoid computational problems, Hastie and Tibshirani (1990) adjusted the GCV criterion

$$GCV(s_1, \ldots, s_p) = \frac{\sum_{t=1}^{n} \hat{\epsilon}_t^2}{n(1 - \text{tr}(\mathbf{H}(s_1, \ldots, s_p))/n)^2}, \tag{2.50}$$

to the modified GCV criterion,

$$GCV^b(s_1, \ldots, s_p) = \frac{\sum_{t=1}^{n} \hat{\epsilon}_t^2}{n(1 - [1 + \sum_{j=1}^{p} \{\text{tr}(\mathbf{S}_j(s_j)) - 1\}]/n)^2}. \tag{2.51}$$

In this way the computational difficulties are circumvented which are associated with the calculation of tr($\mathbf{H}(s_1, \ldots, s_p)$). But, as shown in the previous section,

Giannitrapani et al. (2005) provided a very simple method to keep track of the important projection matrices and when their approach is used in the backfitting algorithm, the projection matrix $\mathbf{H}$ is known and its trace can easily be acquired. Therefore, a modification of the GCV is not required and we have chosen to incorporate the original GCV criterion in the BRUTO algorithm.

The BRUTO algorithm starts with the null fit, where all projection matrices $\mathbf{H}_j = \mathbf{0}$, $j = 1, \ldots, q$. In each iteration one parameter s_j is selected. Hence, the span selection is performed one smoothing parameter at a time, while the other smoothing parameters remain unchanged. In particular the s_j is adjusted which minimises the global GCV. In the cycle (l) this is applied by using the appropriate smoothing parameter $s_j^{(l)}$ to update the projection matrix

$$\mathbf{H}_j^{(l)}(s_j^{(l)}) = \mathbf{S}_j^*(s_j)(\boldsymbol{I} - \sum_{k \neq j} \mathbf{H}_k^{(l-1)}(s_k^{(l-1)})), \tag{2.52}$$

while the other projection matrices are left unaltered. Hence, each iteration only provides for an update of one smoothing parameter s_j and its corresponding projection matrix $\mathbf{H}_j(s_j)$. The BRUTO algorithm is continued until the GCV converges. The convergence is guaranteed, because each iteration produces a decrease in the criterion. The BRUTO algorithm can easily be extended to incorporate model selection. When the GCV is allowed to be optimised by the selection of the null fit, $\mathbf{H}_j = \mathbf{0}$, it enables the removal of the associated explanatory variable from the model. Hence, a particular variable can be included at a certain iteration, its span can be adjusted in a next iteration and the variable can even be omitted from the model later on.

An example on how the BRUTO algorithm proceeds is given in Figure 2.9. The dataset consists of the response nitrate (NO_3^-) and 7 predictor variables: (1) Day number throughout the year, (2) Time, (3) temperature, (4) dissolved oxygen (DO), (5) nitrite (NO_2^-), (6) chemical oxygen demand (COD) and (7) pH. In Figure 2.9, the numbers in the plot indicate which of the predictors was adjusted in each cycle. During the first 4 cycles predictors 1, 6, 7 and 2 are included in the model. From the 5^{th} up to the 9^{th} cycle the spans of the selected predictors are adjusted. During cycle 10 and 11 predictors 5 and 4 are selected. And the last cycles consist of adjusting the spans of predictors 7 and 6. The final model includes predictors 1,2,4,5,6 and 7. Notice that the 3^{th} predictor is never included in the model. At first, the GCV decrease is steep due to the inclusion of extra predictors in the model. This is also reflected in the steep increase of the associated degrees of freedom.

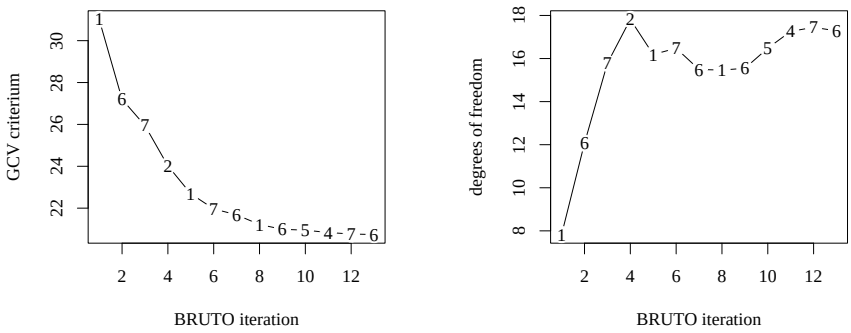

Figure 2.9: Left: Convergence of the GCV criterion in function of the iteration number of the BRUTO algorithm. Right: The evolution of the total degrees of freedom in the model in function the iteration number. The numbers along the curve indicate which of the predictors has been updated and/or included

2.6 Conclusions

A review of additive modelling is given with a special focus on its application to water quality data. Local polynomial smoothers were used as the additive functions of the additive model. The review covers the important issues of model structure, the selection of the smoothing parameters, the derivation of confidence intervals and a brief introduction to model selection.

For researchers who want to apply additive models, the main contribution of this review is that it explicitly includes all mathematical derivations needed to fit the models and to assess their uncertainty. In the existing statistical literature, procedures to obtain analytical pointwise confidence intervals are often only given implicitly. In this review, the analytical and bootstrapped pointwise confidence intervals are included in full detail.

Chapter 3

Data validation

3.1 Introduction

High quality data are essential for an adequate management of the water resources. Therefore, quality assurance is specifically mentioned as an important activity in the WFD guidance document on monitoring (EC, 2003; Højberg et al., 2007). Thus, new data have to be validated before they can be considered for a further evaluation of the water status. Observations can be suspicious due to the lack of the quality of the data, i.e. originating from errors introduced during the analysis in the laboratory, wrong calibration of the equipment or while entering the data. But it is also possible that they are due to a change in the system that causes changes in the water quality.

The detection of suspicious observations in environmental data is not straightfor-

ward because such data typically possess a complex nature. The observations may be dependent, non-normally distributed, may show cyclic variations, flow dependence, and, the trend and relations among the water quality variables may be non-linear (Hirsch et al., 1982; Cai and Tiwari, 2000; Dominici et al., 2002; Wood and Augustin, 2002; McMullan et al., 2003; McMullan, 2004). Therefore it is difficult for experts to validate the large amounts of water quality data originating from these monitoring networks. In this chapter we aim to provide a semi-automatic data validation procedure that can support experts at the environmental agencies to validate their large amounts of monitoring data. Before we elaborate on our data validation method, we first introduce some existing methods which might be used for this purpose. They consist of techniques from time series analysis and statistical process control.

3.1.1 Time series approach

One way to deal with the validation problem is to use models to predict future measurements based on the historical data. In time series literature, this is called forecasting. The new observations can then be compared with forecasts of the model. However, the use of point forecasts to compare with incoming observations is meaningless if the extent of associated uncertainty is unknown. Interval forecasts should be used instead. They provide more information on future uncertainty and take the sampling variance correctly into account. These intervals, characterised by an upper and lower limit, correspond to a specified coverage probability (Kim, 1999; Chatfield, 1993). In time series literature, AutoRegressive Moving Averages (ARMA) and AutoRegressive Integrated Moving Average (ARIMA) models are mainly used. However, in order to obtain stationarity, trends and seasonal variation have to be eliminated first. Subsequently the ARMA model is fitted to the stationary residual time series (Pourahmadi, 2001). The models are then used to compute a forecast and a forecast interval. To reduce the assumptions on the distribution of the residuals, bootstrap-based intervals were developed (Kim, 2004; Clements and Taylor, 2001; Kim, 1999; Chatfield, 1993). In an automated validation procedure, however, the interaction of the user should be limited. This requirement, makes the use of a classical time series approach such as ARMA, ARIMA or ARX difficult. They require expert knowledge to select the proper structure of the time series model. Moreover, the temporal dependence structure is also susceptible to change. One reason, for instance, is that the optimal structure can change over time as the database gets larger. We will now explore methods that are available in statistical process control.

3.1.2 Statistical process control

Statistical process control is developed within the context of quality control and the improvement of manufactured goods and services used by society. Typically, a product should be produced by a stable or repeatable process in order to meet the costumers expectations. In particular, the process must be able to guarantee that the quality of the product fluctuates with little variability around a certain target value (Montgomery, 2005). In this respect, the use of control charts is widespread. A control chart is a graphical tool which displays a certain quality characteristic that was measured in function of the sample number or the time. It contains a center line representing the average of the quality characteristic of the process when it is in control, and, an upper and lower control limit chosen in such a way that most sample points are expected to fall in between them. The following charts are commonly used:

- 'x'-charts which are plots of the observations themselves in function of time (Shewart, 1931).
- EWMA-charts, representing an exponentially weighted moving average of the measurements against time (first proposed by Roberts (1959)).
- CUMSUM-charts, where the cumulative sum of the differences between measurement and a target value is plotted against time (introduced by Page (1954)).
- MA-charts, plotting a moving average of the measurement series against time.

An example of an 'x'-chart and an EWMA-chart applied to the nitrate series at sampling location S5 along the river Yzer is given in Figure 3.1. The 'x'-plot is designed as such that there is a small chance to detect an out-of-control signal when the process is in control, and to have a higher change on a signal when the process is out of control. When only one observation is available at each time instant, the 'x'-plot consists of the individual measurements. The centerline of the chart is the overall process mean when the process is in control, and is assumed to be known. The lower and upper control limit are usually a constant L standard deviations, σ, below and above the centerline (Wardell et al., 1992). For EWMA-charts, the exponentially weighted moving average is defined as $z_i = \gamma y_i + (1-\gamma) z_{i-1}$, where γ is a weight constant between 0 and 1, and the starting value $z_0 = \mu$. Hence,

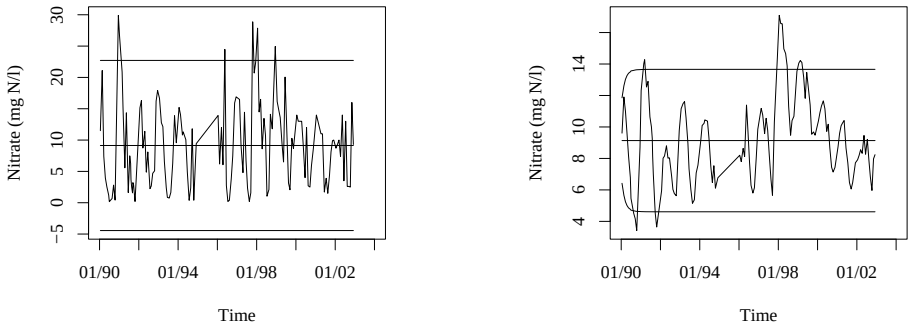

Figure 3.1: Examples of univariate monitoring charts, 'x'-chart (left), EWMA chart (right)

more recent data are receiving heavier weights. EWMA-charts detect shifts in mean more quickly than the 'x'-charts. The centerline is again the process mean, μ and the control limits are $\mu \pm L\sigma\sqrt{\gamma[1-(1-\gamma)^{2i}]/(2-\gamma)}$ (Wardell et al., 1992; Montgomery, 2005). For the construction of these charts the measurements are assumed to be i.i.d and Gaussian. The EWMA charts, however, are known to be robust to deviations of normality (Montgomery, 2005). They are based on a weighted sum of the measurements, and thus allow the use of the central limit theorem when the measured series is long. Environmental agencies often sample the river water quality data at intervals that are larger than two weeks. In this case the observations are often assumed to be independent when seasonality and trend are considered (Van Belle and Hughes, 1984). However seasonality and trend are not taken into account when constructing the basic monitoring charts. Hence, the i.i.d assumption is violated when the monitoring charts are based on the original nitrate measurements. Many authors have reported that this will lead to a false rejection of the data if they are positively correlated (e.g. Montgomery, 2005; Alwan, 1992; Montgomery and Mastrangelo, 1991). To overcome this problem, two general approaches exist to monitor autocorrelated processes. On the one hand, the autocorrelation can be modelled and standard charts are constructed with control limits that have been adjusted to account for the autocorrelation. On the other hand, a time series model can be fitted to the data and the residuals or forecast errors from this model can be used in a control chart (Montgomery, 2005; Reynolds and Lu, 1997). In the latter approach, a quality characteristic y_t is modelled as $y_t = \mu_t + \epsilon_t$

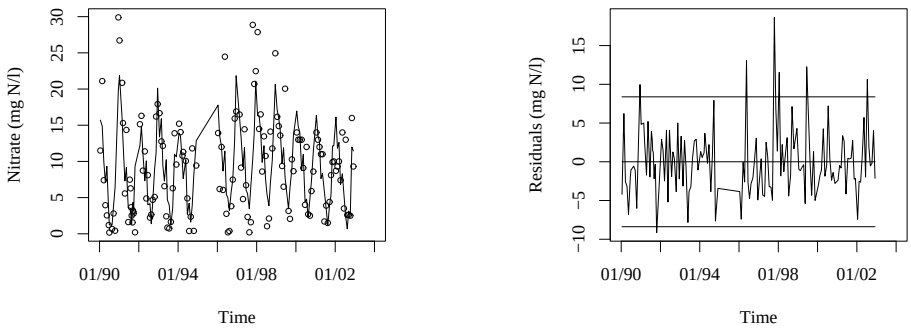

Figure 3.2: Model based univariate monitoring chart: Nitrate series together with the fit of an additive model with a trend and seasonal component (left), Residual based 'x'-chart (right)

where the ϵ_t's are assumed to be i.i.d and Gaussian, say $\epsilon_t \sim N(0, \sigma^2)$. The center line for the residuals is located at 0 and upper and lower control limits are located at $LCL_t = -2\sigma$ and $UCL_t = 2\sigma$. This approach is represented in Figure 3.2. To remove the serial correlation, the nitrate concentration was modelled using the additive models introduced in Chapter 2. Local linear regression smoothers for the trend and the seasonal component were used. The model was fitted by applying the BRUTO algorithm. In the left panel the nitrate data are given along with the model fit. The model-based control chart is represented in the right panel.

In chemical and environmental engineering, history-based methods that require a large amount of historical data are often used. They consist of neural networks or multivariate statistical techniques (e.g. Venkatasubramanian et al., 2003; Penny, 1996; Yoo et al., 2004, 2007). Multivariate process control is mainly based on projection methods. The multivariate observations are then projected on a lower dimensional space which can explain the main features in the multivariate data. Principle component analysis (PCA) is one of the widely used methods for this purpose. The standard multivariate methods imply the presence of a constant number of variables measured simultaneously. However, in many databases not all variables are measured at each time instant. In Flanders, for instance, nutrients are measured with a higher frequency than heavy metals. In this dissertation, we restricted our attention to univariate methods. Compared to multivariate approaches,

univariate control charts will only detect water quality measurements located at the endpoints of the univariate distribution as anomalous, while multivariate approaches can also detect malicious observations in the middle of the univariate distribution in case there is something wrong with their relationship with the other water quality variables.

Classical univariate control charts are not suited for water quality data due to the trend, cyclic variations and other forms of temporal dependences. Model-based control charts can correct for this, because a model can be used to detrend the data and to remove other forms of dependences. But, we think that an important issue is not addressed when using residual-based control charts, because in most applications the model uncertainty is ignored. Another general drawback of the classical methods is that they rely heavily on distributional assumptions and are parametric. The complex nature of water quality data, however, makes it inappropriate to use these existing methods for the validation of new observations.

In this chapter, we introduce a new semi-automatic data validation procedure. In Section 3.2 the method is introduced. A flowchart of the method is presented in Figure 3.3. First, knowledge is extracted from the historical data by the use of a model. To deal with the nonlinear character of the data and to enable an appropriate flexibility of the method towards changes in the process, nonparametric additive models (AM's) are proposed. Next, the AM is used to construct a prediction interval (PI) for a new observation at time $n + 1$. If the new observation is included in the PI, the observation is accepted and can be added to the historical data. Otherwise the observation is rejected and has to be passed on to an expert for further evaluation. Analytical and bootstrap based PI's are proposed. They incorporate both the model uncertainty due to the estimation of the mean model, as well as the additional uncertainty associated with single observations that are typically fluctuating around the modelled mean. In contrast to techniques from time series analysis and statistical process control, the procedure is entirely nonparametric when bootstrapping is used. This reduces the number of assumptions considerably. Since other physico-chemical variables are allowed in the model as predictor variables, it is possible that an outlier in one of these variables results in a false rejection of the incoming response data: A predictor has an additive effect on the outcome of the model, and outliers can result in an extreme value of the predictor function, resulting in a shift in the PI. At first sight this looks like an anomaly of our methodology. However, diagnostic plots to detect such shifts are presented. Moreover, in a practical implementation all the $q+1$ observed variables acquired at time $n + 1$ have to be validated. This is done by repeating the proce-

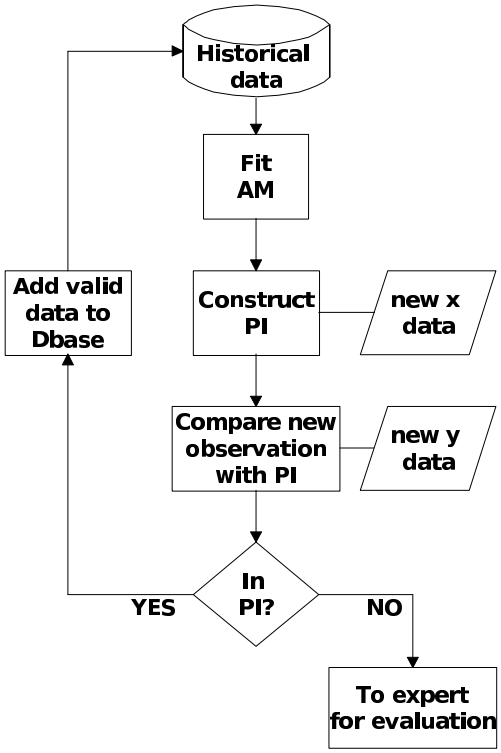

Figure 3.3: Flowchart of the data validation procedure

dure $q+1$ times and each time taking another variables to be the response and the q remaining variables to be the predictors. Due to the use of other water quality variables as predictors, our method also can detect suspicious observations located at the middle of the univariate distribution when there is something wrong with their relationship with the other water quality variables. In Section 3.3 the entire methodology is first illustrated on a real data case. The model that is obtained, is then used to generate synthetic data for a simulation study and a power study. These studies are used to check the coverage and the performance of the prediction intervals. Finally, the method is applied to two case studies to validate the nitrate data of Yzer basin measured in 2003 and 2004.

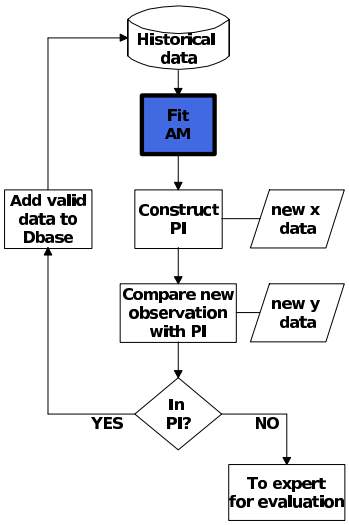

Figure 3.4: Flowchart of the data validation procedure, the fitting step is highlighted

3.2 Methods

In section Section 3.2.1 we start with an outline of the modelling procedure. In Section 3.2.2, the prediction intervals to validate new incoming observations are constructed. Finally, we present diagnostic plots to asses observations that are rejected by the prediction intervals.

3.2.1 Additive modelling of the historical data

The main idea of the procedure is to use the historical data to confront new observations with. In our approach, the information in the historical data is summarised into a fitted model. The position of this modelling step is indicated on the flowchart in Figure 3.4. The model should be able to capture the nonlinear relations between the water quality variables and should also adapt to changes in these relationships. In this respect, nonlinear models such as AM's are commonly used in environmen-

tal applications (e.g. Dominici et al., 2002; Wood and Augustin, 2002; Cai and Tiwari, 2000; McMullan et al., 2003). To be fully functional for the environmental agencies, the interaction of the user should be limited. Thus, the model fitting and selection procedure should be completely data driven, and the methods should rely on a minimum number of assumptions. This requirement, makes the use of a classical time series approach, such as ARMA, ARIMA or ARX, difficult, since expert knowledge is typically needed to select the proper structure of the time series model. Moreover, the structure is also susceptible to change. The optimal temporal structure can for instance change over time as the database gets larger. Further, the data of environmental agencies are often based on samples that are collected at time intervals that are larger than two weeks. With such a sample frequency a large part of the temporal dependence is due to trend and seasonal variations. Water quality data are often considered to be independent when seasonality and trend are accounted for (Van Belle and Hughes, 1984). Finally, it is also desirable that the method can assist the operator to gain insight in the relations between the water quality variables. This can be of great value for the in-depth analysis of rejected data.

Given these considerations, we propose to use additive models for the description of the historical data. They were introduced in Chapter 2. Suppose q predictor variables x_{jt}, $j = 1, \ldots, q$, and a response variable y_t are sampled at times $t = 1, \ldots, n$ and let $\boldsymbol{x}_j$ be an $n \times 1$ vector $\boldsymbol{x}_j = (x_{j1}, \ldots, x_{jn})^T$ and $\boldsymbol{y}$ be an $n \times 1$ vector $\boldsymbol{y} = (y_1, \ldots, y_n)^T$. Then a typical dataset can be represented by an $n \times (q+1)$ matrix $\boldsymbol{D} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_q, \boldsymbol{y})$. Further, $\boldsymbol{y}$ is assumed to be normally distributed with a conditional mean $\mathrm{E}(\boldsymbol{y}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_q) = \boldsymbol{m}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_q)$ and a constant variance σ^2. In the additive model framework, the regression surface $\boldsymbol{m}(.)$ is approximated by the sum of q additive functions and $\boldsymbol{y}$ is modelled by $\boldsymbol{y} = \boldsymbol{\alpha} + \sum_{j=1}^{q} \boldsymbol{f}_j + \boldsymbol{\epsilon}_i$, where the $n \times 1$ vector $\boldsymbol{f}_j = (f_j(x_{j1}), \ldots, f_j(x_{jn}))^T$ contains the contributions corresponding to each x_{jt}. This structure allows additive models to possess a nice interpretation feature. Once the model is fitted, the predictor effects can be studied separately. This enables the operator to get a simple graphical representation of the relationships between the response and each of its predictors (conditional on the other predictors in the model). But a price has to be paid for this additivity, i.e. the model will always remain an approximation of the true regression surface, but hopefully a good one.

Since we want to avoid an ARIMA-like approach, the predictors in the model should be capable of capturing the temporal dependence which is present in the original time series. Therefore we will model the response $\boldsymbol{y}$ by the use of a sea-

sonal effect coded by the day number (1-365), a long-term trend, the temperature and several other water quality variables. Because we want the data to drive the functional relationship between the predictor variables and the response, we use local polynomial regression smoothers to model each relation between a predictor and the mean response. In Section 2.2.3 local polynomial smoothers were introduced. Fan (1992) showed that the local linear regression smoother is the best among linear smoothers. Fan (1992), Fan and Gijbels (1996), and Hasti and Loader (1993) also showed that local polynomial regression adjusts automatically for bias at the boundary and are design adaptive in the sense that they also adjust for bias in regions where the predictors are nonuniform. As another advantage, they also enable straightforward generalisations of classical statistical inference procedures (Cleveland and Devlin, 1988; Fan and Gijbels, 1996; Loader, 1999b). For local polynomial regression smoothers the degree of smoothness is determined by the bandwidth. A choice has to be made between fixed or variable bandwidths. In Section 2.2.4 we have motivated the use of nearest neighbourhood bandwidths. In this case, the size of the neighbourhood is determined by the span, which is a fraction of the total number of data points. On data-rich locations this results in smaller bandwidths, and in data-sparse regions larger bandwidths are used. Since we are interested in the mean model (degree 0), the degree of the local polynomial is chosen to be 1, following Fan and Gijbels (1996) recommendations to use the lowest odd order for the local polynomial (see also Section 2.2.3)

Model selection is a crucial step in the construction of a new model. Here, the model selection involves the selection of the predictor variables and the associated bandwidths of the local linear smoothers. As shown in Section 2.5, the BRUTO algorithm can be used for both model fitting, model selection and tuning of the smoothing parameters. From a practical point of view, this is computationally interesting since the additive model only has to be fitted once. Other model selection algorithms often require fitting multiple candidate models and tuning their corresponding smoothing parameters. The model selected by the BRUTO algorithm is subsequently used for the validation of the new observation. The modelling procedure is illustrated extensively on a real data case in Section 3.3.1.

In the flowchart represented in Figure 3.4 it can be seen that the model is rebuilt each time a new observation is validated and added to the database. Hence, the BRUTO algorithm is executed each time the data series is extended. This approach ensures an optimal model fit at each time instant, i.e. as the data set grows larger, better approximations of the underlying surface m and a lower variance estimator are often obtained. Because the model is used for the construction of prediction

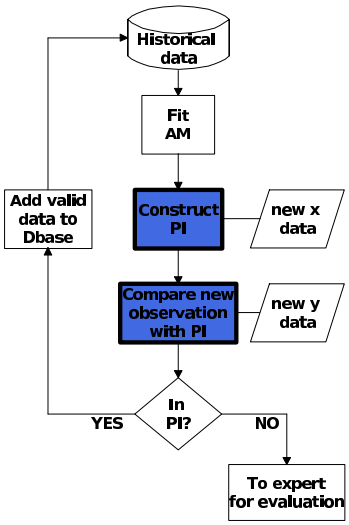

Figure 3.5: Flowchart of the data validation procedure. The construction and the application of the PI are highlighted

bands, a smaller variance ensures a better detection of suspicious observations. Methods to obtain these intervals are given in the next section.

3.2.2 Prediction intervals

To validate new data, a prediction interval (PI) is constructed and the data are considered valid if it is located within the PI. These steps in the data validation process are indicated on the flowchart in Figure 3.5.

A PI, however, differs from the pointwise confidence intervals for the mean derived in Section 2.4. A confidence interval reflects how accurate the mean is estimated. The data validation procedure, however, requires an interval estimate associated with the location of a new single observation. Under the normality assumption, the conditional distribution of an observation at time $n+1$, given the covariates, is $N(m(\boldsymbol{x}_{n+1}), \sigma^2)$. Hence, the prediction interval has to incorporate both the model uncertainty due to the estimation of $m(\boldsymbol{x}_{n+1})$ and the additional variability (σ^2)

associated with single observations that fluctuate around the mean.

Two different approaches are presented to construct prediction intervals: an analytical procedure which only works for AM's with linear smoothers and assumes the errors to be Gaussian, and double bootstrap procedures that work under less stringent conditions. The latter are fully nonparametric and they can cope with any type of AM and non-Gaussian errors. Both methods assume that the residuals are independently distributed. The data used in this study is based on monthly grab samples. When the water quality data are sampled at intervals larger than two weeks, a large amount of the dependences are known to be only related to seasonality and trend (Van Belle and Hughes, 1984). Additionally other water quality variables are used as predictors and they can also model a part of the temporal dependence. Another assumption is that the bias of the estimator is negligible. In the presence of bias, the variance estimate is inflated and this would result in more conservative interval estimates (e.g. Giannitrapani et al., 2005).

3.2.2.1 Analytical prediction intervals

Before the analytical PI's can be constructed, an estimator of the variance of a new prediction is needed. As shown in Section 2.4.1, a projection matrix exists when the AM is build up by linear smoothers. In this case, the prediction by the smoother at a certain predictor value is always a linear combination of the observed values of the responses. From Section 2.2.3 we know that for local linear smoothers (first order polynomial), the prediction corresponding to a predictor value x_0 is

$$\begin{bmatrix} 1 & 0 \end{bmatrix} (\boldsymbol{x}_c^T \boldsymbol{W}_0 \boldsymbol{x}_c)^{-1} \boldsymbol{x}_c^T \boldsymbol{W}_0 \boldsymbol{y}. \tag{3.1}$$

Thus its corresponding (row)smoothing vector can be written as

$$\boldsymbol{S}_0 = \begin{bmatrix} 1 & 0 \end{bmatrix} (\boldsymbol{x}_c^T \boldsymbol{W}_0 \boldsymbol{x}_c)^{-1} \boldsymbol{x}_c^T \boldsymbol{W}_0. \tag{3.2}$$

In the additive model q smoothers are used and to make each function identifiable, an additional constraint was introduced, $\sum_{t=1}^{n} f_j(x_{jt}) = 0$, $j = 1, \ldots, q$. Let $\boldsymbol{S}_{k,n+1}$ be a similar row smoothing vector for the k^{th} smoother evaluated in x_{kn+1}. To calculate the contribution of the k^{th} predictor at time $n+1$ its centered smoothing (row)vector is needed. In Section 2.3 it was shown that the $n \times n$ centered smoother matrix $\boldsymbol{S}_{\boldsymbol{j}}^{*} = (\boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^{\boldsymbol{T}}/n)\boldsymbol{S}_j$ are used for that purpose. The k^{th} centered $1 \times n$ smoothing (row)vector corresponding to x_{kn+1} is given by

$\boldsymbol{S}^*_{k,n+1} = \boldsymbol{S}_{k,n+1} - \mathbf{1}^T\boldsymbol{S}_k/n$. Similar as in Equation 2.23 an estimate of the contribution of the k^{th} predictor function $\hat{f}_{k,n+1}$ of the additive model is given by

$$\begin{aligned}\hat{f}_{k,n+1} &= \boldsymbol{S}^*_{k,n+1}(\boldsymbol{y} - \boldsymbol{\alpha} - \sum_{k\neq j}\hat{\boldsymbol{f}}_j) \\ &= \boldsymbol{S}^*_{k,n+1}(\boldsymbol{I} - \sum_{k\neq j}\boldsymbol{H}_j)\boldsymbol{y} \\ &= \boldsymbol{H}_{k,n+1}\boldsymbol{y}. \end{aligned} \tag{3.3}$$

The estimate of the mean response at time $n+1$, $\hat{y}_{n+1}$, then becomes

$$\begin{aligned}\hat{y}_{n+1} &= \hat{\alpha} + \sum_{j=1}^{q}\hat{f}_{j,n+1} \\ &= (\mathbf{1}^T/n + \sum_{j=1}^{q}\boldsymbol{H}_{j,n+1})\boldsymbol{y} \\ &= \boldsymbol{H}_{n+1}\boldsymbol{y}, \end{aligned} \tag{3.4}$$

and its variance is thus

$$\sigma^2_{\hat{y}_{n+1}} = \boldsymbol{H}_{n+1}\boldsymbol{H}^T_{n+1}\sigma^2. \tag{3.5}$$

This variance refers to the uncertainty associated with prediction of the mean of the new observation at time $n+1$, and not to the variance of a new single observation. The variance needed for the construction of a PI of a new single observation is decomposed into a part related to the uncertainty of the modelled mean, $\sigma^2_{\hat{y}_{n+1}}$, and into the part due to residual variance, σ^2. Thus, the variance for calculating a PI becomes

$$\sigma^2_{y_{n+1}} = (\boldsymbol{H}_{n+1}\boldsymbol{H}^T_{n+1} + 1)\sigma^2, \tag{3.6}$$

and σ^2 is estimated as in Equation (2.34). After plugging this into Equation (3.6), a $1-\alpha$ PI is given by

$$[\hat{y}_{n+1} - z_{(1-\frac{\alpha}{2})}\hat{\sigma}_{y_{n+1}}, \hat{y}_{n+1} + z_{(1-\frac{\alpha}{2})}\hat{\sigma}_{y_{n+1}}], \tag{3.7}$$

and $z_{1-\alpha/2}$ is the $1-\alpha/2$ percentile from the standard normal distribution. This analytical PI is also referred to as aPI.

3.2.2.2 Bootstrap intervals

In general, the estimation of additive models does not have an analytical solution and the errors can deviate from normality. The analytical intervals as described

in Section 3.2.2.1, only exists when linear smoothers are used as building blocks and their coverages are only correct when the errors are Gaussian. In this section, a procedure is proposed for the construction of the prediction intervals that can cope with additive models in general. Unfortunately, an analytical derivation does not exist for the PI for the general case and it implies the use of computationally intensive methods for variance estimation such as bootstrapping. The use of the bootstrap, however, has the advantage that it does not impose strong parametric assumptions on the distribution of the errors.

A general introduction to the bootstrap in a regression context is given in Section 2.4.2. It was used to approximate the distribution of an certain statistic $\hat{\theta} = t(\boldsymbol{D})$. Here the aim is to construct a prediction interval for a new single observation. Hence we put $\hat{\theta} = t(\boldsymbol{D}) = \hat{y}_{n+1}$. In Section 2.4.2 we have motivated to generate bootstrap samples by resampling the errors ($\hat{e}_i$). Resampling cases is not really an option, since it changes the sample design. Water quality data are gathered over time, and so the time covariate is not sampled at random. In this case, bootstrap samples are generated by resampling from the empirical distribution of the residuals, say $\hat{F}$, and creating bootstrapped responses by

$$\boldsymbol{y}^*(b) = \hat{\boldsymbol{y}} + \boldsymbol{e}^*(b), \tag{3.8}$$

where $\boldsymbol{e}^*(b)$ is a bootstrap replicate of the residuals. A bootstrap dataset is then constructed as $\boldsymbol{D}^*(b) = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_q, \boldsymbol{y}^*(b))$. The most straightforward method to obtain $\boldsymbol{e}^*(b)$ is to resample the crude errors $\hat{e}_i$. When a projection matrix $\boldsymbol{H}$ exists for the models, Davison and Hinkley (1997), however, suggested to sample the residuals from the distribution of the centred adjusted residuals $r_t - \bar{r}$, where r_t is defined as

$$r_t = \frac{\hat{\epsilon}_t}{\sqrt{1 - h_{tt}}}, \tag{3.9}$$

where h_t is the t^{th} diagonal element of the projection matrix $\boldsymbol{H}$ and $\bar{r}$ is the average of the r_t.

In Section 2.4.2 we have derived a bootstrap procedure to construct confidence intervals. Here, the aim is to construct a prediction interval on a single new observation. Hence, two sources of variability are involved in the derivation of the PI: the uncertainty due to the model prediction and the variability of the residuals. Therefore a double bootstrap procedure is needed. The main loop takes the variability of the model estimator into account. The second loop adds the additional variability that is associated with a single observation. Two types of bootstrap intervals are considered: a percentile based PI and a standardised prediction error based PI, where the prediction error δ_{n+1} is defined by $\delta_{n+1} = \hat{y}_{n+1} - y_{n+1}$.

The percentile method proceeds as follows:

1. Fit the additive model to the historical dataset $\boldsymbol{D}$
2. Use the fitted model to calculate the prediction $\hat{y}_{n+1}$
3. Extract the empirical distribution $\hat{F}$ of the residuals
4. First bootstrap loop: For $b_1 = 1, \ldots, B_1$
   (a) Take a bootstrap sample of the residuals $\boldsymbol{e}^*(b_1)$ and construct a bootstrapped response $\boldsymbol{y}^*(b_1)$ by adding these residuals to the fitted values of the AM, $\boldsymbol{y}^*(b_1) = \hat{\boldsymbol{y}} + \boldsymbol{e}^*(b_1)$. The bootstrapped dataset $\boldsymbol{D^*}(b_1)$ now becomes $\boldsymbol{D^*}(b_1) = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_p}, \boldsymbol{y^*}(b_1))$.
   (b) Fit an AM model to $\boldsymbol{D}^*(b_1)$, and compute the bootstrapped prediction $\hat{y}^*_{n+1}$.
   (c) Second bootstrap loop: For $b_2 = 1, ..., B_2$
      i. Sample at random a residual $e^*(b_2)$ from the empirical distribution of the residuals ($\hat{F}$).
      ii. The bootstrap estimate $\hat{\theta}^*(b_1, b_2)$ for the new observation is given by $\hat{\theta}^*(b_1, b_2) = \hat{y}^*_{n+1} + e^*(b_2)$.
5. $1 - \alpha$ confidence intervals are calculated from the bootstrap distribution of $\hat{\theta}^*$, say $\hat{G}^*$. First the $\hat{\theta}^*$'s are ordered so that $\hat{\theta}^*_{(1)} \leq \ldots \leq \hat{\theta}^*_{(B_1 B_2)}$. The interval is obtained by taking the $\alpha/2$ and $1 - \alpha/2$ percentiles of $\hat{G}^*$ (Efron and Tibshirani, 1993) and is denoted as

$$[\hat{\theta}^*_{(\lfloor B_1 B_2 \frac{\alpha}{2} \rfloor)}, \hat{\theta}^*_{(\lfloor B_1 B_2 (1-\frac{\alpha}{2}) \rfloor + 1)}]. \tag{3.10}$$

This percentile bootstrap PI is referred to as the %bPI.

Davison and Hinkley (1997) showed for linear models that the PI also can be estimated by computing the bootstrap distribution of the studentised predictions errors, $z = \delta/\hat{\sigma}$, mimicking the standard normal theory, where the prediction error $\delta_{n+1} = \hat{y}_{n+1} - y_{n+1}$ and $\hat{\sigma} = \sqrt{(RSS/df^{err})}$. This idea can easily be adopted to additive models and require steps 4 and 5 of the main bootstrap loop to be replaced by

4. First bootstrap loop: For $b_1 = 1, \ldots, B_1$

   (a) Take a bootstrap sample of the residuals $\boldsymbol{e}^*(b_1)$ and construct a bootstrapped response $\boldsymbol{y}^*(b_1)$ by adding this residuals to the fitted values of the AM. $\boldsymbol{y}^*(b_1) = \hat{\boldsymbol{y}} + \boldsymbol{e}^*(b_1)$. The bootstrapped dataset $\boldsymbol{D^*}(b_1)$ now becomes $\boldsymbol{D^*}(b_1) = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_p}, \boldsymbol{y^*}(b_1))$.

   (b) Fit an AM model to $\boldsymbol{D}^*(b_1)$, and compute the bootstrapped prediction $\hat{y}^*_{n+1}$ and the standard deviation of the corresponding residuals, $\hat{\sigma}^*(b1)$.

   (c) Second bootstrap loop: For $b_2 = 1, ..., B_2$

      i. Sample at random a residual $e^*(b_2)$ from the empirical distribution of the residuals ($\hat{F}$).

      ii. Compute the standardised prediction error $z^*(b_1b_2) = \delta^*(b_1b_2)/\hat{\sigma}^*(b_1)$ with $\delta^*_{n+1}(b_1b_2) = \hat{y}^*_{n+1} - (\hat{y}_{n+1} + e^*(b_2))$.

5. The bootstrap prediction interval, after ranking the z^*'s to $z^*_{(1)} \leq \ldots \leq z^*_{(B_1B_2)}$ is given by

$$[\hat{y}_{n+1} - \hat{\sigma} z^*_{(\lfloor (B_1B_2)(1-\frac{\alpha}{2}) \rfloor + 1)}, y_{n+1} - \hat{\sigma} z^*_{(\lfloor (B_1B_2)\frac{\alpha}{2} \rfloor)}]. \tag{3.11}$$

This standardised prediction error based bootstrap PI is referred to as sbPI.

3.2.3 Diagnostic plots

When an observation is rejected it has to be passed on to an expert for further evaluation. This step requires the interaction of the user and is indicated in the flowchart of Figure 3.6. There are several possible causes for the rejection of incoming data, such as changes in the system, illegal spills, errors during the analysis in the laboratory, wrong calibration of the equipment, outliers in the predictor variables and so on. Since other physico-chemical variables are present in the model as predictor variables, it is possible that an outlier in one of these variables results in a false rejection of the incoming response data: A predictor has an additive effect on the outcome of the model, and outliers can result in an extreme value of the predictor function, resulting in a shift in the PI. At first sight this looks like an anomaly of our methodology. However, such shifts can be detected by simply leaving the predictor out of the model: If the prediction was performed at an outlying observation in a particular predictor variable, the interval will shift back when this predictor variable is omitted from the model. The plots of the PI's made with these reduced

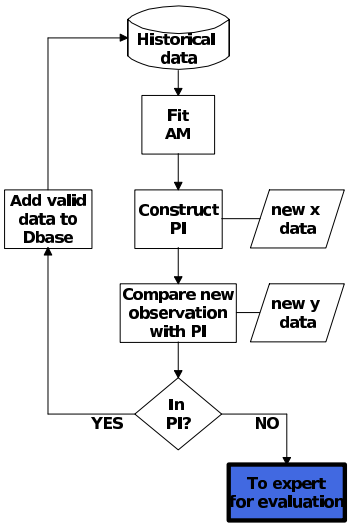

Figure 3.6: Flowchart of the data validation procedure. The expert evaluation stage is highlighted

models can assist the expert in his/her evaluation of rejected data. The use of the diagnostic plots is illustrated in the case study in Section 3.3.4.

3.3 Results and discussion

The data that we use in this section all belongs to the Yzer catchment. A description of the catchment can be found in Section 1.2. First the entire methodology is illustrated on a real data case. The results of this case are then used to generated synthetic data for a simulation study and a power study. These studies are needed to check the coverage and the performance of the derived prediction intervals. Finally, the method is applied to two case studies to validate the nitrate data of the river Yzer measured in 2003 and 2004. In a first case, two years of data are validated at one sampling location. In a second case, the data validation is applied to two years of data on all sampling locations of the river basin that contain enough data to fit the AM models.

3.3.1 Illustration of the methodology on a real data case

The methodology is illustrated on the data of sampling location S5 which was introduced in Section 1.2. The sampling location is located along the river Yzer and its particular location is highlighted in Figure 1.2. The dataset consists of 8 variables, (1) Day number throughout the year, (2) time, (3) temperature, (4) dissolved oxygen (DO), (5) nitrite (NO_2^-), (6) chemical oxygen demand (COD), (7) pH and (8) nitrate (NO_3^-). The observations of the following months were missing: July-September 1990, December 1991, December 1993, November 1994, January-December 1995, January 1997, November 1998, July 1999, December 1999 and September 2001. First, the additive model is built by using all available data before 01/01/2003 and the quality of the model is evaluated in a residual analysis. Then this AM is used to validate a new observation obtained at 14/01/2003 by using the different PI's.

3.3.1.1 Procedure to build the additive model

The nitrate concentration is modelled using an additive model. For the predictor functions of the model only local linear smoothers are used. Hence, the model is fully nonparametric. The first 7 variables are allowed to be included in the final model. In Chapter 1 it was shown that a considerable amount of seasonal variation was present in the data. A common approach to model this variation is to include sinusoidal functions of fixed periods to describe the seasonal cycle within a year (e.g. Hirst, 1998, Cai and Tiwari, 2000 McMullan et al., 2003 and McMullan, 2004). The day of year (support $[0, 365]$) is often used for this purpose. In Figure 1.6 two fits are shown. One by using sinusoidal functions and another by using a smoother to model the seasonal effect. Both approaches use the day of year as predictor. In this section we have chosen for a fully nonparametric approach and use a smoother to model the the seasonal effect. The BRUTO algorithm is used for model selection. The BRUTO algorithm starts with the null fit $\hat{\boldsymbol{y}} = \bar{\boldsymbol{y}}$, where $\bar{\boldsymbol{y}}$ is the $n \times 1$ vector $\bar{\boldsymbol{y}} = (\bar{y}, \ldots, \bar{y})^T$. During each iteration the GCV is optimised either by including a certain variable in the model, by adjusting its span or by removing the variable from the model. For each iteration the change of the GCV and the degrees of freedom of the model are given in Figure 3.7. The numbers in the plot indicate which of the predictors was adjusted in each cycle. During the first 4 cycles predictors 1, 6, 7 and 2 are included in the model. From the 5^{th} up to the 9^{th} cycle the spans of the selected predictors are adjusted. In cycle 10 and 11

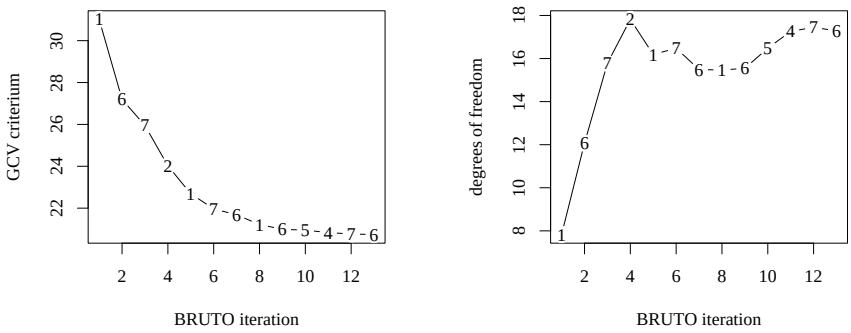

Figure 3.7: Left: Convergence of the GCV criterion when BRUTO is applied to the data of sampling location S5 along the river Yzer. Right: The evolution of the total degrees of freedom in the model in function of the iteration number. The numbers along the curve indicate which of the 7 predictors is updated

predictors 5 an 4 are selected. And the last cycles consist of adjusting the spans of predictors 7 and 6. The final model includes predictors 1,2,4,5,6 and 7. Notice that the temperature (3^{th} predictor) is never included in the model. At first, the GCV decrease is steep, which is due to the take up of extra predictors in the model. This is also reflected in the steep increase of the associated degrees of freedom.

The resulting model is presented in Figure 3.8. To enable a graphical representation of the high dimensional regression surface, we have chosen to represent the fit as a function of the temporal dimension (Figure 3.8 top). The effect of each of the predictors is shown in Figure 3.8 in the remaining panels. All fits are accompanied by 95% pointwise confidence intervals. A fitted value $\hat{y}_t$ is equal to the sum of the general mean α and each of the contributions of the corresponding predictor values $f_j(x_{jt})$. The figure shows a clear seasonal pattern with low contributions in summer and high contributions in winter, and an increasing contribution of the temporal trend (Time) until 1998 and decreasing trend from 1999 on. Low DO concentrations seem to have a negative contribution on the nitrate concentration, while high DO concentrations have a positive contribution. The contribution of COD is inversely related to the nitrate concentration and levels off at high COD concentrations. The contributions of DO and COD can be explained from the biochemical

processes that are taking place in the river. Low dissolved oxygen concentrations limit the nitrification process which converts ammonium to nitrate as it requires oxygen to be completed. Such low oxygen levels are typically occurring at high COD levels. Additionally, in anoxic conditions (in the absence of oxygen and the presence of nitrate), certain micro-organisms can use nitrate to replace oxygen as electron acceptor and in the presence of organic matter they convert nitrate to nitrogen gas which eventually escapes from the water phase. The contribution of nitrite seems to be approximately proportional to the actual nitrate concentration. In Figure 3.8 it can be seen that the model is sufficiently flexible to model a large part of the variation of the original data series.

Once the model is fitted, one can predict the mean response for a new observation by simply adding the individual effects for each of the predictor variables observed at time $n + 1$. In this way a new nitrate value can be calculated, given its day of year, time, DO, NO_2^-, COD and pH values measured for the particular sample under validation.

The model quality is checked in a residual analysis. Residual plots are constructed by plotting the residuals $\hat{\epsilon}_t$'s in function of each predictor. They are presented in Figure 3.9. From the residual plots the data seem more or less homoscedastic. The variance estimate of the residuals is $\hat{\sigma}^2_{S5} = 18.7$. Friedman's supersmoother (Friedman, 1984) is added to each residual plots to assist in visualising the residual pattern. They show that the mean of the residuals is centred around zero, except in data sparse regions at the endpoints. This is likely to be a boundary effect of the smoother. At the boundaries, the data are sparse and a few residuals can have a large influence on the fit of the smoother used in the residual plot. In Figure 3.10 the histogram and the QQ-plot of the residuals indicate deviations from normality in the upper tail and suggest that the residuals are distributed with a slight tail to the right. The boxplot also shows some outliers. When the outliers are removed, the residuals appear to be almost Gaussian (results not shown). Note that in our application these nitrate observations cannot be removed because they might be extreme events which are characteristic for the data-generating process. Moreover, the technique is based on the assumption that all historical data has been validated.

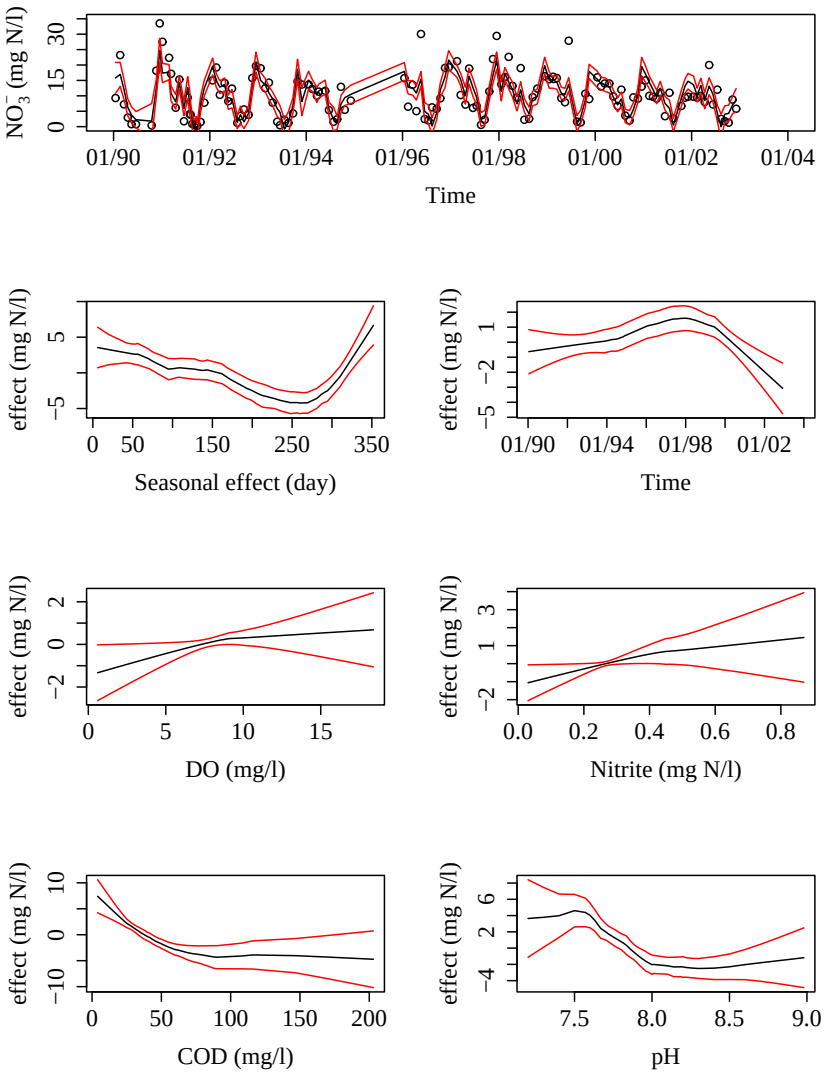

Figure 3.8: AM for nitrate at sampling location S5 at the river Yzer. Nitrate is modelled by a seasonal effect (day), long term trend (Time), DO, COD, nitrite and pH. The top panel shows the data and the lower panels show the contribution of each predictor

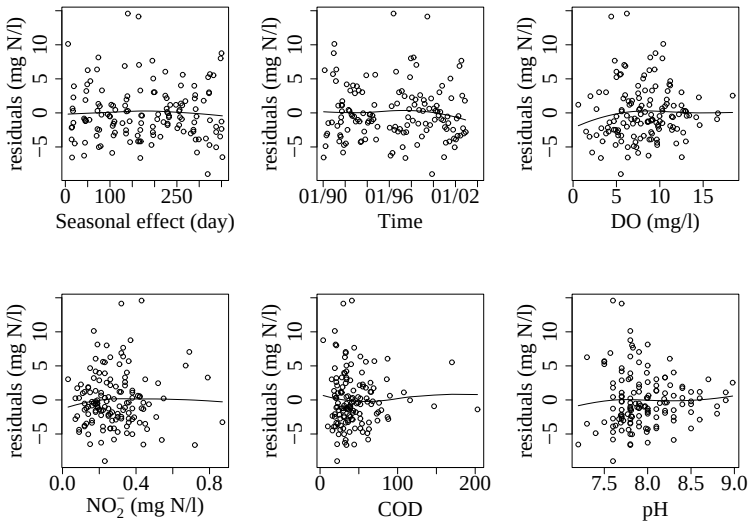

Figure 3.9: Residual plots for the additive model in Figure 3.8. Friedman's super-smoother is added to each plot to assess the residual pattern

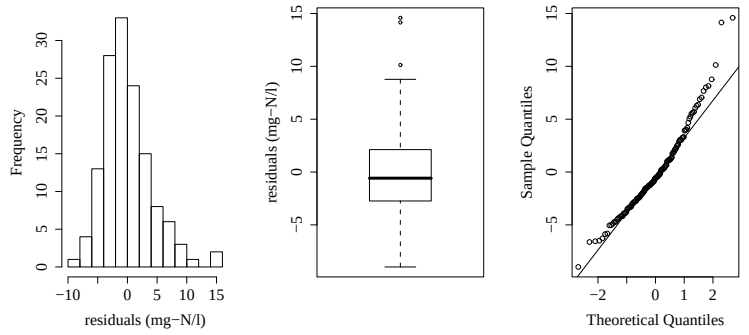

Figure 3.10: Histogram, boxplot and QQ-plot of the residuals from the additive model in Figure 3.8

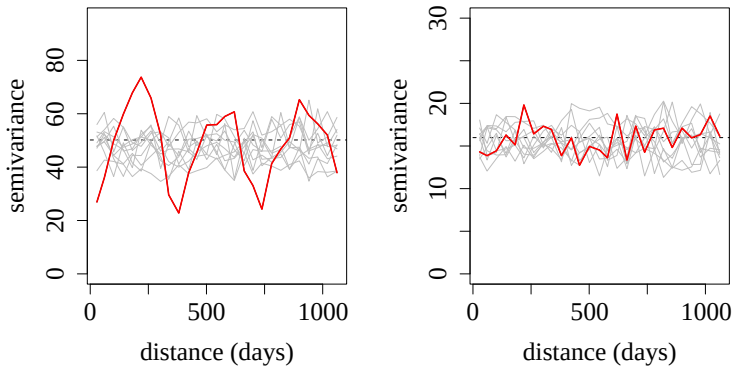

Figure 3.11: Variogram of the original nitrate series (left) and of the residuals after fitting the AM from Figure 3.8 (right). 10 variograms generated from white noise with the same variance are added to the plot (thin grey lines)

The presence of serial correlation in the residuals is checked using the runs test and by making a variogram of the residuals. The runs test is a nonparametric test that checks the randomness hypothesis of a data sequence (see, e.g., (McWilliams, 1990)). The run test on the residuals gives a p-value of 0.78, which clearly accepts the null hypothesis of randomness. A variogram is a tool to visualise autocorrelation in unequally spaced observations. To construct the variogram, first the differences $d(ij) = y_i - y_j$ and the time differences $\Delta_t(ij) = t_i - t_j$ are calculated for all observations i and j. According to their time difference $\Delta_t(ij)$, all differences $d(ij)$ are classified in time distance classes with mean time distance $\Delta_{t,k}$. The distance classes are taken to be equal in size and the bin-length is taken at 30 days. For each distance class k the semivariance is estimated as $\rho_k = \sum_{i=1}^{n_k} d_i^2/(2n_k)$. The semivariance ρ_k is then plotted against $\Delta_{t,k}$. The left panel of Figure 3.11 represents the variogram for the original data series and the right panel displays the variogram for the residuals of the AM. The grey lines in the background are variograms obtained when white noise was created with the same variance as the variograms of interest. The original nitrate measurements are clearly autocorrelated and the seasonal pattern is very obvious. After the AM was fitted, the autocorrelation is completely removed and the variogram behaves in a similar way as the ones originating from white noise.

The additive model for the historical data is fitted and the residuals are shown to be independent. The model can now be used to construct a prediction interval for new observations. In the next section, the validation is performed using the 3 different PI's described in Section 3.2.2.

3.3.1.2 Validation of a new observation by the use of prediction intervals

In the previous section an additive model was established using the data before 01/01/2003. The first new observation is acquired on 14/01/2003 and will be validated. The AM is used to perform a prediction of the fitted response $\hat{y}_{n+1,S5} = 12.3$. The estimated variance corresponding to this prediction is $\hat{\sigma}^2_{\hat{y}_n,S5} = 2.6$. The prediction interval for nitrate on 14/01/2003 is given in Figure 3.12. Instead of creating a two-sided interval, it makes more sense for nitrate to use a one-sided interval by concentrating all the uncertainty in the upper tail. Low nitrate concentrations are not harmful for the environment, so it is more interesting to focus on a faster detection of abnormal high nitrate concentrations. In the double bootstrap procedure 1000 bootstraps are calculated for each bootstrap loop (B_1 and B_2) resulting in 1 million bootstrap replicates (B_1B_2). In the left panel the historical data are presented together with the optimal fitted model. In the right panel, the new observation is represented by a dot and the upper limit of the bootstrap interval is indicated using the 3 different methods. The %bPI seems to be slightly higher than the aPI and the sbPI. The new observation lays in all intervals. Hence, the new observation is declared valid and can be added to the historical database.

In this study, B_1 and B_2 are chosen to be 1000, resulting in 1 million bootstrap replicates (B_1B_2). In the ideal case, however, the number of bootstrap replicates should be taken to be ∞. In practice this is not feasible and the number of bootstrap replicates is set at a large value. This leads to a bootstrap resampling variability. Thus, when the calculation of the bootstrap PI is repeated on the same data, the obtained PI will be slightly different. To stabilise the bootstrap resampling variability, the number of bootstrap replicates should be taken large enough. In a double bootstrap procedure, the bootstrap resampling variability is introduced in both loops. To control the bootstrap resampling variability due to the first loop, the size of B_1 should be appropriate. The bootstrap resampling variability caused by the second loop is controlled by B_1B_2. Hence stable intervals are obtained by taken B_1 and B_1B_2 large enough. The latter can be obtained by taking the number B_1 very large and by taking $B_2 = 1$ or by using moderate values for both B_1 and B_2. In a practical implementation, the computational complexity associated with both

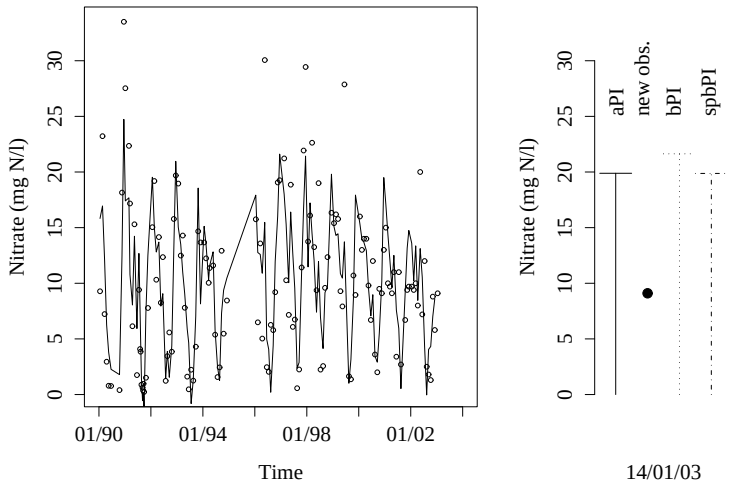

Figure 3.12: Prediction interval for the nitrate concentration on 14/01/2003. Left panel: historical data with model fit. Right panel: The new observation (dot) is accepted by the tree one-sided prediction intervals

bootstrap loops has to be taken into account. Here, the computational load of the second loop is negligible compared to the first loop. Hence, it is interesting to take B_1 as small as possible in order to reduce the computational power. The impact of the sizes of B_1 and B_2 is assessed in Figure 3.13. One sided intervals were calculated to validate nitrate measurements. For the same dataset 50 bootstrap intervals are calculated for (1) $B1 = 1000, B2 = 1$, (2) $B1 = 10000, B2 = 1$, (3) $B1 = 10000, B2 = 100$ and (4) $B1 = 1000, B2 = 1000$. For cases (1) and (4), the time needed to calculate the intervals was almost equal because the computational complexity associated with the calculation of 1000 AM's in the first bootstrap loop is much larger than the complexity needed for the second step. For case (2) and (3), however, 10 times more computational time was needed because the first loop was executed 10 times more. The figure clearly illustrates that for case (4) the one sided interval is estimated much more accurately than in case (1) where there is still a considerable amount bootstrap resampling variability. The stability of the intervals in (4) was slightly better than in case (2). This is because the second loop was only executed 10000 times for case (2) compared to 1000000 times for case (4). In case (3) a small gain in accuracy can be observed compared to case (4). In both cases the second loop is assessed 1000000 times. Hence the bootstrap

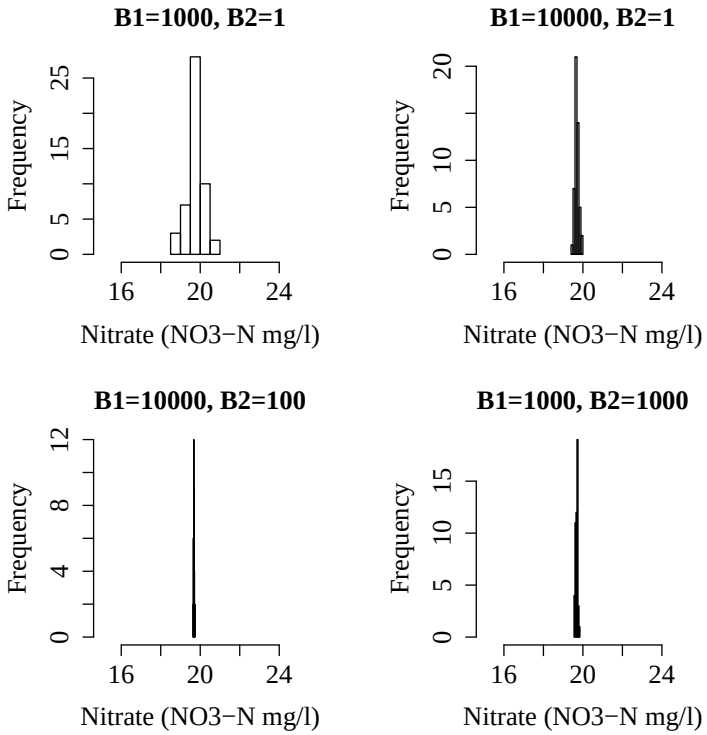

Figure 3.13: Effect of the number of bootstraps in the first and second loop on the bootstrap resampling variability of one-sided 95% sbPI. Each histogram displays the empirical distribution of the upper limit of the one-sided 95% sbPI and is based on 50 PI's. B1 is the number of bootstraps in the main bootstrap loop and B2 is the number of bootstraps in the second bootstrap loop

resampling variability induced by the second loop is controlled at the same level. In case (3) the first loop is executed 10 times as much as in case (4) and therefore a slight reduction of the bootstrap resampling variability is established. But this is at the expense of an increase in the computational time by a factor of 10. In order to reach an acceptable accuracy while keeping the computational time limited we decided to use $B_1 = 1000$ and $B_2 = 1000$.

3.3.2 Evaluation of the coverage of the PI's in a simulation study

In theory, 95% prediction intervals should contain (cover) 95% of the data if they follow the model. In a simulation study we can calculate the coverage empirically. A large number of simulated datasets have to be generated and for each dataset an observation at time $n+1$ should be validated. The data are simulated from a known mean model and a pre-specified distribution of errors. The empirical coverage is then calculated as the ratio between the number of simulations where the validated observation is accepted and the total number of simulations. In this study the empirical coverages of three different PI's derived in Section 3.2.2 are assessed. Five different types of distributions are used in this study, normal residuals, two types of residuals originating from right-tailed distributions and two types of residuals originating from left-tailed distributions. The results of the nitrate dataset at location S5 in Section 3.3.1 are used to generate the data for the simulation study. First we will explain how we obtain samples from right-tailed distributions. Weibull distributions with shape factors of 1 and 2 are used. The scale parameter can be chosen arbitrarily because the simulated residuals are standardised and multiplied with the standard deviation $\hat{\sigma}_{S5}$ of the residuals obtained from the fitted model in Figure 3.8. The residuals from the left-tailed distributions are generated by changing the sign of the residuals from the right-tailed distributions. Plots of the distribution functions that are used in the simulation study are given in Figure 3.14. For the normal residuals we sample from a normal distribution with mean 0 and variance $\hat{\sigma}^2_{S5}$.

Once we can generate new residuals with the same variance as the original data, simulated datasets are constructed. First n residuals are simulated from a particular distribution, and they are denoted by $\boldsymbol{\epsilon^*}$. The simulated datasets $\boldsymbol{D}^*$ then consist of the original predictors ($\boldsymbol{x_1}, \ldots, \boldsymbol{x_q}$) and the simulated response $\boldsymbol{y}^* = \hat{\boldsymbol{y}} + \boldsymbol{\epsilon^*}$. For the simulated datasets, the values of the true underlying function $m(x_{1t}, \ldots, x_{qt})$, $t = 1, \ldots, n$, and the observation under validation at time $t = n + 1$ are known. They are the $\hat{\boldsymbol{y}}_{S5}$ and $\hat{y}_{n+1,S5}$ represented in Figure 3.12, respectively.

For each distribution, 5000 datasets were constructed. Because the simulated y^*_{n+1} originate from a distribution with a mean of $\hat{y}_{n+1,S5}$, the empirical coverage of the 95% PI's should be close 95%. The empirical coverage for the different intervals are given in Table 3.1. The aPI's seem to be slightly too large for the Gaussian case. The coverage of the aPI's reduces when the data are right-tailed and increases when the data are left-tailed. This effect is even more apparent when the distribution becomes more asymmetric. The %bPI seems to have the tendency

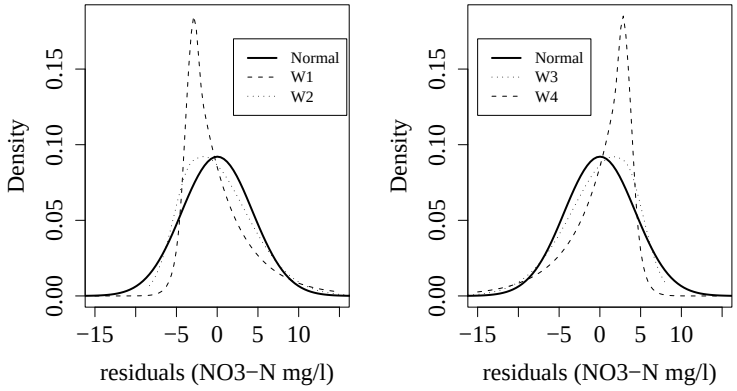

Figure 3.14: Density functions of the residuals used to generate the data for the coverage study

Table 3.1: Coverage (in %) of 95% PI's for data originating from different distributions

	Analytical	Bootstrap	
Distribution	aPI	%bPI	sbPI
Gaussian	96.4	97.2	95.0
Right-tailed, W1	94.1	96.0	94.5
Moderately right-tailed, W2	95.5	96.6	94.8
Moderately Left-tailed, W3	98.8	98.5	95.2
Left-tailed, W4	99.8	99.9	96.6

to be too large, the results for the different distributions are all above 95%. Only the sbPI seems to reach the correct coverage and is robust to deviations from normality. The coverage of %bPI is known to be problematic (Efron and Tibshirani, 1993; Davison and Hinkley, 1997). Corrections for percentile based intervals exist, for instance Efron and Tibshirani (1993) suggested bias and acceleration corrected intervals. But the implementation of the methods they suggested is not straightforward for our the double bootstrap procedure because the second loop consists of adding a random residual. For the semi-automatic data validation procedure, aPI's are preferred from a computational point of view. However, their coverage can be-

have poorly, particularly for the combination of upper bounded one-sided PI's and residuals that follow a left-tailed distribution. The coverage of studentised prediction error based bootstrap PI's (sbPI) however are rather robust to the distribution of the residuals and therefore we suggest to use this PI for data validation purposes.

For all 5 distributions, the coverage of the sbPI is close to the nominal value of 95% and in the data validation procedure, we will use this PI to validate a new observation. Under the null hypothesis H_0, a new observation is valid given the observed historical data when it lays in the PI. Under the alternative hypothesis H_1, the new observation is not valid. The decision error of concluding H_1 when in reality H_0 is true is called the type I error. It may also be called a false positive. When 95% PI's are used to validate the new observation, the corresponding probability $\alpha = 0.05$ is referred to as the type I error rate or the type I error level. Because the empirical coverage of the sbPI is close to the nominal value of 95%, it correctly controls for the type one error. Beside controlling the type one error, the power is another feature which is important in statistical testing. It is the probability to reject H_0 when H_1 is true. Hence, the higher the power, the higher the probability to detect a deviating observation. The power of the validation procedure is assessed in the next section.

3.3.3 Evaluation of the power

Again, the model fitted in Figure 3.8 is used to construct simulated datasets. The residuals, $\boldsymbol{\epsilon^*}$, are simulated from the normal distribution $N(0, \hat{\sigma}^2_{S5})$. The simulated datasets $\boldsymbol{D}^*$ consist of the original predictors ($\boldsymbol{x}_1, \ldots, \boldsymbol{x}_q$) and the simulated response $\boldsymbol{y}^* = \hat{\boldsymbol{y}}_{S5} + \boldsymbol{\epsilon^*}$. Thus for the simulated datasets, the values of the underlying mean function $m(x_{1t}, \ldots, x_{qt})$ evaluated at the predictor points $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_q)$ and $\boldsymbol{x}_{n+1}$ are $\hat{\boldsymbol{y}}_{S5}$ and $\hat{y}_{n+1,S5}$, respectively. Now a systematic deviation is introduced in the simulated data $(\boldsymbol{x}_{n+1}, y^*_{n+1})$ that will be validated. Instead of validating $y^*_{n+1} = \hat{y}_{n+1,S5} + \epsilon^*$, $y^*_n = \hat{y}_{n+1,S5} + \epsilon^* + l\hat{\sigma}_{S5}$ is used and the corresponding power to detect this deviation is calculated. To derive a complete power curve, different values for l are taken ($l \in [0, 4]$). For each value of l, 5000 datasets are generated to calculate the empirical power. The resulting power curve is displayed in Figure 3.15 (thick line). In the same figure a theoretical power curve is represented. The theoretical power was found under the assumption that the model uncertainty could be neglected. In this case, the model prediction $\hat{y}^*_n$ follows a normal distribution $N(\hat{y}_{n+1,S5}, \hat{\sigma}^2_{S5})$. The validated observation y^*_{n+1} however follows a normal distribution $N(\hat{y}_{n+1,S5} + l\hat{\sigma}_{S5}, \hat{\sigma}^2_{S5})$. Hence the power to detect the devia-

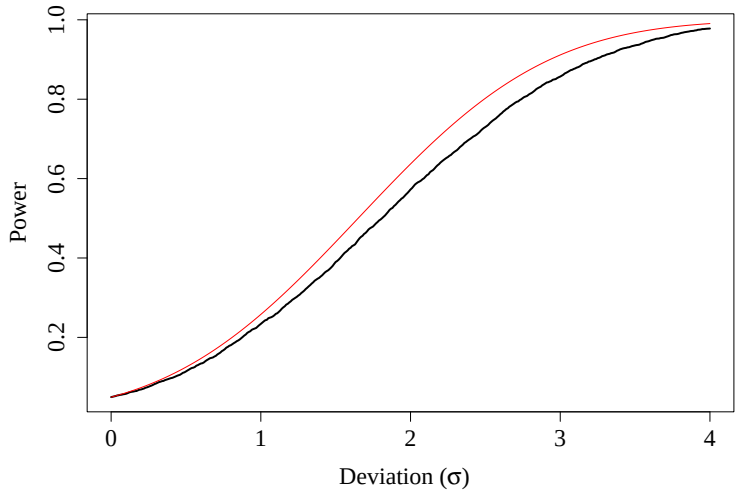

Figure 3.15: Power curve for the detection of deviations in validated data. (Black line: empirical power, thin grey line: theoretical power when the model uncertainty is neglected). The size of the deviations ranges between 0 and 4 times $\hat{\sigma}_{S5}$

tion in y^*_{n+1} is established by using the distribution function $N(\hat{y}_{n+1} + l\hat{\sigma}_{S5}, \hat{\sigma}^2_{S5})$ to calculate the probability $P(y^*_{n+1} > \hat{y}_{n+1,S5} + z_{1-\alpha}\hat{\sigma}_{S5})$. This theoretical power cannot be exceeded because model uncertainty is always present in practical applications. At the beginning, when $l = 0$ both curves start at 5%. This is due to the use of the 95% PI's which correctly control the type I error at the 5% level. For moderate values of l, the empirical power curve is lower than the theoretical one, but the empirical power remains remarkable high. This suggests that our method is well suited for data validation purposes.

3.3.4 Case study I: Validation at one sampling location

The data of sampling location S5 for the years 2003 and 2004 are validated. The dataset at this location contains 8 variables: day number to model the seasonal effect, date to model the long term trend, temperature (T), dissolved oxygen concentration (DO), nitrite concentration (NO_2^-), chemical oxygen demand (COD),

pH and nitrate concentration (NO_3^-). The time series starts at April 1990 and ends in December 2004. All 8 variables are measured on a monthly basis. The data from 1990 until December 2002 are considered as historical data. The nitrate data from 2003 and 2004 are validated in chronological order. In particular if a new observation lays within the 95% PI, then the measurement is accepted and considered as historical data for the validation of the next observation.

The results of the data validation are presented in Figure 3.16. All data from 2003 are accepted. The observations in January and February of 2004 are rejected. To assist the expert with the interpretation of the rejected observations, diagnostic plots can be generated. First reduced models are created by omitting each of the predictors one by one from the fitted model. The diagnostic plots consist of the representation of new PI's that were obtained with the reduced models. If the observation is accepted by the PI constructed with the reduced models, it indicates that there might be something wrong with the relationship between the validated observation and the omitted predictor. Diagnostic plots for the rejected observations are given in Figure 3.17 and 3.18, respectively. In the x-axis, the omitted variable is indicated.

From these diagnostic plots possible explanations for the rejection of the data may become clear. The measurement in January is only accepted when the trend (Time) is omitted from the model, giving a strong indication that this measurement does not follow the expected long-term trend in the data. The measurement in February is accepted when the trend or pH are omitted from the model. This indicates again that a potential cause of the deviation is related to the trend. The nitrate concentrations in the beginning of 2004 are known to be unexpectedly high (Anonymous, 2005). The river Yzer is located in the countryside and 2003 was dry year, which resulted in an accumulation of nitrate in the soil in summer and autumn. The dry summer of 2003 had a beneficial effect on the nitrate concentration, since there was a limited amount of nitrate washed to the water course by rain. Hence, the nitrate accumulated in the soil and was washed out in the winter period. Moreover, January 2004 is recognised to be extremely wet by the Belgian Royal Meteorological Institute (KMI). It can be concluded from their data that this phenomenon at most happens once in a 100 years. The dry summer combined with an extreme wet winter provoked high nitrate concentrations in the receiving water.

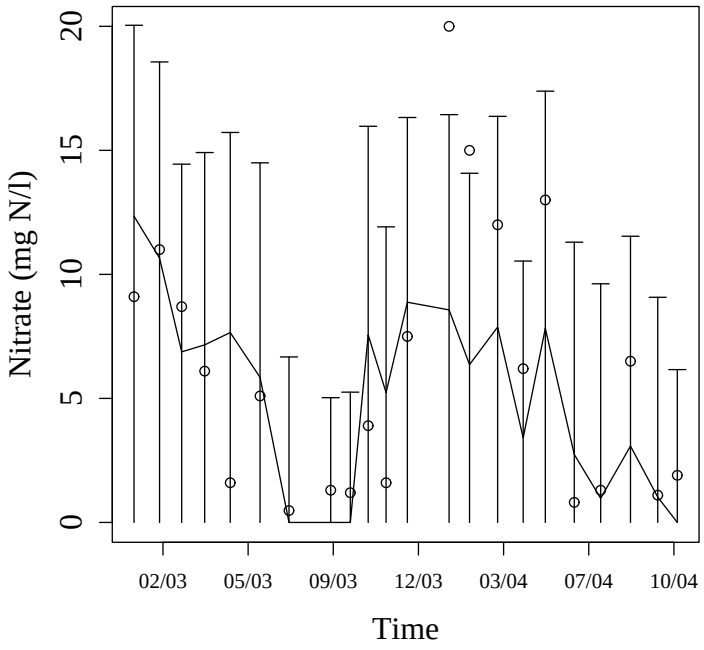

Figure 3.16: Validation of nitrate at sampling location S5 of the Yzer monitoring network. Nitrate concentrations in January and February 2004 are considered as anomalous by the automatic validation procedure. The dots represent the actual measurements, the solid line the predictions by the additive model and the horizontal bars are the 95% PI's

3.3.5 Case study II: Validation of an entire basin

The data from 2003 and 2004 are validated for all sampling locations of the Yzer basin, containing enough data to fit the models. The dataset at each location has information on 8 variables: Day number, date, T, DO, NO_2^-, COD, pH and NO_3^-. Again, all 8 variables are measured on a monthly basis. The data from 1990 until December 2002 are considered as historical data. The nitrate data from 2003 and 2004 are validated in chronological order. If a new observation lays within the PI, then the measurement is accepted and considered as historical data for the validation of the next observation. The data validation is carried out using 95% sbPI's.

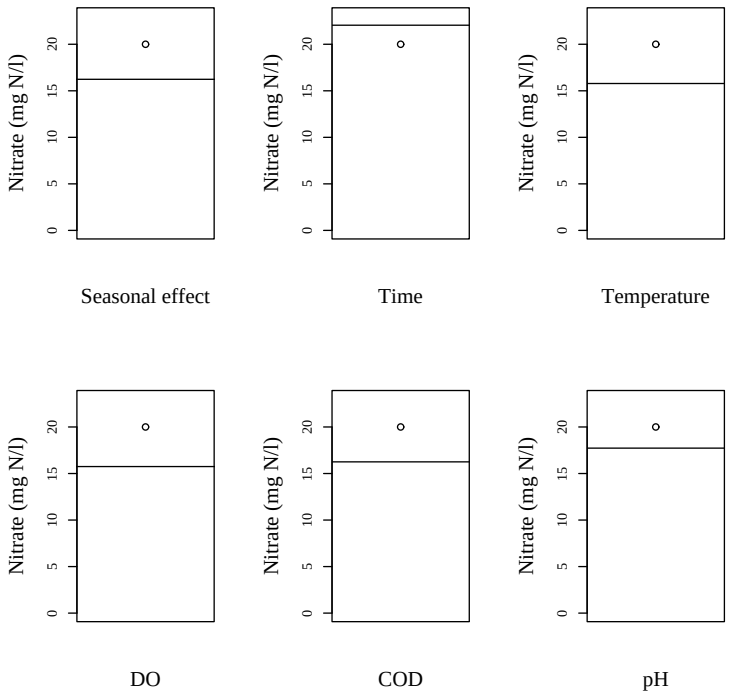

Figure 3.17: Diagnostic plots for rejected nitrate concentration of January 2004 at sampling location S5 of the Yzer monitoring network. The dot represents the observation and the black line indicates the location of the upper limit of the 95% interval

The empirical coverage of the intervals in a certain period is calculated by dividing the number of accepted observations in this period by the total number of validated observations in this period. The coverage of the intervals for the whole validation period, is 91%. However, the coverages for the 2003 data is 94.7% and is close to what is expected from theory when no deviations are present. In 2004 the coverage is only 80 % indicating the presence of a considerable number of anomalous data. In Figure 3.19 the results of the data validation based on the sbPI's is presented. The top panel shows the results of the validation in 2003, in the middle panel the results of 2004 are given and the bottom panel shows the evolution of the coverage of the sbPI's during the whole validation period. Accepted data are indicated with open dots and the rejected data are presented by black dots. From the middle panel

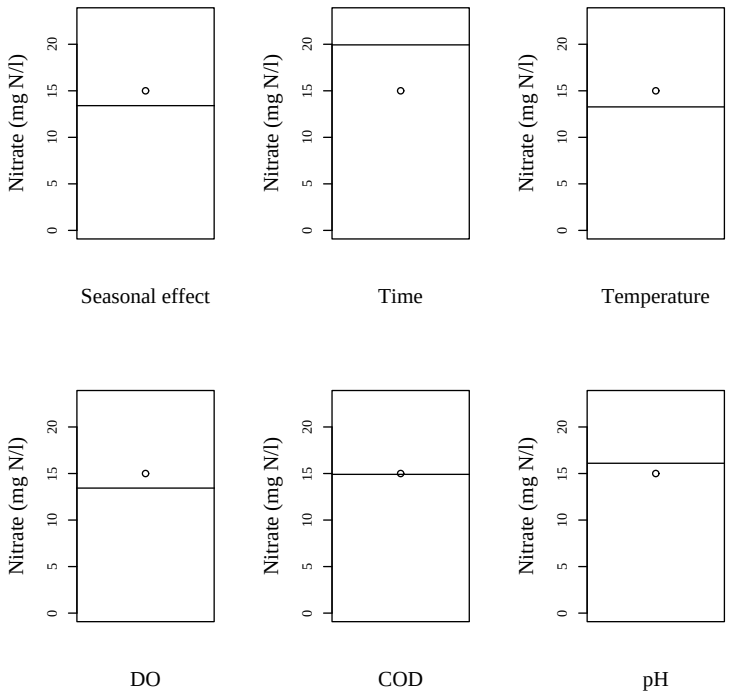

Figure 3.18: Diagnostic plots for rejected nitrate concentration of February 2004 at sampling location S5 of the Yzer monitoring network. The dot indicates the observation and the black line indicates the location of the upper limit of the 95% interval

of Figure 3.19, it can be concluded that a lot of the data in the period of January up to March 2004 are rejected. This is even more obvious in the results presented in the bottom panel. The bottom panel shows the evolution of the empirical coverage in each month. In 2003 the coverage is more or less stable at 95%. In the beginning of 2004 a clear drop of the coverages of the PI's is observed (January 56%, February 66% and March 67%) indicating that there was a change in the system during the first months of 2004.

A more general feature can be derived from Figure 3.19: similar to multivariate techniques, our method can also detect observations to be suspicious even if they are laying in the centre of the univariate distribution of the nitrate concentrations.

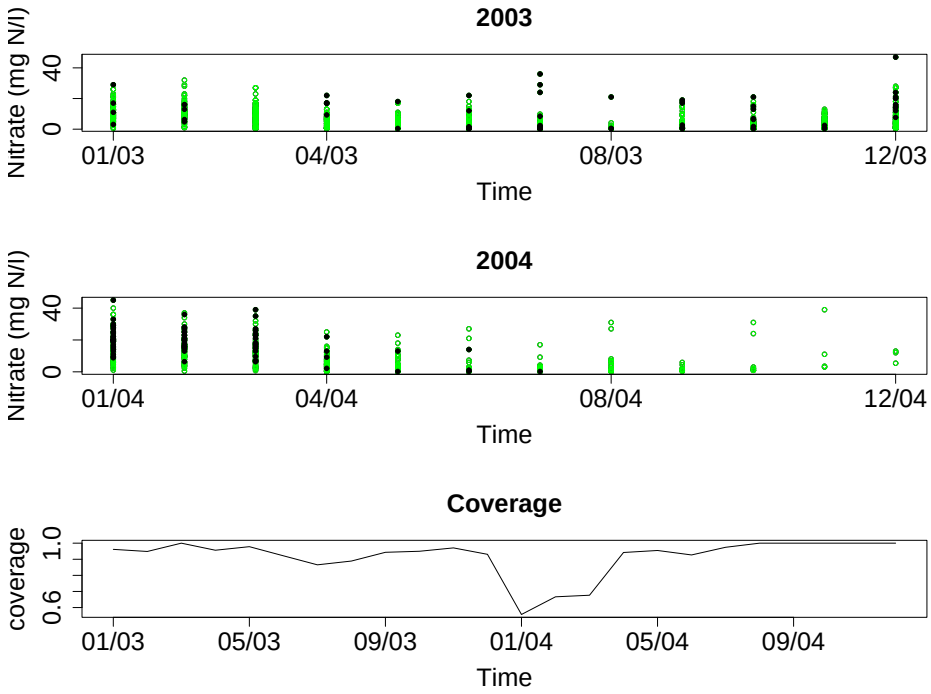

Figure 3.19: Validation of nitrate at at all sampling locations of Yzer monitoring network. The top panel: validation in 2003, middle panel: results for 2004 and the bottom panel: evolution of the coverage of the PI's during the whole validation period. Accepted data are indicated with an open dot and the rejected data are indicated with a grey dot

Hence, our methodology combines the interesting features of multivariate outlier detection without imposing restricted assumptions on the relationship between the response and the predictor variables.

3.4 Conclusions

A method for the validation of river water quality data is proposed. Based on the historical data an additive model is fitted, which is subsequently used to construct prediction intervals for future observations.

Our study indicates that the additive models are clearly able to catch the cyclic pattern present in the data and could model the nonlinear behaviour and relationships typically associated with river water quality data. As an interesting feature, the observed associations between the response and the predictors reflect well known physical and biochemical relationships. Since the model selection is carried out at each time step, the models succeed to adapt to changes in the processes of the underlying river.

From the different prediction intervals which are derived, the studentised prediction error based bootstrap PI's (sbPI's) are most interesting to be used in practice. The coverages of the 95% sbPI's have been assessed in a simulation study and in comparison with analytical intervals, which assume the residuals to be Gaussian, they appear to be much more robust against deviations from normality. The power of the method was also shown to be adequate.

The case studies have illustrated that our method could detect anomalous events, such as an abnormal high nitrate release due to a dry summer which was followed with an extreme wet winter period. The diagnostic plots are also useful to assist the operator with the analysis of the rejected observations: here they indicate that the rejection is related to the trend. In the case studies, the semi-automatic procedure detects suspicious observations laying at the edges as well as observations located in the centre of the univariate distribution of the nitrate observations. Hence, it combines the interesting features of classical multivariate outlier detection tools without having to impose linear relationships typically associated with these methods.

An ICT-tool based on this methodology could be of great value to analyse and maintain environmental databases originating from monitoring networks such as the ones which are implied by the WFD. Such a tool can be used to check the quality of the data and it can also detect abnormal changes in the water quality.

Part II

Spatio-temporal modelling of river monitoring networks

A selection of the presented work is published in

Clement, L., Thas, O., Vanrolleghem, P.A. and Ottoy, J.P. (2006). Spatio-temporal statistical models for river monitoring networks. *Water Science & Technology* 53(1), 9-15.

Clement, L. and Thas, O. (2007). Estimating and modelling spatio-temporal correlation structures for river monitoring networks. *Journal of Agricultural, Biological, and Environmental Statistics*, 12(2), 161-176.

Chapter 4

An introduction to state-space models

4.1 Introduction

Current environmental legislation has triggered the establishment of monitoring networks to assess environmental quality. Environmental processes typically show variability over space and time. Hence, environmental monitoring networks generate vast amounts of spatio-temporal data. In general these data show a rather complicated dependence structure and cannot be treated as a set of independent and identically distributed (i.i.d.) observations. Standard statistical data analysis techniques relying on this i.i.d. assumption are thus not valid. A correct analysis should take the spatio-temporal correlation into account.

In this dissertation, we aim to infer on the data at the sampling locations of river monitoring network and we do not aim to perform predictions at intermediate locations that are not sampled. Therefore the observations of the monitoring network at a certain time instant can be considered as the realisation of a finite-dimensional multivariate random variable with each dimension corresponding to each of the sampling locations. The state-space model framework, is particularly well suited to handle multivariate dynamic data. It can be used to treat a wide range of problems in time series analysis. A nice feature of state-space time series models is that the observations are considered to consist of several distinct components such as a trend, seasonal effect, regression elements and disturbance terms which are all modelled separately. The models for these components are then combined in a single model, the *state-space model* which forms the basis of the analysis (Durbin and Koopman, 2001). State-space modelling assumes that the underlying process is driven by a unobserved series of $m \times 1$ vectors $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_n$, the states, that are associated with a series of $p \times 1$ observed vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$. The states are assumed to follow a stochastic transitional model. Generally, the state-space analysis aims to infer on the properties of the states $\boldsymbol{S}_t$ by the knowledge of the set of observations $\boldsymbol{Y}_s = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s)^T$. The estimation of $\boldsymbol{S}_t$ given $\boldsymbol{Y}_s$ is referred to as

1. filtering for $t = s$,

2. smoothing for $t < s$ and

3. prediction for $t > s$.

When all stochastic processes are Gaussian, the *Kalman filter* can be used to address the filtering and the prediction problem and the *Kalman smoother* solves the smoothing problem. Another interesting feature of the Kalman filter is that it can be used as a computational efficient algorithm to factorise the likelihood of the model.

Before we look to Kalman filtering and smoothing into more detail, we first introduce the state-space representation of the model.

4.2 State-space model

In this section, we assume that no predictor variables are available. Exogenous variables will be introduced later on in Section 4.3.4. The state-space form assumes that a p-dimensional multivariate process $\boldsymbol{y}_t = (y_{1t}, \ldots, y_{pt})^T$, is driven by a m-dimensional state process $\boldsymbol{S_t} = (S_{1t}, \ldots, S_{mt})^T$. This state process is believed to be generated by a first-order Markovian process,

$$\boldsymbol{S}_t = \boldsymbol{\Phi}_t \boldsymbol{S}_{t-1} + \boldsymbol{\delta}_t, \tag{4.1}$$

with $t = 1, \ldots, n$, an $m \times m$ transition matrix $\boldsymbol{\Phi}_t$ and independent $m \times 1$ vectors $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_n$ with zero mean and $m \times m$ variance-covariance matrices $\boldsymbol{Q}_t$. The state process however cannot be observed, instead we only observe a noisy linear transformed version of it, $\boldsymbol{y}_t$. The observations $\boldsymbol{y}_t$ are related to the state variable $\boldsymbol{S}_t$ via the measurement equation

$$\boldsymbol{y}_t = \boldsymbol{Z}_t \boldsymbol{S}_t + \boldsymbol{\epsilon}_t, \tag{4.2}$$

where $\boldsymbol{Z}_t$ is a $p \times m$ matrix and the $\boldsymbol{\epsilon}_t$ $(t = 1, \ldots, n)$ are independent $p \times 1$ vectors with zero mean and $p \times p$ covariance matrix $\boldsymbol{H}_t$. The matrices $\boldsymbol{\Phi}_t$, $\boldsymbol{Q}_t$, $\boldsymbol{Z}_t$ and $\boldsymbol{H}_t$ are also referred to as the system matrices. They are assumed to be non-stochastic and to change over time in a predetermined way. The resulting system is thus linear. When the system matrices do not change over time, the resulting model is time-invariant. For the model to be completely specified, the distribution of the initial state, $\boldsymbol{S}_0$, has to be specified. $\boldsymbol{S}_0$ is assumed to be Gaussian with mean $\boldsymbol{\lambda}_{0|0}$ and covariance matrix $\boldsymbol{P}_{0|0}$. Further the $\boldsymbol{\delta}_t$ and $\boldsymbol{\epsilon}_t$ are assumed to be uncorrelated with each other in all time periods and with the initial state. Hence,

$$\mathrm{E}\left(\boldsymbol{\delta}_s \boldsymbol{\epsilon}_t^T\right) = \boldsymbol{0} \quad \text{for all } s, t = 1, \ldots, n \tag{4.3}$$

and

$$\mathrm{E}\left(\boldsymbol{S}_0 \boldsymbol{\delta}_t^T\right) = \boldsymbol{0}, \quad \mathrm{E}\left(\boldsymbol{S}_0 \boldsymbol{\epsilon}_t^T\right) = \boldsymbol{0} \quad \text{for all } s, t = 1, \ldots, n. \tag{4.4}$$

4.3 Kalman filter and smoother

Most of this section is based on Harvey (1989) and Durbin and Koopman (2001). Once a model is written in its state-space formulation, the Kalman filter and smoother can be applied to find the optimal estimator of the state process $\boldsymbol{S}_t$ at time t.

The *Kalman Filter* provides an estimator of $\boldsymbol{S}_t$ given all observations $\boldsymbol{y}_k$ which are observed at the time steps $k = 1, \ldots, t$. Let $\boldsymbol{Y}_t$ be $\boldsymbol{Y}_t = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_t^T)^T$, the filtered estimator of $\boldsymbol{S}_t$ is then the conditional expectation $\mathrm{E}\,(\boldsymbol{S}_t|\boldsymbol{Y}_t)$. The Kalman filter is important for e.g. online estimation and prediction because it continuously updates our knowledge of the system each time a new observation y_t is brought in. Another interesting feature of the Kalman filter is that it provides a convenient way for calculating the likelihood when the initial state vector and the disturbances are Gaussian.

In offline applications, it is more appropriate to estimate the state vector at a certain time t conditional on all the information which is available, $\boldsymbol{Y}_N = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$. Hence, to provide the optimal predictor of the state process at time t, the measurements that are obtained on later time instants $t + 1, \ldots, n$ should also be considered. In this setting $\mathrm{E}\,(\boldsymbol{S}_t|\boldsymbol{Y}_N)$ is the appropriate estimator and is provided by the *Kalman smoother*.

4.3.1 General form of the Kalman filter

Suppose a system is defined by Equations (4.1) and (4.2), and suppose that all distributions are normal. Let the set $\boldsymbol{Y}_{t-1}$ be the vector of the past observations $\boldsymbol{Y}_{t-1} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_{t-1}^T)^T$. Then the conditional distribution of $\boldsymbol{S}_t$ given $\boldsymbol{Y}_{t-1}$ is also normal, $N(\boldsymbol{\lambda}_{t|t-1}, \boldsymbol{P}_{t|t-1})$, where $\boldsymbol{\lambda}_{t|t-1} = \mathrm{E}\,(\boldsymbol{S}_t|\boldsymbol{Y}_{t-1})$ and $\boldsymbol{P}_{t|t-1} = \mathrm{E}\left((\boldsymbol{S}_t - \boldsymbol{\lambda}_{t|t-1})(\boldsymbol{S}_t - \boldsymbol{\lambda}_{t|t-1})^T\right)$. They can be immediately determined from Equation (4.1),

$$\boldsymbol{\lambda}_{t|t-1} = \boldsymbol{\Phi}_t \boldsymbol{\lambda}_{t-1|t-1}$$

and

$$\boldsymbol{P}_{t|t-1} = \boldsymbol{\Phi}_t \boldsymbol{P}_{t-1|t-1} \boldsymbol{\Phi}_t^T + \boldsymbol{Q}_t.$$

When a new observation becomes available it is our aim to incorporate $\boldsymbol{y}_t$ in the estimation of $\boldsymbol{S}_t$. In this case the conditional distribution of $\boldsymbol{S}_t$ given $\boldsymbol{Y}_t$ has to be defined. We will denote this particular distribution as $N(\boldsymbol{\lambda}_{t|t}, \boldsymbol{P}_{t|t})$. Here $\boldsymbol{\lambda}_{t|t} = \mathrm{E}\,(\boldsymbol{S}_t|\boldsymbol{Y}_t)$ and $\boldsymbol{P}_{t|t} = \mathrm{E}\left((\boldsymbol{S}_t - \boldsymbol{\lambda}_{t|t})(\boldsymbol{S}_t - \boldsymbol{\lambda}_{t|t})^T\right)$ have to be determined.

We first introduce the innovations that are defined as the $p \times 1$ vector $\boldsymbol{v}_t = \boldsymbol{y}_t - \boldsymbol{Z}_t \boldsymbol{\lambda}_{t|t-1}$. They can be regarded as the prediction error. The innovations are normally distributed with a zero mean and variance-covariance matrix $\boldsymbol{F}_t$.

Since all variables are Gaussian, standard formulae from multivariate normal regression theory can be used to obtain $\boldsymbol{\lambda}_{t|t}$ and $\boldsymbol{P}_{t|t}$. Eubank (2006) has shown that these estimators are basically the best linear unbiased predictor (BLUP) of the state $\boldsymbol{S}_t$ based on the innovations.

$$\begin{aligned}\boldsymbol{\lambda}_{t|t} &= \boldsymbol{\lambda}_{t|t-1} + \operatorname{cov}\left(\boldsymbol{S}_t, \boldsymbol{v}_t^T\right) \operatorname{var}\left(\boldsymbol{v}_t\right)^{-1} \boldsymbol{v}_t \\ &= \boldsymbol{\lambda}_{t|t-1} + \boldsymbol{P}_{t|t-1}\boldsymbol{Z}_t^T \boldsymbol{F}_t^{-1}\boldsymbol{v}_t,\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{P}_{t|t} &= \boldsymbol{P}_{t|t-1} - \operatorname{cov}\left(\boldsymbol{S}_t, \boldsymbol{v}_t\right) \operatorname{var}\left(\boldsymbol{v}_t\right)^{-1} \operatorname{cov}\left(\boldsymbol{S}_t, \boldsymbol{v}_t\right)^T \\ &= \boldsymbol{P}_{t|t-1} - \boldsymbol{P}_{t|t-1}\boldsymbol{Z}_t^T \boldsymbol{F}_t^{-1}\boldsymbol{Z}_t\boldsymbol{P}_{t|t-1},\end{aligned}$$

where

$$\operatorname{cov}\left(\boldsymbol{S}_t, \boldsymbol{v}_t\right) = \boldsymbol{P}_{t|t-1}\boldsymbol{Z}_t^T,$$

and

$$\begin{aligned}\operatorname{var}\left(\boldsymbol{v}_t\right) = \boldsymbol{F}_t &= \boldsymbol{Z}_t \operatorname{var}\left(\boldsymbol{S}_t|\boldsymbol{Y}_{t-1}\right) \boldsymbol{Z}_t^T + \operatorname{var}\left(\boldsymbol{\epsilon}_t\right) \\ &= \boldsymbol{Z}_t\boldsymbol{P}_{t|t-1}\boldsymbol{Z}_t^T + \boldsymbol{H}_t.\end{aligned}$$

The recursions for $\boldsymbol{\lambda}_{t|t-1}$, $\boldsymbol{P}_{t|t-1}$, $\boldsymbol{\lambda}_{t|t}$, $\boldsymbol{F}_t$ and $\boldsymbol{P}_{t|t}$ form the hart of the Kalman filter.

Suppose $\boldsymbol{\lambda}_{0|0}$ and $\boldsymbol{P}_{0|0}$ are known, then the Kalman filter can be summarised as follows: for $t = 1, \ldots, n$ the following forward recursions are used and they are started with time $t = 1$,

Prediction step

$$\boldsymbol{\lambda}_{t|t-1} = \boldsymbol{\Phi}_t\boldsymbol{\lambda}_{t-1|t-1} \tag{4.5}$$

$$\boldsymbol{P}_{t|t-1} = \boldsymbol{\Phi}_t\boldsymbol{P}_{t-1|t-1}\boldsymbol{\Phi}_t^T + \boldsymbol{Q}_t. \tag{4.6}$$

Update step

$$\boldsymbol{\lambda}_{t|t} = \boldsymbol{\lambda}_{t|t-1} + \boldsymbol{P}_{t|t-1}\boldsymbol{Z}_t^T\boldsymbol{F}_t^{-1}(\boldsymbol{y}_t - \boldsymbol{Z}_t\boldsymbol{\lambda}_{t|t-1}) \tag{4.7}$$

$$\boldsymbol{F}_t = \boldsymbol{Z}_t\boldsymbol{P}_{t|t-1}\boldsymbol{Z}_t^T + \boldsymbol{H}_t. \tag{4.8}$$

$$\boldsymbol{P}_{t|t} = \boldsymbol{P}_{t|t-1} - \boldsymbol{P}_{t|t-1}\boldsymbol{Z}_t^T\boldsymbol{F}_t^{-1}\boldsymbol{Z}_t\boldsymbol{P}_{t|t-1}. \tag{4.9}$$

4.3.2 Likelihood and the predictor error decomposition

In a classical setting, the observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are i.d.d. This enables the log-likelihood function to be written as a sum,

$$\log L_{\boldsymbol{Y_N}}(\boldsymbol{\Psi}) = \sum_{t=1}^{n} p(\boldsymbol{y_t}), \tag{4.10}$$

where $p(\boldsymbol{y_t})$ is the joint density function evaluated in $\boldsymbol{y}_t$ and indexed by the parameter vector $\boldsymbol{\Psi}$ (dependence is suppressed for notational comfort). The maximum likelihood estimator is then found by maximising Equation (4.10) with respect to the parameter $\boldsymbol{\Psi}$.

When the observations are dependent, the decomposition (4.10) is not applicable. Fortunately, the state-space representation allows a factorisation of the likelihood by using conditional density functions, resulting in a convenient decomposition of the log-likelihood

$$\log L_{\boldsymbol{Y}_N}(\boldsymbol{\Psi}) = \sum_{t=1}^{n} \log p(\boldsymbol{y}_t | \boldsymbol{Y}_{t-1}) \tag{4.11}$$

where $p(\boldsymbol{y}_t|\boldsymbol{Y}_{t-1})$ is the conditional density function of $\boldsymbol{y}_t$ given all previous observations $\boldsymbol{Y}_{t-1} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_{t-1}^T)^T$. When the disturbances and initial state vector $\boldsymbol{S}_0$ are normally distributed, those conditional density functions are explicitly known and also Gaussian. From the Kalman recursion in Equations (4.5)-(4.9) it can be seen that the expected value of $\boldsymbol{S}_t$ conditional on $\boldsymbol{Y}_{t-1}$ is normally distributed with mean $\boldsymbol{\lambda}_{t|t-1}$ and covariance matrix $\boldsymbol{P}_{t|t-1}$. Therefore, from Equation (4.2) it follows that the conditional distribution of $\boldsymbol{y}_t$ is normal with conditional mean

$$\mathrm{E}\left(\boldsymbol{y}_t | \boldsymbol{Y}_{t-1}\right) = \boldsymbol{Z}_t \boldsymbol{\lambda}_{t|t-1} \tag{4.12}$$

and covariance matrix $\boldsymbol{F}_t$. Thus, for a Gaussian model, the Kalman filter can be exploited to formulate the log-likelihood immediately as

$$\log L_{\boldsymbol{Y}_N}(\boldsymbol{\Psi}) = -\frac{pn}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{n}\log|\boldsymbol{F}_t| - \frac{1}{2}\sum_{t=1}^{n}\boldsymbol{v}_t^T \boldsymbol{F}_t^{-1}\boldsymbol{v}_t, \tag{4.13}$$

where the innovations $\boldsymbol{v}_t$ can be interpreted as the vector of the prediction errors. Equation (4.13) is therefore referred to as the prediction error decomposition of the log-likelihood.

When appropriate prior information on $\boldsymbol{S}_0$ is available, the prediction error decomposition of the log-likelihood will yield the exact log-likelihood of all observations, $\boldsymbol{Y}_N$. In most cases, however, genuine prior information is not available. Several possibilities that are commonly used for initialisation are introduced in the next section.

4.3.3 Kalman filter initialisation conditions

In a Bayesian framework and in the absence of genuine prior information, the Kalman filter is often initialised by the use of a diffuse prior. The state variable $\boldsymbol{S}_0$ is then assumed to be Gaussian distributed, say $\boldsymbol{S}_0 \sim MVN(\boldsymbol{0}, \kappa \boldsymbol{I})$, where κ is a positive scalar, $\boldsymbol{I}$ is an $m \times m$ identity matrix. The diffuseness is obtained when κ becomes large. However, κ is not allowed to grow unboundedly because then $\boldsymbol{P}_{0|0}^{-1}$ no longer exists and the distribution does no longer integrate to one. For most practical cases κ is set at an arbitrary large finite value. A large κ makes the variances (diagonal elements of $\boldsymbol{P}_{0|0}$) large, and so it limits the amount of information contained in $\boldsymbol{S}_0$.

Another common approach is to look to $\boldsymbol{\lambda}_{0|0}$ and $\boldsymbol{P}_{0|0}$ as parameters that have to be estimated. In this case, m parameters have to be estimated for $\boldsymbol{\lambda}_{0|0}$ and $m(m+1)/2$ parameter estimates are needed for $\boldsymbol{P}_{0|0}$. To restrict the number of parameters related to the initial conditions, $\boldsymbol{S}_0$ is often considered to be fixed. This is established by setting $\boldsymbol{S}_0 = \boldsymbol{\lambda}_{0|0}$ and $\boldsymbol{P}_{0|0} = \boldsymbol{0}$. In the next section we will show that in this case the state-space model can be reformulated as a state-space model with exogenous predictors and initial conditions $\boldsymbol{\lambda}_{0|0} = \boldsymbol{0}$ and $\boldsymbol{P}_{0|0} = \boldsymbol{0}$. The initial parameters can here be estimated by generalised least squares. Recently, Eubank (2006) showed that this approach, as well as the diffuse prior approach are closely related. Although they start from a completely different viewpoint, he showed they ultimately provide the same predictions.

Up to now, we did not take exogenous predictors into account. In many applications, however, the observations are modelled by taking such predictors into account. In this chapter we assume that the relation between the mean response and its predictors is linear. Due to the dependence structure in the data, the parameters related to the predictors are more efficiently estimated by the use of generalised least squares (GLS). The next section illustrates how the Kalman filter can be used for this purpose.

4.3.4 Using the Kalman filter to perform generalised least squares

Suppose q exogenous predictors at time t, say $\boldsymbol{X}_t = [\boldsymbol{x}_{1t}, \ldots, \boldsymbol{x}_{qt}]$, are used to predict $\boldsymbol{y}_t$, and that the state-space model (4.1)-(4.2) is reformulated as

$$\boldsymbol{S}_t = \boldsymbol{\Phi}_t \boldsymbol{S}_{t-1} + \boldsymbol{\delta}_t, \tag{4.14}$$

$$\boldsymbol{y}_t = \boldsymbol{Z}_t \boldsymbol{S}_t + \boldsymbol{X}_t \boldsymbol{\beta} + \boldsymbol{\epsilon}_t. \tag{4.15}$$

In this case the state-space model can be further reformulated as a regression model

$$\boldsymbol{y}_t = \boldsymbol{X}_t \boldsymbol{\beta} + \boldsymbol{u}_t, \tag{4.16}$$

with correlated error terms $\boldsymbol{u}_t = \boldsymbol{Z}_t \boldsymbol{S}_t + \boldsymbol{\epsilon}_t$. Writing it in matrix notation we find

$$\boldsymbol{Y}_N = \boldsymbol{X}_N \boldsymbol{\beta} + \boldsymbol{U}_N, \tag{4.17}$$

with $\boldsymbol{Y}_N = [\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T]^T$, $\boldsymbol{X}_N = [\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_n^T]^T$ and $\boldsymbol{U}_N = [\boldsymbol{U}_1^T, \ldots, \boldsymbol{U}_n^T]^T$, and let the covariance matrix of $\boldsymbol{U}_N$ be $\boldsymbol{V}$. Thus, the regression problem reduces to a generalised least squares (GLS) problem, where the estimator of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}}_{GLS} = (\boldsymbol{X}_N^T \boldsymbol{V}^{-1} \boldsymbol{X}_N)^{-1} \boldsymbol{X}_N^T \boldsymbol{V}^{-1} \boldsymbol{Y}_N. \tag{4.18}$$

From GLS theory, the variance-covariance matrix of the parameter $\hat{\boldsymbol{\beta}}_{GLS}$ is known to be

$$\text{var}\left(\hat{\boldsymbol{\beta}}_{GLS}\right) = (\boldsymbol{X}_N^T \boldsymbol{V}^{-1} \boldsymbol{X}_N)^{-1}. \tag{4.19}$$

Harvey (1989) showed that the Kalman filter can be used to effectively perform a Cholesky decomposition of $\boldsymbol{V}$. This is done by applying the same Kalman filter to $\boldsymbol{y}_t$ as well as to each of the columns of $\boldsymbol{X}_t$. Hence a $p \times 1$ vector of innovations on the observations $\boldsymbol{y}_t$, say $\boldsymbol{y}_t^*$, and a $p \times q$ matrix of innovations on the explanatory variables $\boldsymbol{X}_t$, say $\boldsymbol{X}_t^*$, are produced. The fact that the same Kalman filter is used for the $\boldsymbol{y}_t$'s and the $\boldsymbol{X}_t$'s suggests that for a given set of parameters $\boldsymbol{\Psi}$, the recursions for $\boldsymbol{P}_{t|t-1}$, $\boldsymbol{P}_{t|t}$ and $\boldsymbol{F}_t$ are run only once, rather than $q+1$ times. The GLS estimator of $\boldsymbol{\beta}$ becomes

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}} = \left[\sum_{t=1}^{n} \boldsymbol{X}_t^{*T} \boldsymbol{F}_t^{-1} \boldsymbol{X}_t^*\right]^{-1} \sum_{t=1}^{n} \boldsymbol{X}_t^{*T} \boldsymbol{F}_t^{-1} \boldsymbol{y}_t^*, \tag{4.20}$$

with innovations

$$\boldsymbol{v}_t = \boldsymbol{y}_t^* - \boldsymbol{x}_t^* \hat{\boldsymbol{\beta}}_{GLS}. \tag{4.21}$$

These innovations can be used for the calculation of the *concentrated log-likelihood*

$$\log L_{\boldsymbol{Y}_N}(\boldsymbol{\Psi}) = -\frac{pn}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{n}\log|\boldsymbol{F}_t| - \frac{1}{2}\sum_{t=1}^{n}\boldsymbol{v}_t^T\boldsymbol{F}_t^{-1}\boldsymbol{v}_t. \tag{4.22}$$

A special application of GLS performed by the Kalman filter is the situation where the initial conditions are considered to be fixed parameters ($\boldsymbol{S}_0 = \boldsymbol{\lambda}_{0|0}$ and $\boldsymbol{P}_{0|0} = \boldsymbol{0}$) which have to be estimated. This approach was first introduced by Wecker and Ansley (2002). They showed that the state vector at time t can be written as

$$\boldsymbol{S}_t = \left[\prod_{j=1}^{t}\boldsymbol{\Phi}_j\right]\boldsymbol{S}_0 + \boldsymbol{S}'_t, \tag{4.23}$$

where $\boldsymbol{S}'_t$ satisfies the following transition equation of the form of Equation (4.14),

$$\boldsymbol{S}'_t = \boldsymbol{\Phi}_t\boldsymbol{S}'_{t-1} + \boldsymbol{\delta}_t, \tag{4.24}$$

which now has the starting value of $\boldsymbol{S}'_0 = \boldsymbol{0}$. Substituting $\boldsymbol{S}_t$ in the measurement equation (4.15) gives

$$\boldsymbol{y}_t = \boldsymbol{Z}_t\boldsymbol{S}'_t + \boldsymbol{X}'_t\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{S}_0\end{bmatrix} + \boldsymbol{\epsilon}_t, \tag{4.25}$$

where

$$\boldsymbol{X}'_t = \left[\boldsymbol{X}_t \;\; \boldsymbol{Z}_t\prod_{j=1}^{t}\boldsymbol{\Phi}_j\right]. \tag{4.26}$$

Thus the model with fixed initial conditions can be written as model (4.24)-(4.26) with initial conditions $\boldsymbol{\lambda}'_{0|0} = \boldsymbol{0}$ and $\boldsymbol{P}'_{0|0} = \boldsymbol{0}$. The parameter vector in this model is simply augmented with $\boldsymbol{S}_0$. An estimate of $\boldsymbol{S}_0$ can then be calculated by the GLS procedure.

4.3.5 The Kalman smoother

In the previous section it was shown how the Kalman filter provides the expected value of the state variable at time t conditional on the information which is available up to this time instant, i.e. $\boldsymbol{\lambda}_{t|t} = \mathrm{E}\left(\boldsymbol{S}_t|\boldsymbol{Y}_t\right)$. In many applications it is useful to incorporate all the information which is available to estimate the state variable at time t. Hence, also the information obtained beyond t has to be incorporated to

estimate $\boldsymbol{S}_t$. This leads to the expected value of the state variable conditional on the entire sample, i.e. $\mathrm{E}\left(\boldsymbol{S}_t|\boldsymbol{Y}_N\right) = \boldsymbol{\lambda}_{t|n}$, which is also referred to as the smoothed estimate and which can be found by applying the Kalman smoother. Because the smoother is based on more information than the filtered estimator, it has a mean squared error which is generally smaller than that of the filtered estimator.

To obtain the smoothed estimates, the Kalman filter should be followed by a set of recursions which are known as the Kalman smoother. The Kalman smoother recursions start with the final quantities, $\boldsymbol{\lambda}_{n|n}$ and $\boldsymbol{P}_{n|n}$ and proceeds backwards. For $t = n-1, \ldots, 0$, it consists of the following backward recursions (Harvey, 1989; Shumway and Stoffer, 2006),

$$\boldsymbol{\lambda}_{t|n} = \boldsymbol{\lambda}_{t|t} + \boldsymbol{J}_t(\boldsymbol{\lambda}_{t+1|n} - \boldsymbol{\lambda}_{t+1|t}) \tag{4.27}$$

$$\boldsymbol{P}_{t|n} = \boldsymbol{P}_{t|t} + \boldsymbol{J}_t(\boldsymbol{P}_{t+1|n} - \boldsymbol{P}_{t+1|t})\boldsymbol{J}_t^T \tag{4.28}$$

$$\boldsymbol{J}_t = \boldsymbol{P}_{t|t}\Phi_{t+1}^T\boldsymbol{P}_{t+1|t}^{-1}. \tag{4.29}$$

Digalakis et al. (1993) provided recursions for the calculation of the lag one covariance estimators $\boldsymbol{P}_{t,t-1|s} = \mathrm{cov}\left(\boldsymbol{S}_t, \boldsymbol{S}_{t-1}|\boldsymbol{Y}_s\right)$. Filtered values can be calculated by the additional forward recursion

$$\boldsymbol{P}_{t,t-1|t} = (\boldsymbol{I} - \boldsymbol{P}_{t|t-1}\boldsymbol{Z}_t^T\boldsymbol{F}_t^{-1}\boldsymbol{Z}_t)\boldsymbol{\Phi}_t\boldsymbol{P}_{t-1|t-1}, \tag{4.30}$$

and smoothed values can be obtained by the additional backward recursion

$$\boldsymbol{P}_{t,t-1|n} = \boldsymbol{P}_{t,t-1|t} + (\boldsymbol{P}_{t|n} - \boldsymbol{P}_{t|t})\boldsymbol{P}_{t|t}^{-1}\boldsymbol{P}_{t,t-1|t}. \tag{4.31}$$

4.4 Maximum likelihood estimation

4.4.1 Introduction

We have already introduced the Kalman filter as a tool for the calculation of the likelihood. For obtaining maximum likelihood parameter estimates, the likelihood function has to be maximised. This can be done numerically by using classical algorithms such as the Newton Raphson approach. When the state-space model is time-invariant, another possibility was introduced by Shumway and Stoffer (1982). They derived an Expectation-Maximisation (EM) type algorithm to obtain maximum likelihood estimates of the parameters. The basic idea of EM algorithms was

introduced by Dempster et al. (1977). It provides maximum likelihood estimates in incomplete data situations. A nice property of EM algorithms is that under certain conditions, the likelihood cannot decrease throughout the iterations. Hence, the likelihood always converges to a local maximum (McLachlan and Krishnan, 1997). An EM algorithm can be specified for the state-space setting, as the unobservable state can be considered as missing data.

4.4.2 EM algorithm

The EM algorithm which is considered here, is based on Shumway and Stoffer (2006). They presented an EM algorithm for time-invariant state-space models without exogenous predictors. For time-invariant state-space models, the system matrices $\boldsymbol{\Phi}_t$, $\boldsymbol{Q}_t$, $\boldsymbol{H}_t$ and $\boldsymbol{Z}_t$ are constant and the index t can thus be dropped. First, we should act as if the state vector is observable. In this case we may consider $(\boldsymbol{S}_N, \boldsymbol{Y}_N)$ as the complete data, and their joint log-likelihood is given by

$$\begin{aligned}\log L_{\boldsymbol{Y},\boldsymbol{S}}(\boldsymbol{\Psi}) \sim & -\frac{1}{2}\log|\boldsymbol{\Sigma}_{S_0}| - \frac{1}{2}(\boldsymbol{S}_0 - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{S_0}^{-1}(\boldsymbol{S}_0 - \boldsymbol{\mu}_0) \\ & - \frac{n}{2}\log|\boldsymbol{Q}| - \frac{1}{2}\sum_{t=1}^{n}(\boldsymbol{S}_t - \boldsymbol{\Phi}\boldsymbol{S}_{t-1})^T\boldsymbol{Q}^{-1}(\boldsymbol{S}_t - \boldsymbol{\Phi}\boldsymbol{S}_{t-1}) \\ & - \frac{n}{2}\log|\boldsymbol{H}| - \frac{1}{2}\sum_{t=1}^{n}(\boldsymbol{Y}_t - \boldsymbol{Z}_t\boldsymbol{S}_t)^T\boldsymbol{H}^{-1}(\boldsymbol{Y}_t - \boldsymbol{Z}_t\boldsymbol{S}_t), \quad (4.32)\end{aligned}$$

which is referred to as the *completed log-likelihood*.

This likelihood cannot be calculated because the state variable is unobservable. The EM algorithm overcomes this by iterating between two steps, a so-called **E-step** and an **M-step**. In the **E-step** of the $(k+1)^{th}$ iteration the conditional expected value of the completed likelihood given the observed data $\boldsymbol{Y}_N$ and the current value of the parameter estimates $\boldsymbol{\Psi}^k$ is calculated. This conditional expectation is given by

$$Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^k) = \mathrm{E}\left(-2\log L_{\boldsymbol{Y},\boldsymbol{S}}(\boldsymbol{\Psi})|\boldsymbol{Y}_N, \boldsymbol{\Psi}^k\right) \quad (4.33)$$

Hence,

$$
\begin{aligned}
Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^k) \sim & \mathrm{E}\left(\log|\boldsymbol{\Sigma}_{S_0}| + (\boldsymbol{S}_0 - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_{S_0}^{-1}(\boldsymbol{S}_0 - \boldsymbol{\mu}_0)|\boldsymbol{Y}_N, \boldsymbol{\Psi}^k\right) \\
& + \mathrm{E}\left(n\log|\boldsymbol{Q}| + \sum_{t=1}^{n}(\boldsymbol{S}_t - \boldsymbol{\Phi}\boldsymbol{S}_{t-1})^T \boldsymbol{Q}^{-1}(\boldsymbol{S}_t - \boldsymbol{\Phi}\boldsymbol{S}_{t-1})|\boldsymbol{Y}_N, \boldsymbol{\Psi}^k\right) \\
& + \mathrm{E}\left(n\log|\boldsymbol{H}| + \sum_{t=1}^{n}(\boldsymbol{Y}_t - \boldsymbol{Z}_t\boldsymbol{S}_t)^T \boldsymbol{H}^{-1}(\boldsymbol{Y}_t - \boldsymbol{Z}_t\boldsymbol{S}_t)|\boldsymbol{Y}_N, \boldsymbol{\Psi}^k\right). \quad (4.34)
\end{aligned}
$$

Shumway and Stoffer (2006) showed that this yields

$$
\begin{aligned}
Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^k) \sim & \log|\boldsymbol{\Sigma}_{S_0}| + \mathrm{tr}\left\{\boldsymbol{\Sigma}_{S_0}^{-1}\left[\boldsymbol{P}_{0|n} + (\boldsymbol{\lambda}_{0|n} - \boldsymbol{\mu}_0)(\boldsymbol{\lambda}_{0|n} - \boldsymbol{\mu}_0)^T\right]\right\} \\
& + n\log|\boldsymbol{Q}| + \mathrm{tr}\left\{\boldsymbol{Q^{-1}}\left[\boldsymbol{A}_{11} - \boldsymbol{A}_{10}\boldsymbol{\Phi}^T - \boldsymbol{\Phi}\boldsymbol{A}_{10}^T + \boldsymbol{\Phi}\boldsymbol{A}_{00}\boldsymbol{\Phi}^T\right]\right\} + n\log|\boldsymbol{H}| \\
& + \mathrm{tr}\left\{\boldsymbol{H^{-1}}\sum_{t=1}^{n}\left[(\boldsymbol{y}_t - \boldsymbol{Z}_t\boldsymbol{\lambda}_{t|n})(\boldsymbol{y}_t - \boldsymbol{Z}_t\boldsymbol{\lambda}_{t|n})^T + \boldsymbol{Z}_t\boldsymbol{P}_{t|n}\boldsymbol{Z}_t^T\right]\right\}, \quad (4.35)
\end{aligned}
$$

where

$$\boldsymbol{A}_{11} = \sum_{t=1}^{n}(\boldsymbol{\lambda}_{t|n}\boldsymbol{\lambda}_{t|n}^T + \boldsymbol{P}_{t|n}), \quad (4.36)$$

$$\boldsymbol{A}_{10} = \sum_{t=1}^{n}(\boldsymbol{\lambda}_{t|n}\boldsymbol{\lambda}_{t-1|n}^T + \boldsymbol{P}_{t,t-1|n}), \quad (4.37)$$

and

$$\boldsymbol{A}_{10} = \sum_{t=1}^{n}(\boldsymbol{\lambda}_{t-1|n}\boldsymbol{\lambda}_{t-1|n}^T + \boldsymbol{P}_{t-1|n}). \quad (4.38)$$

(4.39)

In the subsequent **M-step**, the expected log-likelihood has to be maximised, or, alternatively, $Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^k)$ has to be minimised so as to obtain the update of the parameter set $\boldsymbol{\Psi}^{k+1}$. This yields

$$\boldsymbol{\Phi}^{k+1} = \boldsymbol{A}_{10}\boldsymbol{A}_{00}^{-1}, \quad (4.40)$$

$$\boldsymbol{Q}^{k+1} = n^{-1}(\boldsymbol{A}_{11} - \boldsymbol{A}_{10}\boldsymbol{A}_{00}^{-1}\boldsymbol{A}_{10}^T), \quad (4.41)$$

and

$$\boldsymbol{H}^{k+1} = n^{-1}\sum_{t=1}^{n}\left[(\boldsymbol{y}_t - \boldsymbol{Z}_t\boldsymbol{\lambda}_{t|n})(\boldsymbol{y}_t - \boldsymbol{Z}_t\boldsymbol{\lambda}_{t|n})^T + \boldsymbol{Z}_t\boldsymbol{P}_{t|n}\boldsymbol{Z}_t^T\right]. \quad (4.42)$$

The updates for the initial mean and covariance matrix are

$$\boldsymbol{\mu}_0^{k+1} = \boldsymbol{\lambda}_{0|n} \text{ and } \boldsymbol{\Sigma}_0^{k+1} = \boldsymbol{P}_{0|n}. \tag{4.43}$$

Thus, the overall procedure alternates between the **E-step** where the Kalman filter and the Kalman smoother are calculated, and the **M-step** which consists of the parameter estimates updates (4.40)-(4.43).

Now that it is clear how the parameters can be estimated, a method is needed to calculate the variance of these estimators so that their uncertainty can be assessed and inference becomes possible. The variance-covariance matrix of the parameter estimators can be obtained by perturbation or by the derivation of the Fisher Information Matrix (FIM) as explained below.

4.4.3 Fisher information matrix

When parameter estimators are obtained by maximum likelihood, the Fisher information matrix (FIM) provides an estimate of the inverse of their covariance matrix. The prediction error decomposition can be used to obtain the expected FIM. Let $\boldsymbol{I}$ denote the FIM, and its ij^{th} element is given by

$$\boldsymbol{I}_{ij}(\boldsymbol{\Psi}) = \frac{1}{2}\sum_{t=1}^{n}\left[tr\left(\boldsymbol{F}_t^{-1}\frac{\partial \boldsymbol{F}_t}{\partial \Psi_i}\boldsymbol{F}_t^{-1}\frac{\partial \boldsymbol{F}_t}{\partial \Psi_j}\right)\right] + \mathrm{E}\left(\sum_{t=1}^{n}\left(\frac{\partial \boldsymbol{v_t}}{\partial \Psi_i}\right)^T \boldsymbol{F}_t^{-1}\frac{\partial \boldsymbol{v_t}}{\partial \Psi_j}\right). \tag{4.44}$$

In some cases the observed FIM is easier to evaluate, and is obtained by dropping the expectation operator in the second term. The derivatives of $\boldsymbol{F}_t$ and $\boldsymbol{v}_t$ may be obtained numerically by perturbation. This requires an additional pass of the Kalman filter for each parameter value Ψ_j to obtain the perturbed version $\boldsymbol{F}_t^{per_j}$ and $\boldsymbol{v}_t^{per_k}$. The derivatives are then approximated by

$$\frac{\partial \boldsymbol{F}_t}{\partial \Psi_j} \approx \frac{\left[\boldsymbol{F}_t^{per_j} - \boldsymbol{F}_t\right]}{\Psi_j^{per_j} - \Psi_j} \tag{4.45}$$

$$\frac{\partial \boldsymbol{v}_t}{\partial \Psi_j} \approx \frac{\left[\boldsymbol{v}_t^{per_j} - \boldsymbol{v}_t\right]}{\Psi_j^{per_j} - \Psi_j}. \tag{4.46}$$

4.5 Summary

The state-space model representation enables the modeller to use the Kalman filter and smoother recursions. These recursions can be exploited to perform generalised least squares estimation of the parameter of the mean model. They are also useful for the calculation of the likelihood when all error terms and the initial conditions are Gaussian. Maximum likelihood estimators of the parameters of the system matrices can be obtained by the use of an EM algorithm and a covariance matrix of these estimators can easily be calculated by the evaluation of the observed FIM. In the next chapters, the state-space representation is used to formulate a spatio-temporal model for river monitoring networks. The specific spatial structure of river networks, however, imposes some restrictions on the system matrices. Therefore, an adjusted version of the EM algorithm is needed.

Chapter 5

Spatio-temporal modelling of river monitoring networks, a parametric approach

5.1 Introduction

In the light of the European Water Framework Directive (WFD)(EC, 2000), it is important for environmental agencies and policy makers to dispose of ICT tools to assess the evolution/change of the water quality. This assessment should be possible at individual sampling locations as well as on a more regional scale. Existing statistical techniques cannot be used for this purposes because the data originating from environmental monitoring networks are clearly not independent. They are

sampled from a dynamic process that evolves over space and time. Hence, the statistical methodology should incorporate these spatio-temporal dependences so as to allow valid statistical inference. Traditional approaches to address this problem have focused on the geo-statistical paradigm (Bilonick, 1983; Cressie and Majure, 1997) and on multivariate time-series methods, which specify dynamic models that are linked spatially (Rouhani and Wackernagel, 1990). If both temporal and spatial components are present, it is natural to combine them in a statistical model that is temporally dynamic and spatially descriptive. Such a model is referred to as a space-time dynamic model (Wikle and Cressie, 1999).

Wikle and Cressie (1999) classify time-series as dynamic since the temporal dependence arises from a unidirectional correlation; the AR(1) model is a clear example. This unidirectional structure is often utilised in time series techniques. Geo-statistical methods, on the other hand, are classified as descriptive because of the non-directional correlation, there is no causal interpretation associated with the observed spatial correlation. Based on these considerations Huang and Cressie (1996) derived a temporal dynamic and spatially descriptive Kalman filter.

In this chapter we develop a spatio-temporal model for the analysis of river monitoring network data. With respect to the spatial dependence structure an important distinction has to be made with the classical spatial structures. Since the water flows only in one direction within the river reaches, a causal interpretation can be given to the correlations. However, in contrast to time, rivers can join or split. This implies a more general branched unidirectional structure. Therefore, according to Cressie's terminology, the presented spatio-temporal model is dynamic w.r.t. both the spatial and the temporal dependence structure. Once we know how to build spatio-temporal models that can deal with the specific dependence structure of a river network, we can perform an assessment that incorporates the dependence structure correctly.

To answer the question of interest we still need a model for the mean. Two paradigms can be used for this purpose: the marginal and the conditional modelling paradigm. Diggle et al. (1994) suggested that the choice between both paradigms should be motivated by the research question. Since many environmental problems are clearly related to the marginal mean, we have adopted the marginal modelling paradigm and we model the marginal mean and dependence structure separately.

The focus in this dissertation is on the assessment of the observed data at the sampling locations and we do not aim to perform predictions at intermediate locations

that are not sampled. Therefore the observations of the monitoring network at a certain time instant can be considered as the realisation of a finite-dimensional multivariate random variable with each dimension corresponding to each of the p sampling locations. This enables us to write the model as a p-dimensional state-space model. The state-space model representation allows the use of the Kalman filter and smoother recursions for estimation purposes. In particular, the Kalman filter provides a convenient factorisation of the likelihood (e.g. Harvey, 1989 and Shumway and Stoffer, 2006). For the maximisation of the likelihood function, we use an expectation-maximisation (EM) algorithm (e.g. Shumway and Stoffer, 1982, Harvey, 1989 and Shumway and Stoffer, 2006). A general introduction to state-space models, the Kalman filter and smoother and the EM algorithm for parameter estimation can be found in Chapter 4.

To deal with specific spatio-temporal structures or to reduce the computational burden in large monitoring networks, restrictions are imposed on the model matrices of the state-space model. Xu and Wikle (2005) proposed several parametrisations for spatio-temporal models with a descriptive spatial component. The EM algorithm specified by Shumway and Stoffer (1982), however, assumes a saturated parametrisation of the state-space model and updates all elements of the system matrices. However, in a restricted model specification, a number of elements of the system matrices are known (fixed), e.g. when certain elements of the state process can be assumed to be independent, the covariance matrix of the state process will contain a number of zeroes. Applying the EM algorithm of Shumway and Stoffer (1982) on such a restricted state-space model, will also update these fixed parameters. Hence, after each update of the algorithm, the fixed parameters should immediately be imputed by the known values. Xu and Wikle (2005) argued that it is not clear whether this approach leads to maximum likelihood estimates. Therefore they suggested to adjust the EM algorithm so as to take the restrictions directly into account.

In this chapter we develop a spatio-temporal model for the analysis of data originating from river monitoring networks. The river topology is used to define a spatio-temporal model that is dynamic with respect to both the spatial and the temporal dependence structure. In reality the environmental conditions may obscure the unidirectional spatial dependence structure implied by the river topology. Due to this confounding factor, the state, say $\boldsymbol{S}$, of the underlying river process cannot be observed. We therefore propose to embed the latent variable $\boldsymbol{S}$ in an observation model that allows cross-correlation between sampling locations that are located at different branches of the river. To formulate the model for the mean, we adopt the

marginal modelling paradigm and we model the marginal mean and dependence structure separately. In this chapter a linear mean model is used. For parameter estimation, we adjust the EM-algorithm to take the specific restrictions implied by the river topology explicitly into account. To handle the exogenous predictors in the mean model, the EM-algorithm is further modified to an expectation-conditional-maximisation (ECM) algorithm. An ECM algorithm is a natural extension of the EM algorithm obtained by replacing its M-step by a number of computationally simpler conditional maximisation (CM) steps. It keeps the attractive property that, under suitable conditions, the likelihood does not decrease at any iteration (Meng and Rubin, 1993; McLachlan and Krishnan, 1997).

This chapter is organised as follows. In Section 5.2 the model is formulated and in Section 5.3 the ECM algorithm for parameter estimation is given. Finally, in Section 5.4, the model is applied to real data where the annual mean of nitrate in a certain region is compared to the annual means of previous years.

5.2 Spatio-temporal model

First, in subsections 5.2.1 and 5.2.2 a zero-mean model is constructed. The complete model is given in subsection 5.2.3.

5.2.1 Spatial dependence structure

Let the $p \times 1$ vector $\boldsymbol{S} = (S_1, \ldots, S_p)^T$ denote a stationary spatial process, where S_i $(i = 1, ..., p)$ represents the response variable at sampling location i. The correlation structure of $\boldsymbol{S}$ is completely defined by its conditional dependence structure which is directly derived from the river monitoring network architecture. This is illustrated in Figure 5.1 which shows 5 sampling locations along 2 joining river reaches. The direction of the flow is also indicated. The same figure can also be interpreted as a Directed Acyclic Graph (DAG) (see e.g. Whittaker, 1990) in which the circles represent the graph's vertices associated with the corresponding S_i's. Missing edges or arrows indicate the conditional independences. Thus from Figure 5.1 we read $S_1 \perp\!\!\!\perp S_3$; $S_2 \perp\!\!\!\perp S_3$; $S_4 \perp\!\!\!\perp S_1 | S_2$; $S_5 \perp\!\!\!\perp S_1 | S_2$; $S_5 \perp\!\!\!\perp S_1 | S_4$; $S_5 \perp\!\!\!\perp S_2 | S_4$ and $S_5 \perp\!\!\!\perp S_3 | S_4$. The DAG implies zeroes in the variance-covariance matrix of $\boldsymbol{S}$. Thus it can equivalently be represented by a recursive system of equations (Wer-

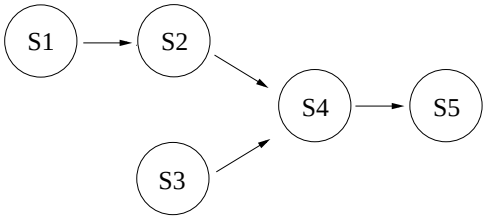

Figure 5.1: Directed Acyclic Graph (DAG) of five sampling locations along two joining river reaches

muth, 1980),

$$\boldsymbol{S} = \boldsymbol{A}\boldsymbol{S} + \boldsymbol{\gamma}, \tag{5.1}$$

where the order of the elements of $\boldsymbol{S}$ can always be rearranged so that $\boldsymbol{A}$ is a $p \times p$ lower triangular square matrix with zeroes at the diagonal, and $\boldsymbol{\gamma}$ is a $p \times 1$ multivariate zero-mean random vector with a diagonal variance-covariance matrix $\boldsymbol{\Sigma}_\gamma$. We further assume that $\boldsymbol{\gamma} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_\gamma)$. For the DAG represented in Figure 5.1, $\boldsymbol{A}$ becomes

$$\boldsymbol{A} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ a_{21} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & a_{42} & a_{43} & 0 & 0 \\ 0 & 0 & 0 & a_{54} & 0 \end{bmatrix}$$

where a_{ij} models the dependence between sampling location S_i and S_j.

5.2.2 Spatio-temporal dependence structure

In a river monitoring network the data are gathered over time. Vector $\boldsymbol{S_t} = (S_{1t}, \ldots, S_{pt})^T$ now represents the observations at the sampling locations at time t ($t = 1, \ldots, n$). So the dependence structure has to be extended with a temporal component which we assume to be autoregressive of order 1 (AR(1)). After fitting the model, the quality of the proposed temporal structure has to be assessed in an analysis of the innovations. To incorporate the temporal structure, Equation (5.1) is extended to

$$\boldsymbol{S}_t = \boldsymbol{A}\boldsymbol{S}_t + \boldsymbol{B}\boldsymbol{S}_{t-1} + \boldsymbol{\eta}_t, \tag{5.2}$$

where $\boldsymbol{B}$ is a $p \times p$ matrix containing the temporal autocorrelation coefficients (diagonal elements) and the spatio-temporal cross-correlation coefficients (off-diagonal elements), and $\boldsymbol{\eta}_t \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_\eta)$ with a $p \times p$ diagonal variance-covariance matrix $\boldsymbol{\Sigma}_\eta$. Similar to matrix $\boldsymbol{A}$, we propose to use only cross-correlations between sampling locations which are directly connected according to the DAG structure. The off-diagonal elements of $\boldsymbol{B}$ are thus structured in a similar way as the elements of matrix $\boldsymbol{A}$. Hence $\boldsymbol{B}$ can be written as

$$\boldsymbol{B} = \begin{bmatrix} b_{11} & 0 & 0 & 0 & 0 \\ b_{21} & b_{22} & 0 & 0 & 0 \\ 0 & 0 & b_{33} & 0 & 0 \\ 0 & b_{42} & b_{43} & b_{44} & 0 \\ 0 & 0 & 0 & b_{54} & b_{55} \end{bmatrix}.$$

When $i \neq j$ the b_{ij} model the spatio-temporal dependence between S_{it} and S_{jt-1} and the b_{ii} model the temporal dependence between S_{it} and S_{it-1}.

Again, this assumption has to be assessed in an analysis of the innovations. Equation (5.2) can be reorganised so that the model can be written in its general state-space model representation ,

$$\boldsymbol{S}_t = \boldsymbol{\Phi}\boldsymbol{S}_{t-1} + \boldsymbol{\delta}_t, \tag{5.3}$$

where $\boldsymbol{\Phi} = (\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{B}$ and $\boldsymbol{\delta}_t \sim MVN(\mathbf{0}, \boldsymbol{Q})$ with covariance matrix $\boldsymbol{Q} = (\boldsymbol{I}-\boldsymbol{A})^{-1}\boldsymbol{\Sigma}_\eta(\boldsymbol{I}-\boldsymbol{A})^{-T}$ and $t = 1, \ldots, n$. For the model to be completely defined, we assume $\boldsymbol{S}_0$ to be multivariate normally distributed, i.e. $\boldsymbol{S}_0 \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_{S_0})$.

5.2.3 Observation model

In reality, however, the dependence structure presented in Model (5.3) might possibly be obscured by common environmental influences such as rainfall or climatological conditions in general. The rather strict structure implied by Model (5.3) is therefore assumed to hold only for an isolated river system. To allow for common environmental disturbances, the unobservable state variable $\boldsymbol{S}$ is embedded into an observation model,

$$\boldsymbol{y}_t = \boldsymbol{S}_t + \boldsymbol{\epsilon}_t, \tag{5.4}$$

($t = 1, \ldots, n$), where $\boldsymbol{y}_t$ is the $p \times 1$ observation vector corresponding to $\boldsymbol{S}_t$, and $\boldsymbol{\epsilon}_t$ is a zero-mean error term. In particular $\boldsymbol{\epsilon}_t \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$. No restrictions are imposed on $\boldsymbol{\Sigma}_\epsilon$ which enables cross-correlations between sampling locations

that are not connected according to the river topology. This specification makes the spatio-temporal model given by Equations (5.3) and (5.4), a state-space model.

So far we have assumed that the mean of $\boldsymbol{y}_t$ is zero, i.e. $\mathrm{E}(\boldsymbol{y}_t) = \mathbf{0}$ for all t. This can be further extended to a linear model, e.g. $\mathrm{E}(\boldsymbol{y}_t) = \boldsymbol{X}_t\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1, ..., \beta_q)^T$ is the $q \times 1$ parameter vector and $\boldsymbol{X_t}$ is the $p \times q$ design matrix which may contain time-dependent covariates. After embedding the mean model into Model (5.4) we obtain

$$\boldsymbol{y}_t = \boldsymbol{S}_t + \boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{\epsilon}_t, \tag{5.5}$$

which specifies together with Model (5.3) the complete spatio-temporal state-space model. Note that this state-space model is time-invariant because the system matrices $\boldsymbol{\Phi}$, $\boldsymbol{Q}$ and $\boldsymbol{\Sigma}_\epsilon$ do not change over time.

Another equivalent formulation of the spatio-temporal model is accomplished by recognising that the Model (5.3) and (5.5) can be written as a Structural Equation Model (SEM) (see e.g. Maruyama, 1997),

$$\boldsymbol{C}\boldsymbol{S}_N = \boldsymbol{\zeta} \tag{5.6}$$

$$\boldsymbol{Y}_N = \boldsymbol{X}_N\boldsymbol{\beta} + \boldsymbol{S}_N + \boldsymbol{\psi}, \tag{5.7}$$

where $\boldsymbol{S}_N = (\boldsymbol{S}_1^T, \ldots, \boldsymbol{S}_n^T)^T$, $\boldsymbol{Y}_N = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$, $\boldsymbol{X}_N = (\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_n^T)^T$, $\boldsymbol{C}$ is a $pn \times pn$ square matrix constructed from the elements of the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{\zeta} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_\zeta)$, where $\boldsymbol{\Sigma}_\zeta$ is a diagonal matrix built from the corresponding elements of $\boldsymbol{\Sigma}_\eta$, and $\boldsymbol{\psi} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_\psi)$ where $\boldsymbol{\Sigma}_\psi$ is block-diagonal with blocks $\boldsymbol{\Sigma}_\epsilon$. From this SEM formulation the covariance structure of the observation vector $\boldsymbol{Y}$ is readily found,

$$\boldsymbol{\Sigma}_{Y_N}(\boldsymbol{\Psi}_\alpha) = \mathrm{var}(\boldsymbol{Y}_N) = \boldsymbol{C}^{-1}\boldsymbol{\Sigma}_\zeta\boldsymbol{C}^{-T} + \boldsymbol{\Sigma}_\psi, \tag{5.8}$$

with $\boldsymbol{\Psi}_\alpha$ a vector containing all the parameters in $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{\Sigma}_{S_0}$, $\boldsymbol{\Sigma}_\eta$, and $\boldsymbol{\Sigma}_\epsilon$. Both representations will lead to slightly different approaches to estimate the parameters.

5.3 Parameter estimation and statistical inference

The parameter estimation is based on maximum likelihood. In Section 5.3.1 we formulate the likelihood. For the calculation of the likelihood we apply the Kalman

filter (Section 5.3.2) for it enables a very natural factorisation of the likelihood of state-space models. Numerical maximisation of the likelihood is done by an ECM algorithm. In Section 5.3.3 an ECM algorithm is derived in case the state-space representation is used. In Section 5.3.4 this algorithm is adjusted to provide the use of the SEM representation. In Section 5.3.5 we conclude with a brief account on model selection criteria and the joint asymptotic distribution of the parameter estimators.

5.3.1 Likelihood

Our state-space model is basically a statistical model representation of the observation vector $\boldsymbol{Y}_N$. It implies that

$$\boldsymbol{Y}_N \sim MVN(\boldsymbol{X}_N\boldsymbol{\beta}, \boldsymbol{\Sigma}_{Y_N}(\boldsymbol{\Psi}_\alpha)). \tag{5.9}$$

The variance-covariance matrix $\boldsymbol{\Sigma}_{Y_N}(\boldsymbol{\Psi}_\alpha)$ is completely parameterised by the elements of $\boldsymbol{\Psi}_\alpha$. Maximum likelihood is thus a natural framework for parameter estimation. The log-likelihood is given by

$$\log L_{Y_N}(\boldsymbol{\Psi}) \sim -\frac{1}{2}(\boldsymbol{Y}_N - \boldsymbol{X}_N\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}_{\boldsymbol{Y}_N}(\boldsymbol{\Psi}_{\boldsymbol{\alpha}})(\boldsymbol{Y}_N - \boldsymbol{X}_N\boldsymbol{\beta}) - \frac{1}{2}\log|\boldsymbol{\Sigma}_{Y_N}(\boldsymbol{\Psi}_\alpha)|, \tag{5.10}$$

where $\boldsymbol{\Psi}$ is the vector containing all parameters of the model ($\boldsymbol{\Psi}_\alpha, \boldsymbol{\beta}$).

Conditional on $\boldsymbol{\Psi}_\alpha$, the likelihood is maximised by the general least squares (GLS) estimator

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\Psi}_\alpha) = (\boldsymbol{X}_N^T\boldsymbol{\Sigma}^{-1}_{Y_N}(\boldsymbol{\Psi}_{\boldsymbol{\alpha}})\boldsymbol{X}_N)^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}_{Y_N}(\boldsymbol{\Psi}_\alpha)\boldsymbol{Y}_N. \tag{5.11}$$

Substitution of $\hat{\boldsymbol{\beta}}(\boldsymbol{\Psi}_\alpha)$ into Equation (5.10) gives the concentrated log-likelihood for $\boldsymbol{\Psi}_\alpha$,

$$\log L_{Y_N}(\boldsymbol{\Psi}_\alpha) \sim -\frac{1}{2}(\boldsymbol{Y}_N - \boldsymbol{X}_N\hat{\boldsymbol{\beta}}(\boldsymbol{\Psi}_\alpha))^T\boldsymbol{\Sigma}^{-1}_{Y_N}(\boldsymbol{\Psi}_\alpha)(\boldsymbol{Y}_N - \boldsymbol{X}_N\hat{\boldsymbol{\beta}}(\boldsymbol{\Psi}_\alpha)) - \frac{1}{2}\log|\boldsymbol{\Sigma}_{Y_N}(\boldsymbol{\Psi}_\alpha)|. \tag{5.12}$$

Maximisation of $\log L_{Y_N}(\boldsymbol{\Psi}_{\boldsymbol{\alpha}})$ yields $\hat{\boldsymbol{\Psi}}_\alpha$ and by substitution of $\hat{\boldsymbol{\Psi}}_\alpha$ in Equation (5.11), the maximum likelihood estimator $\hat{\boldsymbol{\beta}} \equiv \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\Psi}}_\alpha)$ is obtained. If in Equation (5.11), $\boldsymbol{\Sigma}_{Y_N}(\boldsymbol{\Psi}_\alpha)$ is substituted by an estimator, then $\hat{\boldsymbol{\beta}}$ is known as the *feasible generalised least squares estimator* (FGLS, e.g. Prucha (1984)). The state-space representation of the model admits the use of the Kalman filter and smoother. In particular, the Kalman filter enables a further factorisation of the likelihood in Equation (5.10) because the initial conditions and all residual processes are Gaussian. This distributional assumption has to be checked for in the assessment of the innovations. Another interesting feature of the Kalman filter is its use to calculate the GLS estimates of the parameters of the mean model.

5.3.2 Kalman filter and smoother

When all the parameters of the state-space model $\boldsymbol{\Psi} = (\boldsymbol{\Psi}_{\boldsymbol{\alpha}}, \boldsymbol{\beta})$ are known, the Kalman filter and smoother recursions can be used to calculate the conditional mean and covariance of the state variables (e.g. Harvey, 1989) which will be used in the algorithm to maximise the log-likelihood. Although these recursions are already introduced in Chapter 4, they are presented here for completeness.

First, the conditional mean of the state variable $\boldsymbol{S}_t$, given $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s$ is denoted by $\boldsymbol{\lambda}_{t|s} = E[\boldsymbol{S}_t|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s]$. In particular $\boldsymbol{\lambda}_{t|t-1}$, $\boldsymbol{\lambda}_{t|t}$ and $\boldsymbol{\lambda}_{t|n}$ are referred to as the predicted, filtered and smoothed values, respectively. Similarly, the conditional covariance matrix is denoted by $\boldsymbol{P}_{t|s} = \text{var}\,(\boldsymbol{S}_t|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s)$ and the lag one covariance matrix $\boldsymbol{P}_{t,t-1|s} = \text{cov}\,(\boldsymbol{S}_t, \boldsymbol{S}_{t-1}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s)$. Finally, the innovations are defined as $\boldsymbol{v}_t = \boldsymbol{y}_t - \boldsymbol{\lambda}_{t|t-1} - \boldsymbol{X}_t\boldsymbol{\beta}$ and they have the corresponding covariance matrix $\boldsymbol{F}_t = \boldsymbol{P}_{t|t-1} + \boldsymbol{\Sigma}_\epsilon$. The predicted and filtered values are given by the Kalman filter (e.g. Harvey, 1989). For $t = 1, \ldots, n$ the following forward recursions are used and they are started with time $t = 1$,

$$\boldsymbol{\lambda}_{t|t-1} = \boldsymbol{\Phi}\boldsymbol{\lambda}_{t-1|t-1} \tag{5.13}$$

$$\boldsymbol{P}_{t|t-1} = \boldsymbol{\Phi}\boldsymbol{P}_{t-1|t-1}\boldsymbol{\Phi}^T + \boldsymbol{Q} \tag{5.14}$$

$$\boldsymbol{\lambda}_{t|t} = \boldsymbol{\lambda}_{t|t-1} + \boldsymbol{P}_{t|t-1}\boldsymbol{F}_t^{-1}\boldsymbol{v_t} \tag{5.15}$$

$$\boldsymbol{P}_{t|t} = \boldsymbol{P}_{t|t-1} - \boldsymbol{P}_{t|t-1}\boldsymbol{F}_t^{-1}\boldsymbol{P}_{t|t-1} \tag{5.16}$$

$$\boldsymbol{F}_t = \boldsymbol{P}_{t|t-1} + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}, \tag{5.17}$$

where $\boldsymbol{\lambda}_{0|0} = E[\boldsymbol{S}_0] = \boldsymbol{0}$ and $\boldsymbol{P}_{0|0} = \text{var}\,(\boldsymbol{S}_0) = \boldsymbol{\Sigma}_{S_0}$.

Smoothed estimates are given by the Kalman smoother. It starts on time $t = n$ with the final quantities, $\boldsymbol{\lambda}_{n|n}$ and $\boldsymbol{P}_{n|n}$ and then proceeds backwards. The Kalman smoother consists of the following backward recursions (Harvey, 1989), for time $t = n-1, \ldots, 0$

$$\boldsymbol{\lambda}_{t|n} = \boldsymbol{\lambda}_{t|t} + \boldsymbol{J}_t(\boldsymbol{\lambda}_{t+1|n} - \boldsymbol{\lambda}_{t+1|t}) \tag{5.18}$$

$$\boldsymbol{P}_{t|n} = \boldsymbol{P}_{t|t} + \boldsymbol{J}_t(\boldsymbol{P}_{t+1|n} - \boldsymbol{P}_{t+1|t})\boldsymbol{J}_t^T \tag{5.19}$$

$$\boldsymbol{J}_t = \boldsymbol{P}_{t|t}\Phi^T\boldsymbol{P}_{t+1|t}^{-1}. \tag{5.20}$$

Digalakis et al. (1993) provided recursions for the calculation of the lag one covariance estimators. Filtered values can be calculated by the additional forward recursion

$$\boldsymbol{P}_{t,t-1|t} = (\boldsymbol{I} - \boldsymbol{P}_{t|t-1}\boldsymbol{F}_t^{-1})\boldsymbol{\Phi}\boldsymbol{P}_{t-1|t-1}, \tag{5.21}$$

and smoothed values can be obtained by the additional backward recursion

$$\boldsymbol{P}_{t,t-1|n} = \boldsymbol{P}_{t,t-1|t} + (\boldsymbol{P}_{t|n} - \boldsymbol{P}_{t|t})\boldsymbol{P}_{t|t}^{-1}\boldsymbol{P}_{t,t-1|t}. \tag{5.22}$$

In our application, the system matrices $\boldsymbol{\Phi}$, $\mathbf{Q}$ and $\boldsymbol{\Sigma}_\epsilon$ are time-invariant. Time-invariant state-space systems are stationary when the eigenvalues of $\boldsymbol{\Phi}$ are located in the unit circle (Harvey, 1989). For a time-invariant stationary state-space system with a positive semidefinite initial covariance matrix $\mathbf{P}_{1|0}$, the $\mathbf{P}_{t|t-1}$ becomes time-invariant and the Kalman filter is known to converge exponentially to steady state (Harvey, 1989). Hence, once the Kalman filter has converged, Equations (5.14), (5.16) and (5.17) become redundant. Computationally, this is very interesting because the calculations of $\mathbf{P}_{t|t-1}$, $\mathbf{P}_{t|t}$ and $\mathbf{F}_t$ are the most time-consuming part of the Kalman filter.

Harvey (1989) also showed that the Kalman filter can be used to perform GLS to obtain parameter estimates of the mean model. When the same Kalman filter is applied to $\boldsymbol{y}_t$ and each of the columns of $\boldsymbol{X}_t$, the filter can be used to effectively perform a Cholesky decomposition of $\boldsymbol{\Sigma}_{Y_N}(\boldsymbol{\Psi}_\alpha)$ (Harvey, 1989). Hence, a $p \times 1$ vector of innovations $\boldsymbol{y}_t^*$, on the observations $\boldsymbol{y}_t$ and a $p \times q$ matrix of innovations $\boldsymbol{X}_t^*$, on the explanatory variables $(\boldsymbol{x}_{t1}, \ldots, \boldsymbol{x}_{tq})$ are produced. Applying the same Kalman filter to the $\boldsymbol{y}_t$'s and the $\boldsymbol{x}_{tk}$'s means that for a given set of parameters $\boldsymbol{\Psi}$, the recursions for $\boldsymbol{P}_{t|t-1}$, $\boldsymbol{P}_{t|t}$ and $\boldsymbol{F}_t$ are run only once rather than $q+1$ times. The GLS estimator of $\boldsymbol{\beta}$ becomes

$$\hat{\boldsymbol{\beta}}_{GLS} = \left[\sum_{t=1}^{n} \boldsymbol{X}_t^{*T}\boldsymbol{F}_t^{-1}\boldsymbol{X}_t^*\right]^{-1} \sum_{t=1}^{n} \boldsymbol{X}_t^{*T}\boldsymbol{F}_t^{-1}\boldsymbol{y}_t^*. \tag{5.23}$$

For small networks this is computationally much more efficient, for it consists of inverting low dimensional matrices. The actual innovations $\boldsymbol{v}_t$ can be rewritten as $\boldsymbol{v}_t = \boldsymbol{y}_t^* - \boldsymbol{X}_t^* \hat{\boldsymbol{\beta}}_{GLS}$. Because all the disturbances and the initial state vector are multivariate normally distributed, the Kalman filter can also be used to decompose the log-likelihood as $-2 \log L_{Y_N}(\boldsymbol{\Phi}) \sim \sum_{t=1}^{n} |\boldsymbol{F}_t| + \sum_{t=1}^{n} \boldsymbol{v}_t^T \boldsymbol{F}_t^{-1} \boldsymbol{v}_t$ (Harvey, 1989). This allows the use of classical numerical algorithms for direct maximisation of the likelihood. In this dissertation, however, we consider an ECM algorithm for this purpose. The derivation of the algorithm is presented in the next subsection.

5.3.3 The ECM algorithm using the state-space representation

The river monitoring network topology imposes a restricted parametrisation of the dependence structure of the spatial process. As Xu and Wikle (2005) showed for more general spatio-temporal dependence structures, the expectation-maximisation (EM) algorithm of Shumway and Stoffer (1982) presented in Section 4.4.2 should be modified to deal with these restrictions. Since $\boldsymbol{Q}$ and $\boldsymbol{\Phi}$ have some parameters in common, our particular parametrisation is not covered by the theory of Xu and Wikle (2005), and thus we have to adapt the EM algorithm so that it can deal with the specific restrictions induced by our spatio-temporal process. Due to the presence of the exogenous variables, we further extend the EM algorithm to an expectation conditional maximisation (ECM) algorithm. In particular, the M-step is split into a sequence of two CM-steps. First the parameters of the dependence structure $\boldsymbol{\Psi}_{\boldsymbol{\alpha}}$ are estimated given the current values of the parameter estimates of the mean model $\boldsymbol{\beta}$. Next GLS is used to obtain an estimate of $\boldsymbol{\beta}$ using the updated values of $\boldsymbol{\Psi}_{\boldsymbol{\alpha}}$. Before each step of the ECM algorithm is discussed in detail, we first present an overview of the different steps that are used in the algorithm. Let $l_c(\boldsymbol{\Psi}) = \log L_{Y_N, S_N}(\boldsymbol{\Psi})$ denotes the *completed log-likelihood* given by the joint log-likelihood of $\boldsymbol{Y}_N$ and $\boldsymbol{S}_N$. Due to the unobservable state process $\boldsymbol{S}$, this likelihood can not be calculated. In the k^{th} iteration, the ECM algorithm starts with an **E-step** to calculate the conditional expectation of the completed log-likelihood given the observations $\boldsymbol{Y}_N$ and given the current values of the parameters $\boldsymbol{\Psi}^k$. In the succeeding **CM-steps**, new parameter values are calculated that maximise the conditional expected log-likelihood $\mathrm{E}\left(l_c(\boldsymbol{\Psi}) | \boldsymbol{Y}_N, \boldsymbol{\Psi}^k\right)$. The ECM algorithm can be summarised as follows ($k = 0, 1, \ldots$),

1. Choose initial estimates: $\boldsymbol{\Psi}^0$
2. **E-step**: Calculate $Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}_{\alpha}^k, \boldsymbol{\beta}^k) = \mathrm{E}\left(l_c(\boldsymbol{\Psi}) | \boldsymbol{Y}_N, \boldsymbol{\Psi}^k\right)$

3. **CM-step 1**: Find the covariance parameters $\boldsymbol{\Psi}_\alpha^{k+1}$ that maximise $Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}_\alpha^k, \boldsymbol{\beta}^k)$
4. **CM-step 2**: Find $\boldsymbol{\beta}^{k+1}$ that maximises $Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}_\alpha^{k+1}, \boldsymbol{\beta}^k)$
5. Repeat steps 2-4 until convergence

For our particular state-space model, both the E- and the CM-steps can be simplified. Details are provided in the next paragraphs.

E-step Consider first the factorisation

$$\begin{aligned} Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}_\alpha^k, \boldsymbol{\beta}^k) &= \mathrm{E}\left(l_c(\boldsymbol{\Psi})|\boldsymbol{Y}_N, \boldsymbol{\Psi}^k\right) \\ &= \mathrm{E}\left(\log L_{S_0}(\boldsymbol{\Psi})|\boldsymbol{Y}_N, \boldsymbol{\Psi}^k\right) + \mathrm{E}\left(\log L_{S_N|S_0}(\boldsymbol{\Psi})|Y_N, \boldsymbol{\Psi}^k\right) \\ &\quad + \mathrm{E}\left(\log L_{Y_N|S_N}(\boldsymbol{\Psi})|\boldsymbol{Y}_N, \boldsymbol{\Psi}^k\right). \end{aligned} \tag{5.24}$$

Neglecting the parameter independent term, we find

$$\begin{aligned} Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}_\alpha^k, \boldsymbol{\beta}^k) \sim & -\frac{1}{2}\,\mathrm{E}\left(\log|\boldsymbol{\Sigma}_{S_0}| + \boldsymbol{S}_0^T\boldsymbol{\Sigma}_{S_0}^{-1}\boldsymbol{S}_0|\boldsymbol{Y}_N, \boldsymbol{\Psi}^k\right) \\ & -\frac{1}{2}\,\mathrm{E}\Bigg(n\log|\boldsymbol{\Sigma}_\eta| + \sum_{t=1}^{n}(\boldsymbol{S}_t - \boldsymbol{A}\boldsymbol{S}_t - \boldsymbol{B}\boldsymbol{S}_{t-1})^T\boldsymbol{\Sigma}_\eta^{-1} \\ & \qquad (\boldsymbol{S}_t - \boldsymbol{A}\boldsymbol{S}_t - \boldsymbol{B}\boldsymbol{S}_{t-1})|\boldsymbol{Y}_N, \boldsymbol{\Psi}^k\Bigg) \\ & -\frac{1}{2}\,\mathrm{E}\Bigg(n\log|\boldsymbol{\Sigma}_\epsilon| + \sum_{t=1}^{n}(\boldsymbol{y}_t - \boldsymbol{X}_t\boldsymbol{\beta} - \boldsymbol{S}_t)^T\boldsymbol{\Sigma}_\epsilon^{-1} \\ & \qquad (\boldsymbol{y}_t - \boldsymbol{X}_t\boldsymbol{\beta} - \boldsymbol{S}_t)|\boldsymbol{Y}_N, \boldsymbol{\Psi}^k\Bigg). \end{aligned} \tag{5.25}$$

Because the distributions of both $\boldsymbol{Y}$ and $\boldsymbol{S}$ belong to the regular exponential family, the calculation of Equation (5.24) can be reduced to the replacement of the sufficient statistics by their conditional expectations into $l_c(\boldsymbol{\Psi})$ (McLachlan and Krishnan, 1997). For conditioning on $\boldsymbol{Y}_N$, only the expectations of the sufficient statistics based on $\boldsymbol{S}_t$ and on $\boldsymbol{S}_{t-1}$ have to be determined. In particular,

$$E[\boldsymbol{S}_t|\boldsymbol{Y}_N] = \boldsymbol{\lambda}_{t|n} \tag{5.26}$$

$$E[\boldsymbol{S}_t\boldsymbol{S}_t|\boldsymbol{Y}_N] = \boldsymbol{P}_{t|n} + \boldsymbol{\lambda}_{t|n}\boldsymbol{\lambda}_{t|n}^T \tag{5.27}$$

$$E[\boldsymbol{S}_t\boldsymbol{S}_{t-1}|\boldsymbol{Y}_N] = \boldsymbol{P}_{t,t-1|n} + \boldsymbol{\lambda}_{t|n}\boldsymbol{\lambda}_{t-1|n}^T. \tag{5.28}$$

They are calculated using the Kalman filter and smoother recursions (Equations (5.13)-(5.22)).

CM-step 1 In this step the covariance parameters $\boldsymbol{\Psi}_\alpha$ are estimated conditional on $\boldsymbol{\beta}^k$. Because $\boldsymbol{\Sigma}_\eta$ is diagonal, the second term in Equation (5.25) can be further factorized,

$$\begin{aligned}&\mathrm{E}\left(\log L_{\boldsymbol{S}_N}(\boldsymbol{\Psi})|\boldsymbol{Y}_N, \boldsymbol{\Psi}^k\right) \sim \\ &-\frac{1}{2}\sum_{i=1}^{p}\mathrm{E}\left(n\log\sigma^2_{\eta_i} + \frac{1}{\sigma^2_{\eta_i}}\sum_{t=1}^{n}(S_{it} - \boldsymbol{A}_{i.}^{[a_i]}\boldsymbol{S}_t^{[a_i]} - \boldsymbol{B}_{i.}^{[b_i]}\boldsymbol{S}_{t-1}^{[b_i]})^2|\boldsymbol{Y}_N, \boldsymbol{\Psi}^k\right),\end{aligned} \tag{5.29}$$

where $[a_i]$ represents the index set $(j_1, \ldots, j_q)$ corresponding to the non-zero elements of the i^{th} row of $\boldsymbol{A}$, and $[b_i]$ is a similar set for the i^{th} row of matrix $\boldsymbol{B}$. The elements of $[a_i]$ can be derived from the DAG. The index set $[b_i]$ expands the index set $[a_i]$ by sampling location i under consideration, because the state variable S_{it} does not only depend on S_{it-1} but also on its parents in the DAG on time $t-1$. Thus, $\boldsymbol{A}_{i.}^{[a_i]}$ and $\boldsymbol{B}_{i.}^{[b_i]}$ are the non-zero elements of the i^{th} row of $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively, $\boldsymbol{S}_t^{[a_i]} = (S_{j_1t}, \ldots, S_{j_qt})^T$ and $\boldsymbol{S}_{t-1}^{[b_i]} = (S_{j_1t-1}, \ldots S_{j_qt-1}, S_{it-1})^T$.

Furthermore, from Equation (5.25) it is observed that the $\mathrm{E}\left(.|\boldsymbol{Y}_N, \boldsymbol{\Psi}^k\right)$ operation cannot introduce any other parameters contained in $\boldsymbol{\Psi}_{\boldsymbol{\alpha}}$, and the first and third term at the right hand side of Equation (5.25) do not contain any parameters from $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{\Sigma}_\eta$. Therefore the ECM-approach considered here implies an estimation orthogonality and each term in Equation (5.25) may be maximised separately. We use a result of Ansley and Kohn (1985) for the maximisation of Equation (5.29). In particular, they showed that the variance components can be concentrated out of the likelihood, resulting in

$$\mathrm{E}_{\boldsymbol{\Psi}^k}\left(\log L_{\boldsymbol{S}_N}(\boldsymbol{\Psi})|\boldsymbol{Y}\right) \sim -\frac{1}{2}\sum_{i=1}^{p}\mathrm{E}_{\boldsymbol{\Psi}^k}\left(n\log\frac{\mathrm{RSS}_i}{n} + n|\boldsymbol{Y}\right) \tag{5.30}$$

where $\mathrm{RSS}_i = \sum_{t=1}^{n} I_{it}^2$, and $I_{it} = S_{it} - \boldsymbol{A}_{i.}^{[a_i]}\boldsymbol{S}_t^{[a_i]} - \boldsymbol{B}_{i.}^{[b_i]}\boldsymbol{S}_{t-1}^{[b_i]}$. Sampling locations S_i for which $[a_i] = \phi$ imply that $\boldsymbol{A}_{i.}^{[a_i]} = \phi$, $\boldsymbol{S}_t^{[a_i]} = \phi$, $[b_i] = i$, and also imply that the vectors $\boldsymbol{B}^{[b_i]}$ and $\boldsymbol{S}_{t-1}^{[b_i]}$ reduce to the diagonal element $\boldsymbol{B}_{ii}$ and the scalar S_{it-1}, respectively. The maximum likelihood estimator (MLE) of $\sigma^2_{\eta_i}$ is given by $\frac{\mathrm{RSS}_i}{N}$. Note that maximising Equation (5.30) in the CM step is equivalent to the minimisation of RSS_i. The solution is obtained by equating the partial

derivatives $\frac{\partial \text{RSS}_i}{\partial \boldsymbol{A}_{i.}^{[a_i]}}$ and $\frac{\partial \text{RSS}_i}{\partial \boldsymbol{B}_{i.}^{[b_i]}}$ to zero after the sufficient statistics are replaced by their conditional expectations as calculated in the E-step (Equations (5.26)-(5.28)). In the E-step, these sufficient statistics were calculated conditionally on the parameters $\boldsymbol{\Psi}^k$ of the previous iteration. Thus this replacement does not introduce any additional $\boldsymbol{A}_{i.}^{[a_i]}$, $\boldsymbol{B}_{i.}^{[b_i]}$ nor $\sigma_{\eta_i}^2$. We propose to solve $\frac{\partial \text{RSS}_i}{\partial \boldsymbol{A}_{i.}^{[a_i]}} = 0$ and $\frac{\partial \text{RSS}_i}{\partial \boldsymbol{B}_{i.}^{[b_i]}} = 0$ first, and subsequently replace the sufficient statistics in the expressions for the MLE's of $\boldsymbol{A}_{i.}^{[a_i]}$ and $\boldsymbol{B}_{i.}^{[b_i]}$, eventually leading to ${\boldsymbol{A}_{i.}^{[a_i]}}^{k+1}$ and ${\boldsymbol{B}_{i.}^{[b_i]}}^{k+1}$. Some matrix algebra gives

$$
\begin{aligned}
{\boldsymbol{B}_{i.}^{[b_i]}}^{k+1} = & \left[\sum_{t=1}^{n} S_{it} {\boldsymbol{S}_{t-1}^{[b_i]}}^T - \left(\sum_{t=1}^{n} S_{it} \boldsymbol{S}_t^{[a_i]T} \right) \left(\sum_{t=1}^{n} \boldsymbol{S}_t^{[a_i]} \boldsymbol{S}_t^{[a_i]T} \right)^{-1} \right. \\
& \left. \left(\sum_{t=1}^{n} \boldsymbol{S}_t^{[a_i]} {\boldsymbol{S}_{t-1}^{[b_i]}}^T \right) \right] \times \\
& \left[\sum_{t=1}^{n} \boldsymbol{S}_{t-1}^{[b_i]} {\boldsymbol{S}_{t-1}^{[b_i]}}^T - \left(\sum_{t=1}^{n} \boldsymbol{S}_{t-1}^{[b_i]} \boldsymbol{S}_t^{[a_i]T} \right) \left(\sum_{t=1}^{n} \boldsymbol{S}_t^{[a_i]} \boldsymbol{S}_t^{[a_i]T} \right)^{-1} \right. \\
& \left. \left(\sum_{t=1}^{n} \boldsymbol{S}_t^{[a_i]} {\boldsymbol{S}_{t-1}^{[b_i]}}^T \right) \right]^{-1}
\end{aligned}
\tag{5.31}
$$

$$
\begin{aligned}
{\boldsymbol{A}_{i.}^{[a_i]}}^{k+1} = & \left(\sum_{t=1}^{n} S_{it} \boldsymbol{S}_t^{[a_i]T} - {\boldsymbol{B}^{[b_i]}}^{k+1} \sum_{t=1}^{n} \boldsymbol{S}_{t-1}^{[b_i]} \boldsymbol{S}_t^{[a_i]T} \right) \\
& \left(\sum_{t=1}^{n} \boldsymbol{S}_t^{[a_i]} \boldsymbol{S}_t^{[a_i]T} \right)^{-1},
\end{aligned}
\tag{5.32}
$$

in which all the sufficient statistics have to be replaced by their corresponding conditional expectations given in Equations (5.26)-(5.28). The derivation of these results can be found in the appendix of this chapter. In case sampling location Si has no parents according to the DAG, $[a_i] = \phi$ and $\boldsymbol{A}_{i.}^{[a_i]} = \phi$. This causes the terms containing $\boldsymbol{S}_t^{[a_i]}$ in Equation (5.31) to disappear.

After ${\boldsymbol{B}_{i.}^{[b_i]}}^{k+1}$ and ${\boldsymbol{A}_{[i.]}^{[a_i]}}^{k+1}$ are computed, the estimate of $\sigma_{\eta_i}^2$ is calculated as

$$
{\sigma_{\eta_i}^2}^{k+1} = \frac{\text{RSS}_i^{k+1}}{n},
\tag{5.33}
$$

in which the sufficient statistics are replaced, as before.

Shumway and Stoffer (2006) showed that the first term in Equation (5.25) is optimised by

$$\hat{\boldsymbol{\Sigma}}_{S_0} = \boldsymbol{P}_{0|n}. \tag{5.34}$$

The maximisation of the third term in the right hand side of Equation (5.25) gives

$$\begin{aligned}\boldsymbol{\Sigma}_{\epsilon}^{k+1} &= \frac{1}{n}\sum_{t=1}^{n}(\boldsymbol{y}_t' - \boldsymbol{S}_t)(\boldsymbol{y}_t' - \boldsymbol{S}_t)^T \\ &= \frac{1}{n}\left[\sum_{t=1}^{n}(\boldsymbol{y}_t'\boldsymbol{y}_t'^T) - \sum_{t=1}^{n}(\boldsymbol{y}_t'\boldsymbol{S}_t^T) - \sum_{t=1}^{n}(\boldsymbol{S}_t\boldsymbol{y}_t'^T) + \sum_{t=1}^{n}(\boldsymbol{S}_t\boldsymbol{S}_t^T)\right], \end{aligned} \tag{5.35}$$

where $\boldsymbol{y}_t' = \boldsymbol{y}_t - \boldsymbol{X}_t\boldsymbol{\beta}^k$ and in which the sufficient statistics are again to be replaced by their conditional expectations given in Equations (5.26)-(5.28).

CM-step 2 In this CM step, the parameters of the mean model are estimated by means of the Kalman filter resulting in the feasible generalised least squares estimate $\hat{\beta}(\hat{\boldsymbol{\Psi}}_{\alpha})$. The estimation is conditional on $\boldsymbol{\Psi}_{\alpha}^{k+1}$. In Section 5.3.2 it was shown that the Kalman filter can be used to perform the GLS and its solution is given in Equation (5.23). For the calculation of Equation (5.23), the forward recursions displayed in Equations (5.14), (5.16) and (5.17) have to be calculated and their results can be used in the next iteration step.

To conclude, we summarise the whole algorithm ($k = 0, 1, \ldots$):
(1) Choose initial estimates: $\boldsymbol{\Psi}^0$
(2) **E-step**: Calculate the expected sufficient statistics using Equations (5.26)-(5.28)
(3) **CM-step 1**: Estimate the covariance parameters $\boldsymbol{\Psi}_{\alpha}^{k+1}$ using Equations (5.31) - (5.35)
(4) **CM-step 2**: Use the covariance parameters $\boldsymbol{\Psi}_{\alpha}^{k+1}$ to calculate the FGLS estimator $\hat{\boldsymbol{\beta}}^{k+1}$ by using Equation (5.23)
(5) Repeat steps 2-4 until convergence

To obtain initial parameter values $\boldsymbol{\Psi}^0$ we suggest to perform an ordinary least squares (OLS) regression. As an initial estimate for $\boldsymbol{\beta}^0$ the parameter vector obtained by OLS can be used. The residuals, say $\boldsymbol{e}_t$'s, of the OLS regression can be used to provide the initial values for the parameters in $\boldsymbol{A}$ and $\boldsymbol{B}$, by fitting the

following regression model: $\boldsymbol{e}_t = \boldsymbol{A}^0\boldsymbol{e}_t + \boldsymbol{B}^0\boldsymbol{e}_{t-1} + \boldsymbol{e}'_t$. $\boldsymbol{\Sigma}^0_\epsilon$ can be obtained from the $\boldsymbol{e}'_t$'s by the method of moments and finally the parameters in $\boldsymbol{\Sigma}_\eta$ should be set on an arbitrary value.

5.3.4 ECM algorithm using the SEM representation

We now use the structural equation model (SEM) representation presented in Equations (5.6) and (5.7) to derive an ECM algorithm. This representation implies some adaptations of the ECM algorithm presented in Section 5.3.3. Instead of using the Kalman filter and smoother to calculate the expectations of the sufficient statistics in the E-step, the SEM representation has to be used. The sufficient statistics are now calculated as follows,

$$\mathrm{E}_{\boldsymbol{\Psi}^k}\left(\boldsymbol{S}_N|\boldsymbol{Y}_N\right) = \boldsymbol{\Sigma}^k_{S_N}\left(\boldsymbol{\Sigma}^k_{S_N} + \boldsymbol{\Sigma}^k_\epsilon\right)^{-1}\left(\boldsymbol{Y}_N - \boldsymbol{X}_N\boldsymbol{\beta}^k\right) \tag{5.36}$$

$$\mathrm{E}_{\boldsymbol{\Psi}^k}\left(\boldsymbol{S}_N\boldsymbol{S}_N^T|\boldsymbol{Y}_N\right) = \boldsymbol{Z}^k{\boldsymbol{Z}^k}^T + \boldsymbol{\Sigma}^k_{S_N} - \boldsymbol{\Sigma}^k_{S_N}\left(\boldsymbol{\Sigma}^k_{S_N} + \boldsymbol{\Sigma}^k_\epsilon\right)^{-1}\boldsymbol{\Sigma}^k_{S_N}, \tag{5.37}$$

where $\boldsymbol{\Sigma}^k_{S_N} = {\boldsymbol{C}^k}^{-1}\boldsymbol{\Sigma}^k_\zeta{\boldsymbol{C}^k}^{-T}$ and $\boldsymbol{Z}^k = \mathrm{E}_{\boldsymbol{\Psi}^k}\left(\boldsymbol{S}_N|\boldsymbol{Y}_N\right)$. In the CM step-1, the calculation of the estimate of the covariance parameters $\boldsymbol{\Psi}^{k+1}_\alpha$ remains unaltered. These estimates are then used to construct $\boldsymbol{C}^{k+1}$, $\boldsymbol{\Sigma}^{k+1}_\zeta$ and $\boldsymbol{\Sigma}^{k+1}_\psi$. By plugging $\boldsymbol{C}^{k+1}$, $\boldsymbol{\Sigma}^{k+1}_\zeta$ and $\boldsymbol{\Sigma}^{k+1}_\psi$ into Equation (5.8), an estimate $\boldsymbol{\Sigma}^{k+1}_{Y_N}$ is obtained. In the original CM step-2, the Kalman filter was used to perform the GLS. In the SEM approach, the GLS has to be performed explicitly by

$$\boldsymbol{\beta}^{k+1} = \left(\boldsymbol{X}_N^T(\boldsymbol{\Sigma}^{k+1}_{Y_N})^{-1}\boldsymbol{X}_N\right)^{-1}\boldsymbol{X}_N^T(\boldsymbol{\Sigma}^{k+1}_{Y_N})^{-1}\boldsymbol{Y}_N. \tag{5.38}$$

The SEM approach involves the calculations of the inverse of the matrices $\boldsymbol{C}^{k+1}$ and $\boldsymbol{\Sigma}^{k+1}_{Y_N}$. The dimensions of these matrices are $np \times np$. In our application, the number of time instants n is much larger than the number of sampling locations p. For such networks, it is computationally less attractive to calculate the inverse of $\boldsymbol{C}^{k+1}$ and $\boldsymbol{\Sigma}^{k+1}_{Y_N}$ than to calculate n times the inverse of the $p \times p$ matrices $\boldsymbol{F}_t$ and $\boldsymbol{P}_{t+1|t}$ needed to evaluate the Kalman filter and smoother. Moreover, the Kalman filter also reaches a steady state after a certain time instant, say w. This implies that for instance the matrices $\boldsymbol{P}_{t|t-1}$, $\boldsymbol{P}_{t|t}$ and $\boldsymbol{F}_t$ do not alter anymore for time instants $t > w$. Therefore we use the ECM algorithm derived for the state-space representation to perform the parameter estimation. However, for the

statistical inference procedures presented in the next section, we will make use the SEM representation to calculate an estimate of the variance-covariance matrix $\boldsymbol{\Sigma}_{Y_N}(\boldsymbol{\Psi}_\alpha)$.

5.3.5 Statistical Inference

Model selection can be performed by using a modified AIC criteria which accounts for the dependence of the data. Akaike (1973) showed that $-2l(\hat{\boldsymbol{\Psi}}) = -2\log L_{Y_N}(\hat{\boldsymbol{\Psi}})$ is a biased estimator of the exact Kullback-Leibler divergence between the true and the fitted model. The bias adjustment for this estimator is often approximated by $2df$, where df is the number of degrees of freedom used by the model. State-space models usually require a large number of observations to make this asymptotic approximation work well. Bengtsson and Cavanaugh (2006) defined an *improved* AIC criterion for state-space model selection, referred to as AICi. They suggested a Monte Carlo approximation of the bias, say $\hat{B}$. Their criterion is then defined as AICi$= -2l(\hat{\boldsymbol{\Psi}}) + \hat{B}$. Unfortunately, for our river monitoring network model this method is computationally too demanding because the maximum likelihood estimates are needed for each Monte Carlo simulation run. We therefore suggest to use the original AIC criterion for model selection. Besides model selection criteria, diagnostics are useful to check the quality of the model structure. In the state-space framework, plots of the standardised innovations $\boldsymbol{v_t F_t^{-1/2}}$ can be used as diagnostic plots for the temporal dependence structure and the mean model (Harvey, 1989).

Statistical inference on the parameters of the mean model requires the joint sampling distribution of $\hat{\boldsymbol{\beta}}$. Since we deal with a linear model for the mean and with a stationary Gaussian process sampled on regular time steps, theorem 3 of Mardia and Marshall (1984) can be applied to establish that the maximum likelihood estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Psi}}_{\boldsymbol{\alpha}}$ are consistent and asymptotically normally distributed. In particular,

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\Psi}_{\boldsymbol{\alpha}}) \xrightarrow{d} MVN(\boldsymbol{\beta}, (\boldsymbol{X}^T\boldsymbol{\Sigma}_{Y_N}^{-1}\boldsymbol{X})^{-1}), \tag{5.39}$$

when $n \to \infty$. Finally, since $\hat{\boldsymbol{\Psi}}_{\boldsymbol{\alpha}}$ is a consistent estimator, the variance of $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\Psi}}_{\boldsymbol{\alpha}})$ is estimated consistently by

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{Y_N}}^{\boldsymbol{-1}}(\hat{\boldsymbol{\Psi}}_{\boldsymbol{\alpha}})\boldsymbol{X})^{-1}. \tag{5.40}$$

To obtain the standard errors of the parameter estimators of the dependence structure, it is possible to evaluate the Hessian matrix after convergence. Another possi-

bility is to perturbate the likelihood function and to apply numerical differentiation to find the observed Fisher information matrix (e.g. Harvey, 1989 and Shumway and Stoffer, 2006). In the case study we apply the latter approach.

5.4 Case study

One of the key actions of the WFD is the design and maintenance of water quality monitoring networks. In Flanders, several water quality monitoring networks are maintained by the VMM. An example is the physico-chemical monitoring network of the surface waters. The VMM reports on the water quality on an annual basis. In their annual reports they use yearly averages of the water quality. It would be very informative if they could use a statistical tool to compare the mean of the current year with that of the general mean and with the means of recent years. Preferably, such a tool would incorporate statistical tests on the level of the individual sampling locations as well as on a more regional scale. In this case study we will use our spatio-temporal model for river monitoring networks for this purpose. The data of 5 sampling locations of the physico-chemical monitoring network of the Flemish surface waters are used. They are located along 2 joining reaches in the Yzer catchment. Their DAG and location in the catchment is indicated on the map in Figure 5.2. Sampling locations S1, S2, S4 and S5 are located on the Yzer while sampling location S3 is located on a joining creek. Every sampling location is monitored on a monthly basis. Nitrate data between 1990 and 2003 are available. Hence, the 5 sampling locations are monitored on 168 time instants and the entire dataset consists of 840 observations.

The observations are taken at time intervals that are much larger than the timescale of the water flow. Therefore we make the assumption that the matrix $\boldsymbol{B}$, used to describe the temporal correlation, is diagonal. Hence, we only model the temporal autocorrelations for a particular state S_{it} at time t and not the spatio-temporal cross correlations between S_{it} and its parents in the DAG $\boldsymbol{S}_{t-1}^{[a_i]}$ at time $t-1$. This leads to the reduction of the parent set $[b_i]$ to $[b_i] = i$, containing only the current sampling location. Instead of assessing the annual mean at the level of individual sampling locations, we aim to perform an assessment on a more regional scale and therefore we will calculate the annual mean based on all 5 sampling locations. In this case study two questions will be addressed. On the one hand we want to test whether this "regional" annual mean for nitrate in 2003 is different from the general mean. On the other hand we also want to test whether the "regional" annual mean in 2003

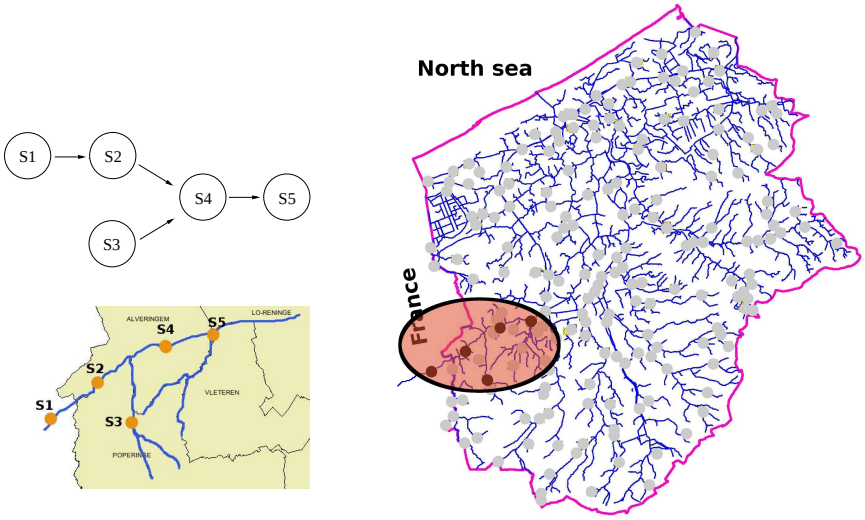

Figure 5.2: Top Left: Directed Acyclic Graph (DAG) of the sampling locations along the 2 river reaches. Bottom Left: Map of the river reaches considered in this case study. Locations S1, S2, S4 and S5 are located on the Yzer river while location S3 is located on a joining creek. Sampling location S1 is located in France. Right: Map of the part of the Yzer catchment located in Flanders, Belgium. The sampling locations are indicated by the dots. The area considered in this study is indicated with the ellipse and the black dots are the sampling locations included in this study

is different from the "regional" mean of the two most recent years (2001 and 2002).

The annual mean is modelled by a factor with one level for each year. Seasonal variation also is typically present in water quality data and the model has to account for it. This was illustrated in Chapter 1. In the introduction, the seasonal variation is illustrated in Figure 1.6 where nitrate data of all years is plotted in function of the day of the year. A common approach to deal with this seasonal variation is to include sinusoidal functions of fixed periods to describe the seasonal cycle within a year (e.g. Hirst, 1998, Cai and Tiwari, 2000, McMullan et al., 2003 and McMullan, 2004). A common function which is used for this purpose is $\alpha\cos(2\pi(t/P)+\theta)$, where P is the period which is taken to be 1 year, α is the amplitude of the seasonal trend and θ is a parameter to allow for a phase shift. This function however is nonlinear in the parameter θ. But, it can be expressed

in a linear form by using a standard trigonometric expansion. With a period P of one year we get $\gamma_1 \sin(2\pi t/12) + \gamma_2 \cos(2\pi t/12))$. To enable the assessment on a regional scale, the interaction between sampling location and year should be neglected. This assumption will be checked in the analysis of the standardised innovations. The models that are considered are given in Table 5.1, where μ is the general mean, α_i is the effect for the i^{th} sampling location, $\beta_{\lfloor t/12 \rfloor}$ is the effect of the $\lfloor t/12 \rfloor^{th}$ year, γ_k are the parameters for the seasonal component modelled by Fourier terms, and the $(\alpha\gamma)_{ik}$ and $(\beta\gamma)_{\lfloor t/12 \rfloor k}$ are the parameters for the sampling location-season and year-season interactions respectively. The interactions between year and season are included because the seasonal variation of water quality variables often changes from year to year (e.g. Hirst, 1998; McMullan et al., 2003; McMullan, 2004). The models are estimated by using the ECM-algorithm from Section 5.3.3.

Model III has the lowest AIC and is selected. The results of the GLS estimation of the mean model are visualised in Figure 5.3. The model indicates that a seasonal pattern changes over time. The amplitude drops from 1999 on. From Figure 5.3 it also seems that the annual mean is decreasing in the most recent years. The marginal mean at the joining creek (S3) seems to be overestimated in the more recent years and this deviation increases as time evolves. The parameter estimates of the mean model, $\hat{\boldsymbol{\beta}}$, are presented in Table 5.2. Along with the parameter value, the standard deviation and a p-value are given. This two-sided p-value corresponds to the null-hypothesis that the particular parameter value is equal to zero.

Table 5.1: Mean models to assess the evolution in the "regional" annual mean

Model	E (y_{it})	AIC
I	$\mu + \alpha_i + \beta_{\lfloor t/12 \rfloor} + \gamma_1 \sin(2\pi t/12) + \gamma_2 \cos(2\pi t/12)$	5100.5
II	$\mu + \alpha_i + \beta_{\lfloor t/12 \rfloor} + \gamma_1 \sin(2\pi t/12) + \gamma_2 \cos(2\pi t/12)$ $+(\alpha\gamma)_{i1} \sin(2\pi t/12) + (\alpha\gamma)_{i2} \cos(2\pi t/12)$	5096.4
III	$\mu + \alpha_i + \beta_{\lfloor t/12 \rfloor} + \gamma_1 \sin(2\pi t/12) + \gamma_2 \cos(2\pi t/12)$ $+(\beta\gamma)_{\lfloor t/12 \rfloor 1} \sin(2\pi t/12) + (\beta\gamma)_{\lfloor t/12 \rfloor 2} \cos(2\pi t/12)$	5063.9
IV	$\mu + \alpha_i + \beta_{\lfloor t/12 \rfloor} + \gamma_1 \sin(2\pi t/12) + \gamma_2 \cos(2\pi t/12)$ $+(\alpha\gamma)_{i1} \sin(2\pi t/12) + (\alpha\gamma)_{i2} \cos(2\pi t/12)$ $+(\beta\gamma)_{\lfloor t/12 \rfloor 1} \sin(2\pi t/12) + (\beta\gamma)_{\lfloor t/12 \rfloor 2} \cos(2\pi t/12)$	5179.1

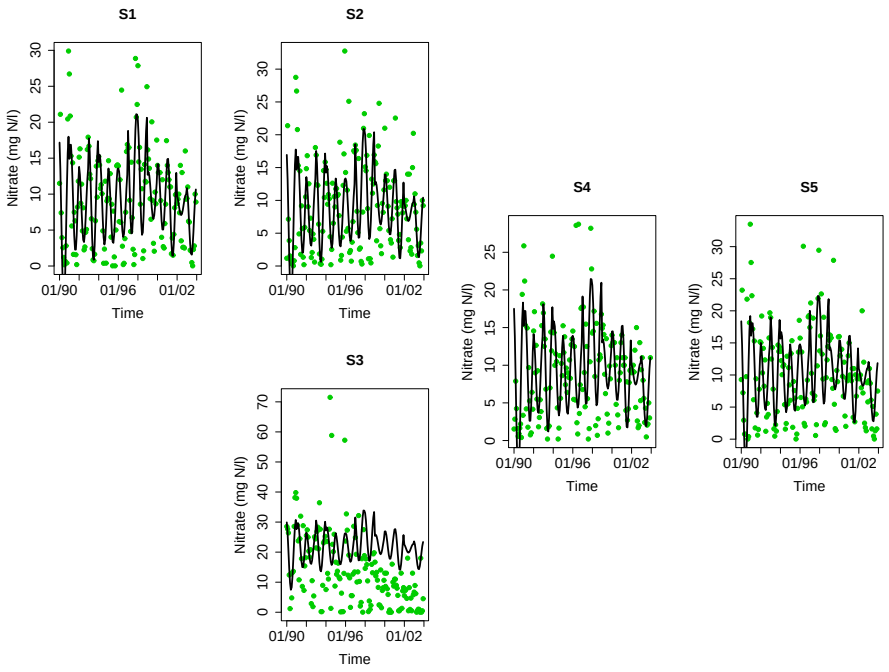

Figure 5.3: Evolution of the water quality at five sampling locations of the river Yzer. Sampling locations S1, S2, S4, S5 are located on the main river, sampling location S3 is located on a tributary which drains into the Yzer between S2 and S4. The line indicates the model fit according to Model III

Table 5.2: The parameter estimates of the mean model of Model III

parameter	value	std error	p-value
μ	11.96	0.72	$<$0.0001
β_{14}	$-$2.92	0.42	$<$0.0001
β_{13}	$-$0.60	0.34	0.0790
β_{12}	$-$0.99	0.34	0.0035
β_{11}	0.49	0.34	0.1500
β_{10}	0.68	0.34	0.0460
β_{9}	3.35	0.34	$<$0.0001

Table 5.2 – Continued

parameter	value	std error	p-value
β_8	3.70	0.34	<0.0001
β_7	−0.34	0.34	0.3100
β_6	−0.75	0.34	0.0260
β_5	0.09	0.34	0.7800
β_4	0.20	0.34	0.5600
β_3	−0.56	0.34	0.0960
β_2	0.45	0.34	0.1800
β_1	−2.79	0.24	<0.0001
α_5	−1.62	0.46	0.0005
α_4	−2.50	0.49	<0.0001
α_2	−3.04	0.59	<0.0001
α_1	−2.82	0.58	<0.0001
α_3	9.98	1.82	<0.0001
γ_1	2.55	0.25	<0.0001
γ_2	4.99	0.25	<0.0001
$(\beta\gamma)_{2,1}$	−0.99	0.83	0.2300
$(\beta\gamma)_{3,1}$	−2.52	0.78	0.0013
$(\beta\gamma)_{4,1}$	2.44	0.78	0.0018
$(\beta\gamma)_{5,1}$	0.79	0.78	0.3100
$(\beta\gamma)_{6,1}$	0.22	0.78	0.7800
$(\beta\gamma)_{7,1}$	−1.06	0.78	0.1700
$(\beta\gamma)_{8,1}$	−4.81	0.78	<0.0001
$(\beta\gamma)_{9,1}$	0.24	0.78	0.7600
$(\beta\gamma)_{10,1}$	−0.81	0.78	0.3000
$(\beta\gamma)_{11,1}$	3.00	0.78	0.0001
$(\beta\gamma)_{12,1}$	0.40	0.78	0.6100
$(\beta\gamma)_{13,1}$	−0.60	0.78	0.4400
$(\beta\gamma)_{14,1}$	4.82	0.78	<0.0001
$(\beta\gamma)_{2,2}$	−0.59	1.06	0.5800
$(\beta\gamma)_{3,2}$	−3.61	0.98	0.0002
$(\beta\gamma)_{4,2}$	−0.21	0.98	0.8300
$(\beta\gamma)_{5,2}$	−1.68	0.98	0.8400
$(\beta\gamma)_{6,2}$	−3.29	0.98	0.0008
$(\beta\gamma)_{7,2}$	3.06	0.98	0.0017
$(\beta\gamma)_{8,2}$	3.17	0.98	0.0011
$(\beta\gamma)_{9,2}$	−1.13	0.98	0.2500
$(\beta\gamma)_{10,2}$	−0.05	0.98	0.9600

Table 5.2 – Continued

parameter	value	std error	p-value
$(\beta\gamma)_{11,2}$	-2.21	0.98	0.0240
$(\beta\gamma)_{12,2}$	3.08	0.98	0.0016
$(\beta\gamma)_{13,2}$	-0.07	0.98	0.9400
$(\beta\gamma)_{14,2}$	-3.04	0.98	0.0019

As mentioned earlier, the effect of the year is modelled by the use of a factor with one level for each of the 14 years. The size of each of these annual effect is modelled by the parameters $(\beta_1, \ldots, \beta_{14})$. According to the p-values the mean nitrate level of a number of years is not significantly different from the general mean (e.g. for β_2, β_4, β_5, β_7, β_{11} and β_{13}). For the seasonal-year interactions $(\beta\alpha)_{j,k}$ also a number of non-significant parameters occur. Note, however that the seasonal effect is coded by two parameters to provide an amplitude and a phase shift. Here, this is done by the use of a sine and a cosine term. Hence, for a particular year j there is a season-year interaction as soon as one of the parameters $(\beta\alpha)_{j,1}$ or $(\beta\alpha)_{j,2}$ is different from zero. The non-significant parameters are not eliminated from the model because other parameters of the main and the interaction effect are (highly) significant. Moreover, the conclusion that certain parameters are non-significant is a weak conclusion as we do not have any information about the power of the tests.

For the spatio-temporal parameters in $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{\Sigma_\eta}$ and $\boldsymbol{\Sigma_\epsilon}$ the following estimates are obtained (standard error between brackets)

$$\hat{\boldsymbol{A}} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0.78(0.11) & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1.31(0.21) & 0.10(0.05) & 0 & 0 \\ 0 & 0 & 0 & 0.63(0.11) & 0 \end{bmatrix},$$

$$\hat{\boldsymbol{B}} = \begin{bmatrix} -0.14(0.08) & 0 & 0 & 0 & 0 \\ 0 & 0.72(0.07) & 0 & 0 & 0 \\ 0 & 0 & 1.01(0.01) & 0 & 0 \\ 0 & 0 & 0 & -0.21\ (0.11) & 0 \\ 0 & 0 & 0 & 0 & 0.35(0.11) \end{bmatrix},$$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}} = \begin{bmatrix} 11.1(1.8) & 0 & 0 & 0 & 0 \\ 0 & 0.2\ 10^{-6}(0.1\ 10^{-3}) & 0 & 0 & 0 \\ 0 & 0 & 0.31(0.3) & 0 & 0 \\ 0 & 0 & 0 & 3.8\ 10^{-6}(0.9\ 10^{-3}) & 0 \\ 0 & 0 & 0 & 0 & 0.4\ 10^{-6}(0.1\ 10^{-3}) \end{bmatrix}$$

and

$$\hat{\boldsymbol{\Sigma}}_{\epsilon} = \begin{bmatrix} 7.5(0.5) & & & & \\ 8.3(1.3) & 23.9(2.8) & & & \\ 4.2(2.3 & 13.6(3.7) & 89.8(10) & & \\ 2.9(1.0) & 7.3(2.1) & 6.0(3.2) & 16.6(2.7) & \\ 13.7(0.01) & 17.2(1.7) & 13.3(3.7) & 11.8(1.8) & 27.8(0.9) \end{bmatrix}.$$

Note that the estimate of $\boldsymbol{B}$ at S3 is larger than 1. This provokes an eigenvalue of the transition matrix $\boldsymbol{\Phi}$ that is larger than 1. Hence the estimated state-space model is not stationary.

As mentioned before, the model quality has to be checked and in this work this is done by the use of an assessment of the standardised innovations. These innovations should be independent which can be assessed by a plot of the autocorrelation function (ACF). The ACF plot of the original series is shown in Figure 5.4. From these plots, the correlation in the original nitrate measurements is obvious. Moreover, they also indicate the presence of seasonal correlation. The ACF of the standardised innovations are shown in Figure 5.5. The model succeeds well in reducing a considerable amount of the serial correlation present in the original series. A joint test of significance of the first i autocorrelation coefficients is provided by the Ljung-Box portmanteau test (Ljung and Box, 1978). The p-values for the Ljung-Box portmanteau test of the autocorrelation coefficients of the standardised innovations are given in Table 5.3. Significant p-values appeared to be present at S1. This is due to the negative autocorrelation at lag 2. An ACF plot at S1 up to lag 100 is provided in Figure 5.6 and it can be seen that only 3 large autocorrelations occur during the first 100 lags. Based on the ACF-plots the AR(1) seems to be sufficient to model the temporal correlation.

The quality of the mean model is checked in Figure 5.7 showing the standardised innovations with respect to time. Friedman's supersmoother is added to each plot to study the residual pattern present in the standardised innovations (Friedman, 1984). From Figure 5.7 it can be seen that the smoothers remain close to zero, suggesting that the model quality is good. For S2, S3, S4 and S5 the smoothers give larger predictions near the boundaries. This is probably due to the combination of a boundary effect of the smoother, large nitrate values measured in the

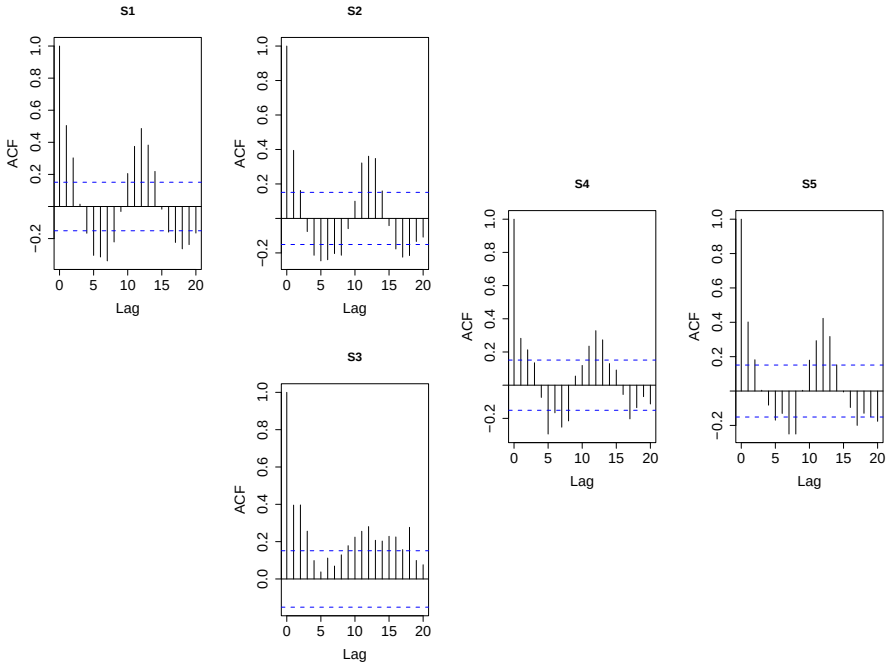

Figure 5.4: Autocorrelation plots of original nitrate series at the different sampling locations

Table 5.3: p-values for the Ljung-Box portmanteau test of the autocorrelation coefficients of the standardised innovations for the first 5 lags

Lag	S1	S2	S3	S4	S5
1	0.78	0.87	0.63	0.90	0.11
2	0.95	0.97	0.10	0.96	0.27
3	0.20	0.99	0.17	0.74	0.37
4	0.05	1.00	0.14	0.46	0.52
5	0.04	0.99	0.06	0.52	0.58

beginning of the time series and the Kalman filter which might not have reached steady state yet. Figure 5.3 indicated a systematic deviation of the estimated mean at sampling location S3. In Figure 5.7, however, the smoother only suggests a small deviation in the standardised innovations at S3. Hence, the systematic deviation in the marginal mean at S3 is modelled by the temporal dependence structure.

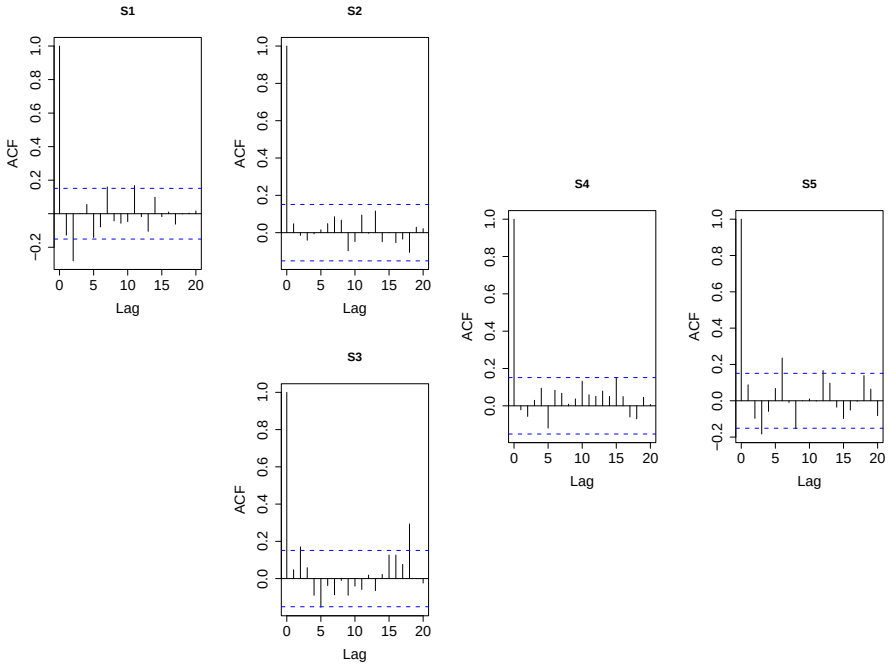

Figure 5.5: Autocorrelation plot of the standardised innovations of Model III

This is reflected by the estimate of the AR(1) coefficient for S3 which is slightly larger than 1 ($b_{3,3} = 1.01$). The deviation of the mean model at S3 and the non-stationary autocorrelation coefficient at S3 might be due to the assumption that there was no interaction between the year and the sampling location. However, when this interaction term would be included in the model, we can not infer on a "regional" scale. Another assumption that has to be checked is related to the distributional assumptions that were imposed. All processes were assumed to be Gaussian. Therefore, the standardised innovations should follow a standard normal distribution and we expect about 95% of standardised innovations to be in the interval $[-2, 2]$. In Figure 5.7 it can be seen that at each sampling location a number of outliers are present. The normality of the innovations is further assessed in Figure 5.8. Both the boxplot and the QQ-plot show a clear departure from normality. The boxplot indicates a considerable amount of outliers and the QQ-plot indicates that the distribution has larger tails than the normal distribution. On the other hand, from all plots it can be seen that the distribution of the standardised innovations is symmetric. To answer the research question, we need to infer on the parameters

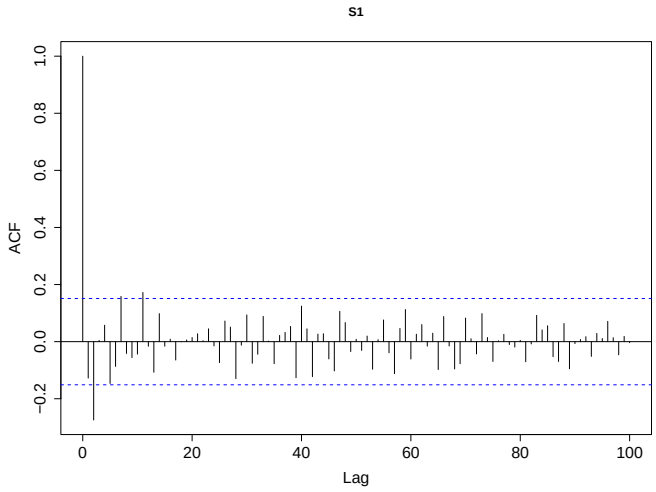

Figure 5.6: Autocorrelation plot at S1 of the standardised innovations of Model III

of the mean model. The mean model is included in the observation equation and it is a deterministic component of the model. When the normality assumption is dropped, the asymptotic distribution associated with the deterministic components is not affected (Harvey, 1989). Hence, the inference on the parameters of the mean model remains approximately valid.

To compare the annual mean of the most recent year with the mean (or a linear combination of means of) the other years, a general linear hypothesis can be formulated. A general linear hypothesis is formulated as $\boldsymbol{H\beta} = \mathbf{0}$ where $\boldsymbol{H}$ is the $r \times q$ hypothesis matrix. Based on the estimate of the variance of $\hat{\boldsymbol{\beta}}$ (Equation (5.40)) the hypotheses can for instance be tested by means of a Wald type test statistic (Casella and Berger, 2002),

$$T = (\boldsymbol{H}\hat{\boldsymbol{\beta}})^T(\boldsymbol{H}\hat{\boldsymbol{\Sigma}}_\beta\boldsymbol{H}^T)^{-1}(\boldsymbol{H}\hat{\boldsymbol{\beta}}), \tag{5.41}$$

which is asymptotically χ^2_r distributed under the general linear null hypothesis. To answer the research question, one test is needed to check whether the mean nitrate level of 2003 at $S1, \ldots, S5$ is different from the mean of the years 2001 and 2002; and another test is needed to check whether the mean of 2003 is different from the general mean. For the first question the contrast $\beta_{14} - (\beta_{13} + \beta_{12})/2$ is assumed to be 0 under H_0. For the second question, we can test for $\beta_{14} = 0$. We will use the Holms correction for multiplicity (see e.g. Shaffer (1995)).

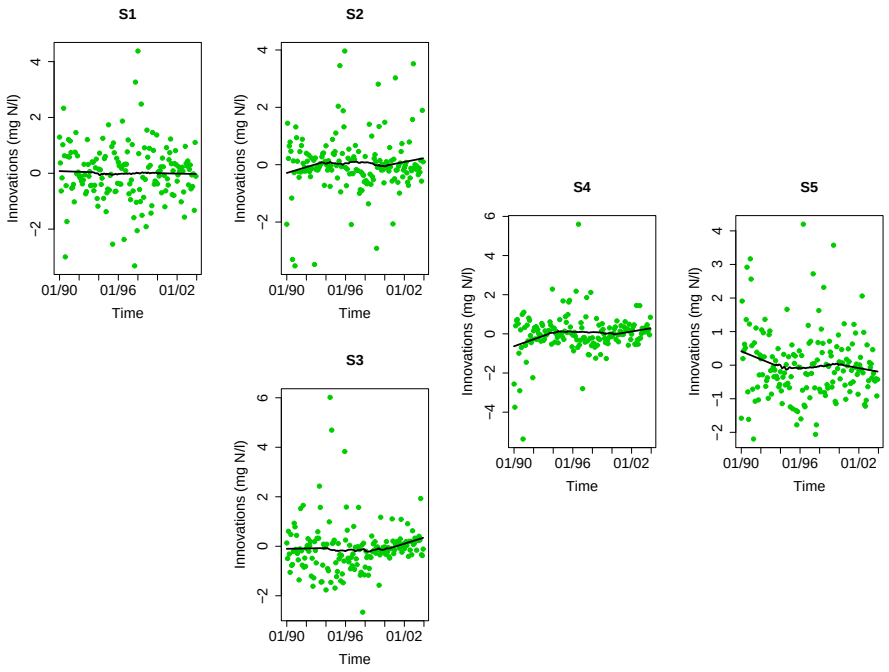

Figure 5.7: Plot of the standardised innovations of Model III. Friedman's super-smoother is added to the plots to assess the residual pattern

When this method is applied to the data, we conclude that for the study region, the mean nitrate concentration in 2003 is very significantly different from the mean of the two years before ($\hat{\beta}_{14} - (\hat{\beta}_{13} + \hat{\beta}_{12})/2 = -2.13$, $p < 0.0001$). The mean concentration in 2003 is also very significantly different from the general mean ($\hat{\beta}_{14} = -2.92$, $p < 0.0001$). The point estimates further show a reduction in the annual mean of the nitrate concentration in the study region. Although the fit of the mean model at S3 might be biased, the p-values of the tests allow us to feel confident about our conclusions.

To improve the fit of the mean model we can extend model III to allow for a different annual mean in the main river ($S1$, $S2$, $S4$ and $S5$) and the joining creek ($S3$). The mean model becomes

$$\begin{aligned} &\mathrm{E}\left(y_{it}\right) = \mu + \alpha_i + \beta_{\lfloor t/12 \rfloor} + \gamma_1 \sin(2\pi t/12) + \gamma_2 \cos(2\pi t/12) \\ &+ (\beta\gamma)_{\lfloor t/12 \rfloor 1} \sin(2\pi t/12) + (\beta\gamma)_{\lfloor t/12 \rfloor 2} \cos(2\pi t/12) + (\alpha\beta)_{i\lfloor t/12 \rfloor} \mathrm{I}(i), \end{aligned} \quad (5.42)$$

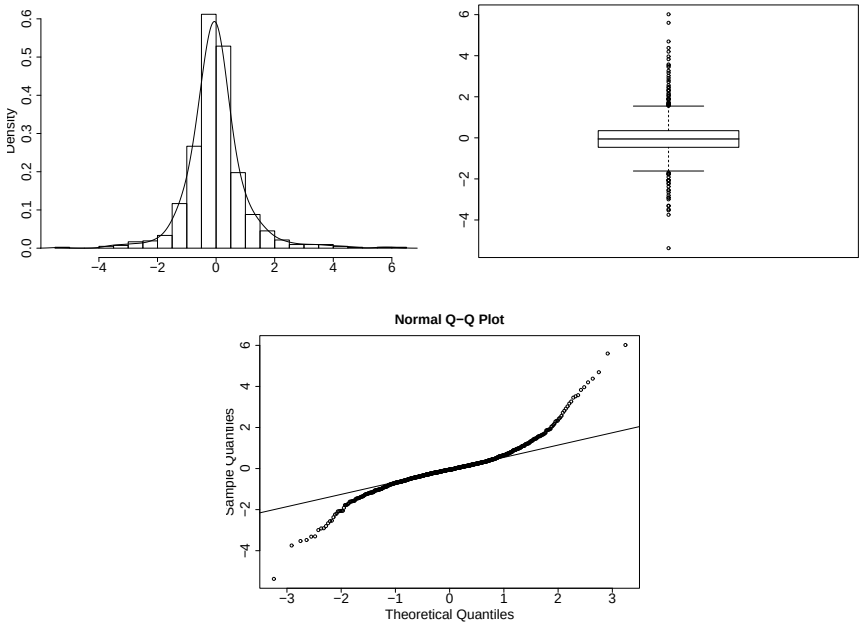

Figure 5.8: Histogram, boxplot and QQ-plot of the standardised innovations of Model III

where I(i) is an indicator function which is -1 for the sampling location S3 and which is 1 elsewhere. This model is referred to as Model IIIb. If this model gives satisfying results, we can infer on a regional scale in the main river and on the level of an individual sampling location in the tributary. The AIC of Model IIIb is 5070.4 which is higher than the AIC of Model III. Hence, according to the AIC the higher complexity of Model IIIb is not adequately reflected in an improved model fit. The GLS fit is shown in Figure 5.9. The fit of the mean model at sampling location S3 seems much better now. The estimates, standard errors and p-values for the parameters of the mean model are given in Table 5.4. The p-value corresponds again to the null-hypothesis that the particular parameter value is equal to zero. Again a number of parameters coding for the main and interaction effects are non-significant. However, this is a weak conclusion. Moreover, the other parameters corresponding of the main and interaction effects are significant. Therefore, the non-significant parameters are not removed from the model.

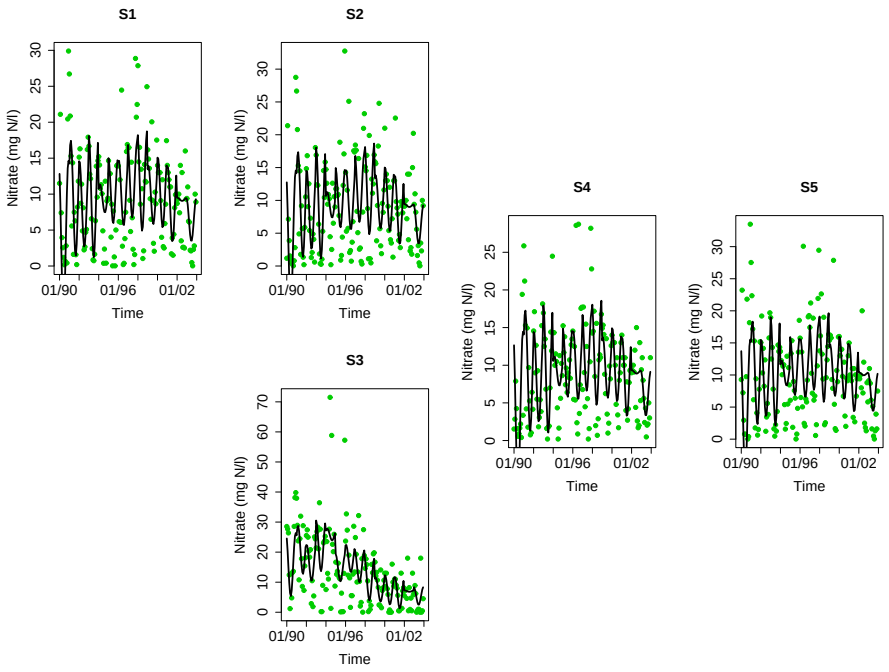

Figure 5.9: Evolution of the water quality at five sampling locations of the river Yzer. Sampling locations S1, S2, S4, S5 are located on the main river, sampling location S3 is located on a tributary which drains into the Yzer between S2 and S4. The line indicates the model fit obtained by Model IIIb

Table 5.4: The parameter estimates of the mean model of Model IIIb

parameter	value	std error	p-value
μ	11.91	0.62	$<$0.0001
β_{14}	-5.65	0.70	$<$0.0001
β_{13}	-3.42	0.66	$<$0.0001
β_{12}	-4.01	0.66	$<$0.0001
β_{11}	-2.72	0.66	$<$0.0001
β_{10}	-2.88	0.66	$<$0.0001
β_9	-0.22	0.66	0.7400
β_8	2.10	0.66	0.0014
β_7	2.46	0.66	0.0002

Table 5.4 – Continued

parameter	value	std error	p-value
β_6	1.08	0.66	0.1000
β_5	6.05	0.66	$<$0.0001
β_4	4.30	0.66	$<$0.0001
β_3	0.97	0.66	0.1400
β_2	3.52	0.66	$<$0.0001
β_1	-1.58	0.56	0.0046
α_1	-2.60	0.60	$<$0.0001
α_2	-2.67	0.44	$<$0.0001
α_3	1.94	0.51	0.0002
α_4	-2.77	0.51	$<$0.0001
α_5	-1.70	0.43	$<$0.0001
γ_1	2.45	0.29	$<$0.0001
γ_2	4.03	0.29	$<$0.0001
$(\alpha\beta)_{3,14}$	2.76	1.19	0.0200
$(\alpha\beta)_{3,13}$	3.39	1.19	0.0042
$(\alpha\beta)_{3,12}$	3.34	1.19	0.0048
$(\alpha\beta)_{3,11}$	3.66	1.19	0.0020
$(\alpha\beta)_{3,10}$	3.43	1.19	0.0038
$(\alpha\beta)_{3,9}$	2.74	1.19	0.0210
$(\alpha\beta)_{3,8}$	0.67	1.19	0.5700
$(\alpha\beta)_{3,7}$	-1.60	1.19	0.1800
$(\alpha\beta)_{3,6}$	0.11	1.19	0.9200
$(\alpha\beta)_{3,5}$	-5.96	1.19	$<$0.0001
$(\alpha\beta)_{3,4}$	-3.94	1.19	0.0009
$(\alpha\beta)_{3,3}$	-1.63	1.19	0.1700
$(\alpha\beta)_{3,2}$	-3.39	1.19	0.0043
$(\alpha\beta)_{3,1}$	-3.59	1.18	$<$0.0001
$(\beta\gamma)_{14,1}$	-1.64	0.97	0.0890
$(\beta\gamma)_{13,1}$	-2.32	0.92	0.1100
$(\beta\gamma)_{12,1}$	1.07	0.92	0.2500
$(\beta\gamma)_{11,1}$	0.75	0.92	0.4200
$(\beta\gamma)_{10,1}$	0.90	0.92	0.3300
$(\beta\gamma)_{9,1}$	-1.65	0.92	0.0730
$(\beta\gamma)_{8,1}$	-2.06	0.92	0.0250
$(\beta\gamma)_{7,1}$	0.90	0.92	0.3300
$(\beta\gamma)_{6,1}$	-0.72	0.92	0.4400
$(\beta\gamma)_{5,1}$	-0.54	0.92	0.5500

Table 5.4 – Continued

parameter	value	std error	p-value
$(\beta\gamma)_{4,1}$	1.47	0.92	0.1100
$(\beta\gamma)_{3,1}$	1.57	0.92	0.0880
$(\beta\gamma)_{2,1}$	5.52	0.92	$<$0.0001
$(\beta\gamma)_{14,2}$	-1.15	1.15	0.3100
$(\beta\gamma)_{13,2}$	-3.85	1.10	0.0005
$(\beta\gamma)_{12,2}$	-0.08	1.10	0.9400
$(\beta\gamma)_{11,2}$	-0.29	1.10	0.7900
$(\beta\gamma)_{10,2}$	-1.80	1.10	0.1000
$(\beta\gamma)_{9,2}$	2.88	1.10	0.0087
$(\beta\gamma)_{8,2}$	1.26	1.10	0.2500
$(\beta\gamma)_{7,2}$	-0.90	1.10	0.4100
$(\beta\gamma)_{6,2}$	0.19	1.10	0.8700
$(\beta\gamma)_{5,2}$	-3.58	1.10	0.0011
$(\beta\gamma)_{4,2}$	3.45	1.10	0.0017
$(\beta\gamma)_{3,2}$	0.49	1.10	0.6500
$(\beta\gamma)_{2,2}$	-3.08	1.10	0.0050

The estimates of the spatial parameters $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{\Sigma_\eta}$ and $\boldsymbol{\Sigma_\epsilon}$ are (standard error and p-value between brackets)

$$\hat{\boldsymbol{A}} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1.1(0.3) & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0.7(0.3) & 0.4(0.2) & 0 & 0 \\ 0 & 0 & 0 & 0.6(0.1) & 0 \end{bmatrix},$$

$$\hat{\boldsymbol{B}} = \begin{bmatrix} -0.008(0.1) & 0 & 0 & 0 & 0 \\ 0 & 0.3(0.1) & 0 & 0 & 0 \\ 0 & 0 & -0.2(0.1) & 0 & 0 \\ 0 & 0 & 0 & -0.03(0.1) & 0 \\ 0 & 0 & 0 & 0 & 0.72(0.7) \end{bmatrix},$$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}} = \begin{bmatrix} 7.0(3.1) & 0 & 0 & 0 & 0 \\ 0 & < 10^{-5}(0.003) & 0 & 0 & 0 \\ 0 & 0 & 34.7(19.3) & 0 & 0 \\ 0 & 0 & 0 & < 10^{-5}(0.002) & 0 \\ 0 & 0 & 0 & 0 & < 10^{-5}(10^{-4}) \end{bmatrix},$$

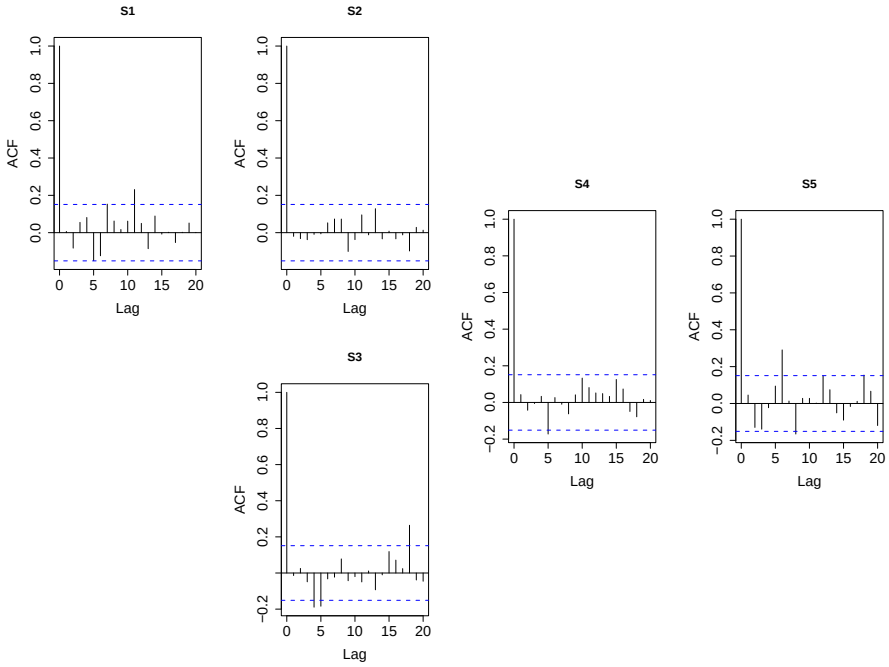

Figure 5.10: Autocorrelation plot of the standardised innovations of Model IIIb

$$\hat{\boldsymbol{\Sigma}}_\epsilon = \begin{bmatrix} 13.4(2.7) & & & & \\ 7.0(3.5) & 19.4(5.5) & & & \\ 3.4(2.6) & 10.3(3.3) & 41.3(19) & & \\ 4.7(0.8) & 5.6(1.7) & -14.9(2.3) & 11.1(1.8) & \\ 16.8(0.05) & 17.6(1.6) & 4.4(2.8) & 10(0.13) & 26.6(0.01) \end{bmatrix}$$

Note that the estimate of $\boldsymbol{B}$ for S3 is now within the unit circle.

ACF plots and a plot of the standardised innovations in function time can be found in Figure 5.10 and Figure 5.11, respectively. The ACF plots seem quite similar to the plots of Model III. In Figure 5.11 the standardised innovations are centered around 0. As compared to Figure 5.7 the smoother at S3 does not alter anymore at the boundaries. Again the smoothers indicate a deviation from zero of the standardised innovations at early dates for S1, S2, S3 and S4. This is again probably due to the combination of a boundary effect of the smoother, large nitrate values measured in the beginning of the time series and the Kalman filter which might not have reached steady state yet. The p-values of Ljung-Box test of the the autocor-

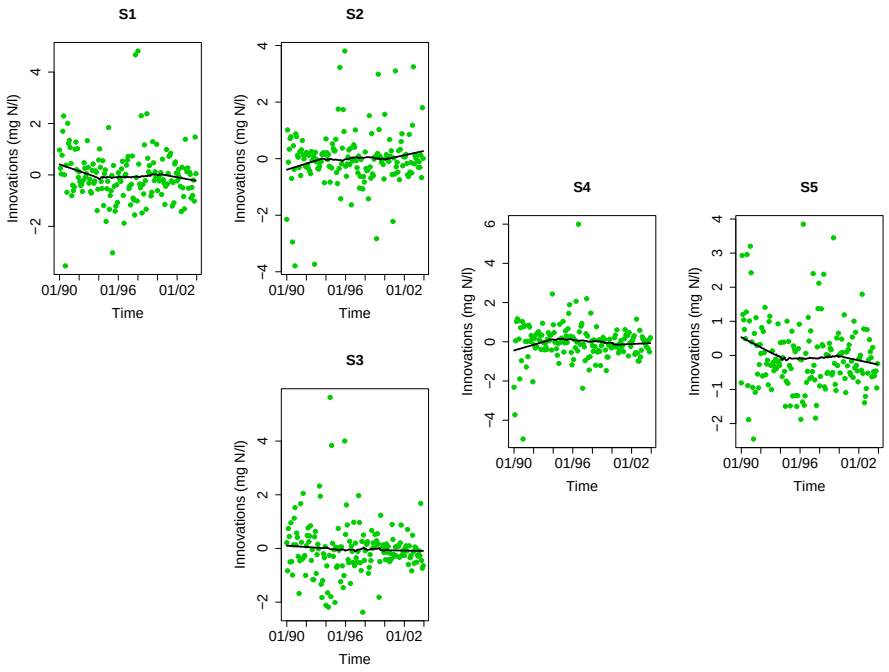

Figure 5.11: Analysis of the standardised innovations of Model IIIb in function of time

Table 5.5: p-values for the Ljung-Box portmanteau test of the autocorrelation coefficients of the standardised innovations of model IIIb at the first 5 lags

Lag	S1	S2	S3	S4	S5
1	0.93	0.80	0.85	0.57	0.55
2	0.55	0.88	0.93	0.73	0.19
3	0.63	0.92	0.90	0.89	0.08
4	0.58	0.97	0.14	0.94	0.15
5	0.23	0.99	0.03	0.30	0.14

relations up to the lag 5 are given in Table 5.5. Only at S3 the Ljung-Box test is significant at lag 5. Note that the AR(1) coefficient at S3 and the eigenvalues of $\mathbf{\Phi}$ are now in the unit circle. We now thus conclude that the state-space model is stationary. Finally, an assessment on the normality of the standardised innovations is presented in Figure 5.12. Again, both the boxplot and the QQ-plot show a clear

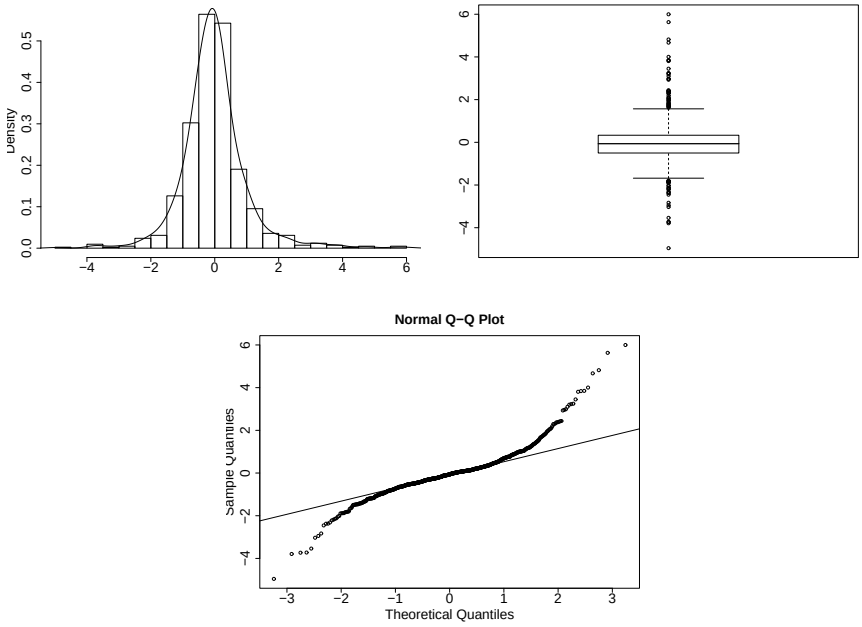

Figure 5.12: Histogram, boxplot and QQ-plot of the innovations of Model IIIb

departure from normality. A considerable amount of outliers is present according to the boxplot and the QQ-plot indicates that the distribution has larger tails that the normal distribution. Similar to the residuals of Model III, from all plots it can be seen that the distribution of the standardised innovations is symmetric. To answer research question, we are again interested in the inference on the parameters of the mean model and their asymptotic distribution is known to be unaffected when the Gaussianity assumption is dropped (Harvey, 1989).

To answer the research question by using model IIIb, four tests are needed:

1. H_0: In the main river, the annual mean of 2003 is equal to the mean of the year 2001 and 2002
   $H_0 : (\beta_{14} + (\alpha\beta)_{3,14} - 1/2(\beta_{12} + (\alpha\beta)_{3,12} + \beta_{13} + (\alpha\beta)_{3,13}) = 0$
2. H_0: In the main river, the annual mean of 2003 is equal to the general mean
   $H_0 : (\beta_{14} + (\alpha\beta)_{3,14}) = 0.$

Table 5.6: p-values of the tests to assess the annual mean of 2003

Test	contrast	p-value	p-holm
	Main river ("Regional")		
2003 $\leftrightarrow$ 2001-2002	-2.54	0.016	0.031
2003 $\leftrightarrow$ general mean	-2.88	0.0005	0.0014
	Joining creek (S3)		
2003 $\leftrightarrow$ 2001-2002	-1.32	0.56	0.56
2003 $\leftrightarrow$ general mean	-8.40	< 0.0001	< 0.0001

3. H_0: In S3 located at the joining creek, the annual mean of 2003 is equal to the mean of the year 2001 and 2002
   $H_0 : (\beta_{14} - (\alpha\beta)_{3,14} - 1/2(\beta_{12} - (\alpha\beta)_{3,12} + \beta_{13} - (\alpha\beta)_{3,13}) = 0$

4. H_0: In sampling location S3, the annual mean of 2003 is equal to the general mean $H_0 : (\beta_{14} - (\alpha\beta)_{3,14}) = 0$.

Again, we use the Holms correction for multiplicity. The contrasts, uncorrected p-values and the corrected p-values are presented in Table 5.6. These results show that in the main river, the mean in 2003 differs significantly from the mean of the last two years and from the general mean. At sampling location $S3$ the mean in 2003 is very significantly different from the general mean, but the mean in 2003 is not different from the mean of the last two years in that sampling location. Again the point estimates indicate that the significant differences correspond to a reduction in the mean nitrate concentration.

5.5 Discussion and Conclusions

A spatio-temporal state-space model is proposed for river monitoring networks where the spatial dependence structure of the state variable is directly derived from the river topology and the temporal dependence structure is modelled by an AR(1) process. The state variable is embedded into an observation model that contains a model for the mean. The latter is needed to answer research questions. With this model it is, for instance, possible to infer on the annual mean nitrate concentration of a river monitoring network. The methodology is shown to be very flexible and

enables the user to test at the level of individual sampling locations as well as on a more regional scale.

A Kalman filter and smoother is formulated for the state-space model, and for the parameter estimation an ECM algorithm is developed. In this algorithm, the parameters of the mean model are estimated by generalised least squares. The parameter estimators are shown to be asymptotically normally distributed. The AIC criterion and an assessment of the standardised innovations are used for model selection and for the evaluation of the quality of the model, respectively.

The temporal correlation structure is restricted to an AR(1) process. In the case study presented here, this seemed to be the right model, but when more complex temporal structures are needed, the methodology can be extended. For instance, Harvey (1989) showed that more general ARMA structures can be handled by the Kalman filter. For example, for an AR(2) process, the state variable $\boldsymbol{S}_t$ has to be replaced by a vector $(S_{1t}, ..., S_{pt}, S_{1t-1}, ..., S_{pt-1})^T$ containing also the state variable at the previous time step. This leads to a reformulation of the observation model and the Kalman filter equations.

The spatial variance-covariance matrix of the observation model $\boldsymbol{\Sigma}_\epsilon$ used a saturated parametrisation. To reduce the complexity in large monitoring networks, $\boldsymbol{\Sigma}_\epsilon$ can be further parameterised (e.g. Xu and Wikle, 2005). Due to the estimation orthogonality in the first CM step, this will only change update Equation (5.35).

The methodology has been applied on a case study at five sampling locations of the river Yzer. Depending on the formulation of the mean model, inference is possible on a regional scale, on the level of a river reach as well as on the level of individual sampling locations. The case study infers on the annual mean of nitrate concentrations of the most recent year. A general linear hypothesis was used to test whether the annual mean of the most recent year was different from the means of the two most recent years and from the general mean. In the study area, the annual average of the nitrate concentration in 2003 is shown to be lower than the general mean ($p < 0.01$). Moreover, in the main river, the mean nitrate concentration of 2003 was also lower than the mean of the two most recent years ($p = 0.03$).

5.6 Appendix: Calculation of the parameters in A and B in CM-step 1

Let $\text{RSS}_i = \sum_{t=1}^{n} \left(S_{it} - \boldsymbol{A}_{i.}^{[a_i]} \boldsymbol{S}_t^{[a_i]} - \boldsymbol{B}_{i.}^{[b_i]} \boldsymbol{S}_{t-1}^{[b_i]} \right)^2$. The estimators in the CM step for $\boldsymbol{A}^{[a_i]}$ and $\boldsymbol{B}^{[b_i]}$ are obtained by maximising Equation (5.30) in the CM step. This is equivalent to the minimisation of RSS_i with respect to $\boldsymbol{A}^{[a_i]}$ and $\boldsymbol{B}^{[b_i]}$, respectively. We find

$$\frac{\partial \text{RSS}_i}{\partial \boldsymbol{A}_{i.}^{[a_i]}} = 0$$

$$\Leftrightarrow 0 = \sum_{t=1}^{n} \left(S_{it} - \boldsymbol{A}_{i.}^{[a_i]} \boldsymbol{S}_t^{[a_i]} - \boldsymbol{B}_{i.}^{[b_i]} \boldsymbol{S}_{t-1}^{[b_i]} \right) \boldsymbol{S}_t^{[a_i]^T}$$

$$\Leftrightarrow \boldsymbol{A}_{i.}^{[a_i]} \sum_{t=1}^{n} \boldsymbol{S}_t^{[a_i]} \boldsymbol{S}_t^{[a_i]^T} = \sum_{t=1}^{n} S_{it} \boldsymbol{S}_t^{[a_i]^T} - \boldsymbol{B}_{i.}^{[b_i]} \sum_{t=1}^{n} \boldsymbol{S}_{t-1}^{[b_i]} \boldsymbol{S}_t^{[a_i]^T}$$

$$\Leftrightarrow \boldsymbol{A}_{i.}^{[a_i]} = \left(\sum_{t=1}^{n} S_{it} \boldsymbol{S}_t^{[a_i]^T} - \boldsymbol{B}_{i.}^{[b_i]} \sum_{t=1}^{n} \boldsymbol{S}_{t-1}^{[b_i]} \boldsymbol{S}_t^{[a_i]^T} \right) \left(\sum_{t=1}^{n} \boldsymbol{S}_t^{[a_i]} \boldsymbol{S}_t^{[a_i]^T} \right)^{-1}$$

$$\frac{\partial \text{RSS}_i}{\partial \boldsymbol{B}_{i.}^{[b_i]}} = 0$$

$$\Leftrightarrow 0 = \sum_{t=1}^{n} \left(S_{it} - \boldsymbol{A}_{i.}^{[a_i]} \boldsymbol{S}_t^{[a_i]} - \boldsymbol{B}_{i.}^{[b_i]} \boldsymbol{S}_{t-1}^{[b_i]} \right) \boldsymbol{S}_{t-1}^{[b_i]^T}$$

$$\Leftrightarrow \boldsymbol{B}_{i.}^{[b_i]} \sum_{t=1}^{n} \boldsymbol{S}_{t-1}^{[b_i]} \boldsymbol{S}_{t-1}^{[b_i]^T} = \sum_{t=1}^{n} S_{it} \boldsymbol{S}_{t-1}^{[b_i]^T} - \boldsymbol{A}_{i.}^{[a_i]} \sum_{t=1}^{n} \boldsymbol{S}_t^{[a_i]} \boldsymbol{S}_{t-1}^{[b_i]^T}$$

$\Updownarrow$ Use (5.32) to replace $\boldsymbol{A}_{i.}^{[a_i]}$

$$\boldsymbol{B}_{i.}^{[b_i]} \sum_{t=1}^{n} \boldsymbol{S}_{t-1}^{[b_i]} \boldsymbol{S}_{t-1}^{[b_i]\,T} = \sum_{t=1}^{n} S_{it} \boldsymbol{S}_{t-1}^{[b_i]\,T} - \left(\sum_{t=1}^{n} S_{it} \boldsymbol{S}_{t}^{[a_i]T} - \boldsymbol{B}_{i.}^{[b_i]} \sum_{t=1}^{n} \boldsymbol{S}_{t-1}^{[b_i]} \boldsymbol{S}_{t}^{[a_i]T} \right) \left(\sum_{t=1}^{n} \boldsymbol{S}_{t}^{[a_i]} \boldsymbol{S}_{t}^{[a_i]T} \right)^{-1} \left(\sum_{t=1}^{n} \boldsymbol{S}_{t}^{[a_i]} \boldsymbol{S}_{t-1}^{[b_i]\,T} \right)$$

$\Updownarrow$

$$\boldsymbol{B}_{i.}^{[b_i]} \left[\sum_{t=1}^{n} \boldsymbol{S}_{t-1}^{[b_i]} \boldsymbol{S}_{t-1}^{[b_i]\,T} - \left(\sum_{t=1}^{n} \boldsymbol{S}_{t-1}^{[b_i]} \boldsymbol{S}_{t}^{[a_i]T} \right) \left(\sum_{t=1}^{n} \boldsymbol{S}_{t}^{[a_i]} \boldsymbol{S}_{t}^{[a_i]T} \right)^{-1} \left(\sum_{t=1}^{n} \boldsymbol{S}_{t}^{[a_i]} \boldsymbol{S}_{t-1}^{[b_i]\,T} \right) \right]$$
$$= \sum_{t=1}^{n} S_{it} \boldsymbol{S}_{t-1}^{[b_i]\,T} - \left(\sum_{t=1}^{n} S_{it} \boldsymbol{S}_{t}^{[a_i]T} \right) \left(\sum_{t=1}^{n} \boldsymbol{S}_{t}^{[a_i]} \boldsymbol{S}_{t}^{[a_i]T} \right)^{-1} \left(\sum_{t=1}^{n} \boldsymbol{S}_{t}^{[a_i]} \boldsymbol{S}_{t-1}^{[b_i]\,T} \right)$$

$\Updownarrow$

$$\boldsymbol{B}_{i.}^{[b_i]} = \left[\sum_{t=1}^{n} S_{it} \boldsymbol{S}_{t-1}^{[b_i]\,T} - \left(\sum_{t=1}^{n} S_{it} \boldsymbol{S}_{t}^{[a_i]T} \right) \left(\sum_{t=1}^{n} \boldsymbol{S}_{t}^{[a_i]} \boldsymbol{S}_{t}^{[a_i]T} \right)^{-1} \left(\sum_{t=1}^{n} \boldsymbol{S}_{t}^{[a_i]} \boldsymbol{S}_{t-1}^{[b_i]\,T} \right) \right]$$
$$\times \left[\sum_{t=1}^{n} \boldsymbol{S}_{t-1}^{[b_i]} \boldsymbol{S}_{t-1}^{[b_i]\,T} - \left(\sum_{t=1}^{n} \boldsymbol{S}_{t-1}^{[b_i]} \boldsymbol{S}_{t}^{[a_i]T} \right) \left(\sum_{t=1}^{n} \boldsymbol{S}_{t}^{[a_i]} \boldsymbol{S}_{t}^{[a_i]T} \right)^{-1} \left(\sum_{t=1}^{n} \boldsymbol{S}_{t}^{[a_i]} \boldsymbol{S}_{t-1}^{[b_i]\,T} \right) \right]^{-1}$$

Chapter 6

Spatio-temporal modelling of river monitoring networks, a semi-parametric approach

6.1 Introduction

The Water Framework Directive (WFD)(EC, 2000) aims to trigger local authorities to improve the aquatic environment. To reach that goal, the Flemish environmental agency (VMM) is developing basin management plans to improve the water quality of the rivers in Flanders (Belgium). A dominant problem in Flemish water bodies is the eutrophication due to nutrient pollution. One of the main nutrient pollution sources originates from agricultural activities. In Flanders there is an intensive

pig farming activity and in the past the produced manure was mainly disposed on agricultural lands. A major action to reduce this nutrient load was the introduction of two Manure Action Plans (MAP's) (Vlaams Parlement, 1995, 1999). The MAP's restrict the amount of fertilisers that may be used by farmers in area's which are susceptible to eutrophication. The first MAP was introduced in 1996 (Vlaams Parlement, 1995) and after an evaluation a new and more restrictive MAP was implemented in 2000 (Vlaams Parlement, 1999). When such actions are taken, it is important to assess whether they indeed have an effect on the water quality. Therefore water quality monitoring networks are needed to assess the evolution of the water quality. In Flanders, the VMM has developed several monitoring networks along the rivers. In their physico-chemical monitoring network, a basic spectrum of physico-chemical variables is evaluated monthly at each sampling location. In this chapter we assess the evolution of the nitrate concentration in a small region of the Yzer basin. This river is located in the Western part of Flanders. It is a rural area with a large agricultural activity.

The focus of this chapter lays on the development of a methodology to detect and to locate trends in the water quality data. Instead of assessing trends at the level of individual sampling locations, our aim is to develop a methodology for trend detection on a more regional scale. Standard techniques cannot be used for this purpose because river monitoring networks typically generate data with a strong spatial and temporal dependence structure. In order for the statistical inference procedure to be formally valid, this dependence has to be taken into account. Many researchers, however, have avoided the estimation of the spatio-temporal dependence in river monitoring network data by simply ignoring it or by using ad hoc methods. Burn and Hag Elnur (2002), for instance, adopted an approach to determine the *field significance* that is involved in the calculation of a regional value for the Mann Kendall statistic. To correct for serial correlation, they proposed to first perform a pre-whitening step which preserves the trend. To correct for the spatial correlation, they suggested a resampling strategy that constructs bootstrapped datasets by selecting the time instants to be included at random until the original number of sampling times is reached. For each of the selected time instants the corresponding data at all sampling locations has to be used to preserve the spatial pattern. Hence, the temporal structure such as trends that existed in the original data, is not reproduced in the resampled datasets, but the spatial pattern remains. Finally, Mann Kendall statistics are calculated for each of the sampling locations of the resampled datasets to derive a kind of field significance level under the null-hypothesis. This field significance is then used to assess the trends on a more regional scale. Beside the stagewise removal of the dependences, another disadvantage of their approach

is the assumption of monotonic trends. In environmental systems, however, the trend is often nonlinear and changes over time. Hence, if there are sign changes in the trend during the period of interest, tests for monotonic trends are not useful.

To control the type I error rate of the trend tests at more than one sampling location, we suggest to use a spatio-temporal model for a river monitoring network that takes the spatio-temporal dependence structure explicitly into account. In contrast with ad hoc methods, the modelling approach provides a very natural way to introduce the spatio-temporal dependence structure into the testing procedure. River monitoring networks, however, possess a specific spatial dependence structure. As compared to classical geostatistical models, an important distinction has to be made with respect to the spatial dependence structure: due to the direction of the flow a causal interpretation can be given to the correlations. Moreover, rivers can join or split, which implies a more general branched unidirectional structure. In reality the environmental conditions may obscure the unidirectional spatial dependence structure implied by the river topology. We therefore only impose this restrictive topology-implied dependence structure on an unobservable state variable $\boldsymbol{S}$. The latent variable $\boldsymbol{S}$ is embedded in an observation model $\boldsymbol{y}$ that allows cross-correlation between sampling locations that are located at different branches of the river, so that more realistic dependence structures are allowed. Besides the dependence structure, we also have to model the marginal mean to assess the trends. Trends in water quality are often nonlinear. Therefore we propose here a trend detection based on local polynomial regression smoothers. To enable an assessment of the trend on a regional scale, a common nonparametric trend is estimated at all sampling locations. The evaluation of the local trend is done by testing that the first derivative of the nonlinear trend is significant. This has to be performed at each time instant, and leads to a large number of simultaneous tests. Therefore, a multiplicity correction procedure is required. In general, observations which are close in time are likely to have similar trends. Thus, in our setting, the trend tests are not independent, and this reduces the actual dimension of the multiplicity problem. In this chapter we present a procedure that corrects for multiplicity and takes the dependence between the tests explicitly into account. This leads to a correct statistical inference procedure that is not too conservative.

The organisation of the chapter is as follows. First the spatio-temporal model is briefly presented in Section 6.2. The difference with the model presented in Chapter 5 is situated in the formulation of the mean model. The mean model is now semi-parametric because a smoother is used for the trend estimation. The parameter estimation procedure is introduced in Section 6.3. Section 6.4 deals with the

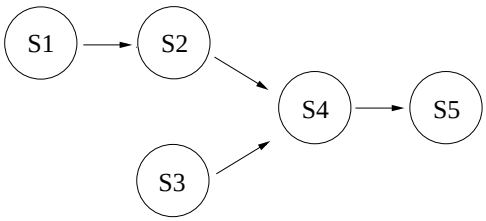

Figure 6.1: Directed Acyclic Graph (DAG) of five sampling locations along two joining river reaches

trend detection method, and in Section 6.5 the methodology is illustrated in a case study. Finally we will formulate some conclusions in Section 6.6.

6.2 Spatio-temporal model

Let $\boldsymbol{S} = (S_1, \ldots, S_p)^T$ represent the $p \times 1$ vector of response variables S_i at sampling locations $i = 1, \ldots, p$. The correlation structure of $\boldsymbol{S}$ is completely defined by the river topology. This is illustrated in Figure 6.1, which shows the river topology of 5 sampling locations. The same figure can also be interpreted as a Directed Acyclic Graph (DAG) (see e.g. Whittaker, 1990) in which the circles represent S_i's and arrows immediately determine the conditional independence structure. For example, observations at sampling location $S4$ are independent of $S1$ given observations at $S2$ because all the water from $S1$ has to pass through $S2$ before it can reach $S4$. The DAG can be modelled by

$$\boldsymbol{S} = \boldsymbol{A}\boldsymbol{S} + \boldsymbol{\gamma}, \tag{6.1}$$

where $\boldsymbol{A} = (a_{ij})_{i,j}$ can be written as a $p \times p$ lower triangular square matrix with zeroes at the diagonal, and $\boldsymbol{\gamma}$ is multivariate normally distributed (MVN): $\boldsymbol{\gamma} \sim MVN(0, \boldsymbol{\Sigma}_\gamma)$ with a diagonal variance-covariance matrix $\boldsymbol{\Sigma}_\gamma$. When the model is applied to the graph in Figure 6.1, it can be seen that $\boldsymbol{A}$ becomes

$$\boldsymbol{A} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ a_{21} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & a_{42} & a_{43} & 0 & 0 \\ 0 & 0 & 0 & a_{54} & 0 \end{bmatrix}$$

where a_{ij} models the dependence between sampling location S_i and S_j.

The river monitoring network, however, generates data over time. The spatial pattern of the DAG is thus repeated over time and we have to extend Equation (6.1) to also take the temporal dependence into account. Let $\boldsymbol{S}_t = (S_{1t}, \ldots, S_{pt})^T$. We assume a Markovian dependence structure and we model $\boldsymbol{S}_t$ by conditioning on $\boldsymbol{S}_{t-1}$. Extending and rearranging Equation (6.1) gives

$$\boldsymbol{S}_t = (\boldsymbol{I}_p - \boldsymbol{A})^{-1}\boldsymbol{B}\boldsymbol{S}_{t-1} + (\boldsymbol{I}_p - \boldsymbol{A})^{-1}\boldsymbol{\eta}_t, \tag{6.2}$$

$t = 1, \ldots, n$, where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix, $\boldsymbol{\eta}_t \sim MVN(0, \boldsymbol{\Sigma}_\eta)$ with a $p \times p$ diagonal variance covariance matrix $\boldsymbol{\Sigma}_\eta$, and $\boldsymbol{B}$ is a $p \times p$ matrix containing the temporal autocorrelation (diagonal elements) and the spatio-temporal cross-correlation coefficients (off-diagonal elements). Similar to the matrix $\boldsymbol{A}$, we propose to only use cross-correlations between sampling locations that are directly connected according to the DAG structure. The off-diagonal elements of $\boldsymbol{B}$ are thus structured in a similar way as the elements of matrix $\boldsymbol{A}$. Hence $\boldsymbol{B}$ can be written as

$$\boldsymbol{B} = \begin{bmatrix} b_{11} & 0 & 0 & 0 & 0 \\ b_{21} & b_{22} & 0 & 0 & 0 \\ 0 & 0 & b_{33} & 0 & 0 \\ 0 & b_{42} & b_{43} & b_{44} & 0 \\ 0 & 0 & 0 & b_{54} & b_{55} \end{bmatrix}.$$

When $i \neq j$ the b_{ij} model the spatio-temporal dependence between S_{it} and S_{jt-1} and the b_{ii} model the temporal dependence between S_{it} and S_{it-1}. For completeness the initial conditions have to be defined at time instant 0. We assume $\boldsymbol{S}_0$ to be $MVN(\boldsymbol{0}, \boldsymbol{\Sigma}_{S_0})$.

In reality, however, the dependence structure might be obscured by common environmental confounders, such as rainfall. Therefore, the model is embedded into an observation model,

$$\boldsymbol{y}_t = \boldsymbol{S}_t + \boldsymbol{\epsilon}_t, \tag{6.3}$$

$t = 1, \ldots, n$, where $\boldsymbol{y}_t$ is the observation vector corresponding to $\boldsymbol{S}_t$, and $\boldsymbol{\epsilon}_t \sim MVN(\boldsymbol{0}, \boldsymbol{\Sigma}_\epsilon)$. No restrictions are imposed on $\boldsymbol{\Sigma}_\epsilon$. This enables cross correlations between sampling locations that are not connected according to the river topology.

Model (6.3) only defines the spatio-temporal dependence structure. It can be easily seen that $E[\boldsymbol{y_t}] = \boldsymbol{0}$ at all time instants. To model the trend, Equation (6.3) is extended with an additive model for the mean. Besides a trend, seasonal variation is typically present in water quality data. In Chapter 1 we introduced some

of the data. The seasonal variation was illustrated in Figure 1.6 where nitrate data of all years was plotted in function of the day of the year (support $[1, 365]$). A common approach to deal with this variation is to include sinusoidal functions of fixed periods to describe the seasonal cycle within a year (e.g. Hirst, 1998; Cai and Tiwari, 2000; McMullan et al., 2003; McMullan, 2004). A common function which is used for this purpose is $\alpha \cos(2\pi(t/P) + \theta)$, where P is the period which is taken to be 1 year, α is the amplitude of the seasonal trend and θ is a parameter to allow for a phase shift. Hence, α and θ have to be estimated. This term, however, is nonlinear in the parameter θ because the parameter appears within the cosine function. However, this term can be expressed in a linear form by using a standard trigonometric expansion of the cosine term. This is also the parametrisation of our choice and therefore we use Fourier basis functions to model the seasonal effect $\gamma_1 \sin(2\pi t/365) + \gamma_2 \cos(2\pi t/365)$. Hence, the following mean model is proposed: $\mathrm{E}[y_{it}] = \boldsymbol{X}_{it}\boldsymbol{\beta} + f(t)$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^T$ is the $q \times 1$ parameter vector and $\boldsymbol{X}_{it}$ is the $1 \times q$ design vector that includes the proper Fourier basis functions and some other linear predictors, and $f(t)$ is a local linear regression smoother for the estimation of the nonlinear trend. Note that $f(t)$ does not depend on the sampling location because we want to assess the trend on a regional scale. After embedding the mean model into Model (6.3), we obtain

$$\boldsymbol{y}_t = \boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{f}(t) + \boldsymbol{S}_t + \boldsymbol{\epsilon}_t, \tag{6.4}$$

which specifies together with Model (6.2) the complete time-invariant spatio-temporal state-space model.

An equivalent formulation of the spatio-temporal model is accomplished by recognising that the Model (6.2) and (6.4) can be written as a Structural Equation Model (SEM) (see e.g. Maruyama, 1997),

$$\boldsymbol{C}\boldsymbol{S}_N = \boldsymbol{\zeta} \tag{6.5}$$

$$\boldsymbol{Y}_N = \boldsymbol{X}_N\boldsymbol{\beta} + \boldsymbol{f}_N + \boldsymbol{S}_N + \boldsymbol{\psi}, \tag{6.6}$$

where $\boldsymbol{S}_N = (\boldsymbol{S}_1^T, \ldots, \boldsymbol{S}_n^T)^T$, $\boldsymbol{Y}_N = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$, $\boldsymbol{X}_N = (\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_n^T)^T$, $\boldsymbol{f}_N = (\boldsymbol{f}^T(1), \ldots, \boldsymbol{f}^T(n))^T$, $\boldsymbol{C}$ is a $pn \times pn$ square matrix constructed from the elements of the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{\zeta} \sim MVN(\boldsymbol{0}, \boldsymbol{\Sigma}_\zeta)$, where $\boldsymbol{\Sigma}_\zeta$ is a diagonal matrix built from the corresponding elements of $\boldsymbol{\Sigma}_\eta$, and $\boldsymbol{\psi} \sim MVN(\boldsymbol{0}, \boldsymbol{\Sigma}_\psi)$ where $\boldsymbol{\Sigma}_\psi$ is block-diagonal with blocks $\boldsymbol{\Sigma}_\epsilon$. From this SEM formulation the covariance structure of the observation vector $\boldsymbol{Y}_N$ is easily found,

$$\boldsymbol{\Sigma}_{Y_N}(\boldsymbol{\Psi}_{\boldsymbol{\alpha}}) = \mathrm{var}\,(\boldsymbol{Y}_N) = \boldsymbol{C}^{-1}\boldsymbol{\Sigma}_\zeta\boldsymbol{C}^{-T} + \boldsymbol{\Sigma}_\psi, \tag{6.7}$$

with $\boldsymbol{\Psi}_{\boldsymbol{\alpha}}$ the vector that contains all parameters in $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{\Sigma}_{\boldsymbol{S}_0}$, $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$.

6.3 Parameter estimation and statistical inference procedure

As in Chapter 5 it is possible to perform the parameter estimation and the inference procedure completely in the likelihood framework. However, in order to control the computational burden we will consider a slightly different approach where the mean model is estimated by ordinary least squares (OLS). OLS also provides an unbiased estimator but it is asymptotically less efficient than generalised least squares (GLS) (e.g. Shin and Oh, 2002). Thus, the variance of the OLS estimators will be larger. The parameter estimation of the mean model is given in Section 6.3.1. In Section 6.3.2 the estimation procedure for the parameters of the dependence structure is briefly discussed.

6.3.1 Mean model

In Chapter 5 a linear spatio-temporal model was used to model the mean. The parameter estimation was done within the likelihood framework which implies the use of generalised least squares (GLS) for the estimation of the parameters of the mean model. In this chapter the approach of Chapter 5 is extended by introducing a smoother in the mean model for the estimation of a nonlinear trend. This nonlinear trend is estimated by the use of a polynomial smoother (An overview of fitting local polynomial smoothers can be found in Section 2.2.3). Because a smoother is involved in the mean model, we will have to obtain the smoother matrix to fit the semiparametric mean model. In a GLS context, the dependence structure is involved in the calculation of the smoother matrix (see e.g. Giannitrapani et al., 2005). Hence, the only adjustment which is needed to use the ECM algorithm of Chapter 5, is to adapt the second CM to enable the fit of the semiparametric mean model. However, this would imply the recalculation of the projection matrix of the smoother at each iteration and would lead to a drastic increase of the computational power that is needed to estimate the model parameters. Therefore we will use ordinary least squares (OLS) to fit the mean model. OLS estimators are also unbiased and consistent, but they are asymptotically less efficient (e.g. Shin and Oh, 2002). Thus, the variance of the OLS estimator will be larger than the variance of GLS estimators. From a computational point of view, however, they have a considerable advantage because the parameters of the mean model only have to be estimated once and the parameters of the dependence structure can then be estimated using

the residuals of the OLS estimation procedure. Given these considerations we prefer OLS.

When the OLS procedure is applied to our particular additive model, the estimating equations have an analytical solution (see Section 2.3, and Hastie and Tibshirani (1990)). The following results are obtained for the OLS of our model,

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}_N^T(\boldsymbol{I}_N - \boldsymbol{S}_f)\boldsymbol{X}_N)^{-1}\boldsymbol{X}_N^T(\boldsymbol{I}_N - \boldsymbol{S}_f)\boldsymbol{Y}_N = \boldsymbol{H}_\beta \boldsymbol{Y}_N \tag{6.8}$$

$$\hat{\boldsymbol{f}} = \boldsymbol{S}_f(\boldsymbol{Y}_N - \boldsymbol{X}_N\hat{\boldsymbol{\beta}}), \tag{6.9}$$

where $\boldsymbol{S}_f$ is the smoother matrix and $\boldsymbol{I}_N$ is the $N \times N$ identity matrix. Hence, a projection matrix $\boldsymbol{H}_f$ can be constructed for the smoother term,

$$\boldsymbol{H}_f = \boldsymbol{S}_f(\boldsymbol{I}_N - \boldsymbol{X}_N(\boldsymbol{X}_N^T(\boldsymbol{I}_N - \boldsymbol{S}_f)\boldsymbol{X}_N)^{-1}\boldsymbol{X}_N^T(\boldsymbol{I}_N - \boldsymbol{S}_f)). \tag{6.10}$$

For inference procedures, this is advantageous. Once the covariance matrix of the observations $\boldsymbol{Y}_N$ is available, the covariance matrix of the smoother estimators can be obtained. To assess whether a beneficial trend occurs after a certain time, we have to infer the first derivative of the trend. For local polynomial regression smoothers, a smoother matrix $\boldsymbol{S}_{f^{(1)}}$ for the first derivative $\boldsymbol{f}^{(1)}$ is available (Fan and Gijbels, 1996). The smoother, however, is embedded in an additive model, thus a projection matrix $\boldsymbol{H}_{f^{(1)}}$ for the first derivative has to be calculated. For local polynomial smoothers, this becomes

$$\begin{aligned}\hat{\boldsymbol{f}}^{(1)} &= \boldsymbol{S}_{f^{(1)}}(\boldsymbol{Y}_N - \boldsymbol{X}_N\hat{\boldsymbol{\beta}}) \\ &= \boldsymbol{S}_{f^{(1)}}(\boldsymbol{I}_N - \boldsymbol{X}_N(\boldsymbol{X}_N^T(\boldsymbol{I}_N - \boldsymbol{S}_f)\boldsymbol{X}_N)^{-1}\boldsymbol{X}_N^T(\boldsymbol{I}_N - \boldsymbol{S}_f))\boldsymbol{Y}_N \\ &= \boldsymbol{H}_{f^{(1)}}\boldsymbol{Y}_N,\end{aligned} \tag{6.11}$$

with

$$\boldsymbol{H}_{f^{(1)}} = \boldsymbol{S}_{f^{(1)}}(\boldsymbol{I}_N - \boldsymbol{X}_N(\boldsymbol{X}_N^T(\boldsymbol{I}_N - \boldsymbol{S}_f)\boldsymbol{X}_N)^{-1}\boldsymbol{X}_N^T(\boldsymbol{I}_N - \boldsymbol{S}_f)) \tag{6.12}$$

6.3.2 Dependence structure

To fit the parameters of the dependence structure, we propose to apply a slightly adjusted version of the ECM algorithm of Section 5.3.3. Because the parameters of the mean model are estimated by OLS, the second CM step dealing with the estimation of the parameters of the mean model is redundant and only the first CM step is used. The only adjustment that is needed here is to replace $\boldsymbol{y}_t'$ by $\boldsymbol{y}_t' = \boldsymbol{y}_t - \boldsymbol{X}_t\boldsymbol{\beta} - \boldsymbol{f}(t)$.

6.4 Statistical inference procedure

Since the parameters of the seasonal component and the nonlinear trend are a linear combination of the responses, $\hat{\boldsymbol{\beta}} = \boldsymbol{H}_\beta \boldsymbol{Y}_N$ and $\hat{\boldsymbol{f}} = \boldsymbol{H}_f \boldsymbol{Y}_N$, inference on the mean parameters and the nonlinear trend require an estimator of the complete variance-covariance matrix of the observation vector $\boldsymbol{Y}_N$. From the SEM model representation an estimator of $\boldsymbol{\Sigma}_{Y_N}$ can be calculated directly by plugging in the parameter estimates in Equation (6.7). Let $\hat{\boldsymbol{\Sigma}}_{Y_N}$ denote this estimator. The variance-covariance matrix of the parameter estimators for the seasonal effect is thus consistently estimated by

$$\hat{\boldsymbol{\Sigma}}_\beta = \boldsymbol{H}_\beta \hat{\boldsymbol{\Sigma}}_{Y_N} \boldsymbol{H}_\beta^T. \tag{6.13}$$

To know whether the nonlinear trend is present at a certain time t, an analysis of its derivative, $\boldsymbol{f}^{(1)}(t)$, is proposed. For local linear regression smoothers the derivative can be calculated and is linear in the response. Since projection matrices exist for the local polynomial regression smoother and its first derivative, the calculation of the estimator of the variance-covariance matrix of the nonlinear trend ($\boldsymbol{\Sigma}_f$) and of the derivative ($\boldsymbol{\Sigma}_{f^{(1)}}$) is straightforward. They are given by

$$\hat{\boldsymbol{\Sigma}}_f = \boldsymbol{H}_f \hat{\boldsymbol{\Sigma}}_Y \boldsymbol{H}_f^T \tag{6.14}$$

$$\hat{\boldsymbol{\Sigma}}_{f^{(1)}} = \boldsymbol{H}_{f^{(1)}} \hat{\boldsymbol{\Sigma}}_Y \boldsymbol{H}_{f^{(1)}}^T. \tag{6.15}$$

Simple test statistics can thus be used for asymptotic pointwise inference on the derivative, e.g. $t = f^{(1)}(t)/s_{f^{(1)}}(t)$ is asymptotically standard normally distributed under the null-hypothesis of no trend. Since the test is performed at each time instant, we have to correct for multiplicity. A widely used method to take multiplicity into account is to use adjusted p-values. Well-known examples of this approach are classical methods such as the Bonferroni or Holm procedures. They consider all the tests to be independent and they are known to be too conservative when this is not the case (e.g. Shaffer, 1995). In our application, tests at time instants which are close to one another are likely to be correlated. Thus, the effective dimension of the multiplicity problem is reduced. We therefore propose to use a procedure which can take these dependences explicitly into account. In particular we have chosen to use the free step-down resampling method (algorithm 2.8 of Westfall and Young, 1993). Their procedure proceeds as follows

1. Rank the original p-values: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(n)}$, where (j) denotes the rank number and store the ranked p-values in the vector $(p_{(1)}, \ldots, p_{(n)})$

2. Initialise the count variables: $COUNT_i = 0, i = 1, \ldots, n$

3. Generate a vector $(p^*_{(1)}, \ldots, p^*_{(n)})$ from the same (or at least, approximately the same) distribution of the *original* p-values $(p_{(1)}, \ldots, p_{(n)})$ under the complete null hypothesis. Note that the sequence $\{(j)\}$ is fixed throughout the simulation. Thus the $p^*_{(j)}$ will not have the same monotonicity as the original p-values $p_{(j)}$.

4. Define the successive minima:

$$\begin{aligned} q^*_n &= p^*_{(n)} \\ q^*_{n-1} &= \min(q^*_n, p^*_{(n-1)}) \\ q^*_{n-2} &= \min(q^*_{n-1}, p^*_{(n-2)}) \\ &\vdots \\ q^*_1 &= \min(q^*_2, p^*_{(1)}). \end{aligned}$$

5. If $q^*_i \leq p_{(i)}$, then $COUNT_i = COUNT_i + 1$.

6. Repeat step 3-5 B times, compute the adjusted p-values $\tilde{p}{}^{(B)}_{(i)}$ as $\tilde{p}{}^{(B)}_{(i)} = \frac{COUNT_i}{B}$.

7. Enforce monotonicity using successive maximisation:

$$\begin{aligned} \tilde{p}{}^{(B)}_{(1)} &= \tilde{p}{}^{B}_{(1)} \\ \tilde{p}{}^{(B)}_{(1)} &= \max(\tilde{p}{}^{(B)}_{(1)}, \tilde{p}{}^{(B)}_{(2)}) \\ &\vdots \\ \tilde{p}{}^{(B)}_{(n)} &= \max(\tilde{p}{}^{(B)}_{(n-1)}, \tilde{p}{}^{(B)}_{(n)}). \end{aligned}$$

Westfall and Young (1993) argue that once the monotonicity is enforced and if B is sufficiently large that the $\tilde{p}{}^{(B)}_{(j)}$ are reasonable approximations of the actual $\tilde{p}_{(j)}$. They also recommend to take $B \geq 10000$. One of the possibilities Westfall and Young (1993) proposed to perform step 3 is to sample from a parametric estimate of the null distribution $\hat{F}_0$. When F_0 is a known function that depends on a vector of unknown parameters $\boldsymbol{\Theta}$, $F_0 = F_0(\boldsymbol{\Theta})$, one can sample from $\hat{F}_0 = F_0(\hat{\boldsymbol{\Theta}})$, where $\hat{\boldsymbol{\Theta}}$ is a consistent estimate of $\hat{\boldsymbol{\Theta}}$. In our application, a simulated sample from $\hat{F}_0$ can be obtained by

1. sampling a new set of derivatives $\boldsymbol{f}^{(1)*}$ under the null-hypothesis of no trend from $MVN(\mathbf{0}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{f}^{(1)}})$,
2. calculating the p-values p_k^* that correspond to each of the simulated derivatives $f_k^{(1)*}$, and
3. ranking these p-values according to the *original* p-values $(p_{(1)}, \ldots, p_{(n)})$ to obtain $(p^*_{(1)}, \ldots, p^*_{(n)})$.

In the above, inference is provided for the components of the mean model. To obtain standard errors of the parameter estimators of the dependence structure, we will estimate the observed Fisher information matrix. In this dissertation, this is done by numerical perturbation of the likelihood function (e.g. Harvey, 1989 and Shumway and Stoffer, 2006).

Model selection will be based on the AIC criterion and the quality of the model will be checked in an analysis of the standardised innovations (for more details see Section 5.3.5).

6.5 Case study

The data used in this case study is part of a public database of the Flemish environmental agency (http://www.vmm.be). Five sampling locations along two joining river reaches located in the Yzer basin (Belgium) are used to assess whether there exists a trend in the nitrate concentration between January 1990 and December 2003. Their DAG and locations in the catchment are shown in Figure 6.2. Sampling locations S1, S2, S4 and S5 are located on the Yzer while sampling location S3 is located on a joining creek. For each sampling location, monthly nitrate measurements are available between January 1990 and December 2003. Hence the five sampling locations are sampled on 168 different time instants resulting in a total sample size of 840 observations. Since the observations are taken at time intervals which are much larger than the time scale of the water flow, the matrix B describing the temporal correlation, can be assumed to be diagonal, i.e. an AR(1) structure. Instead of looking for trends at the level of individual sampling locations, we aim to detect the trend on a more regional scale and impose the restriction that all locations have the same trend in common. This assumption is later assessed in the analysis of the innovations. The nonlinear trend is estimated by means of a local

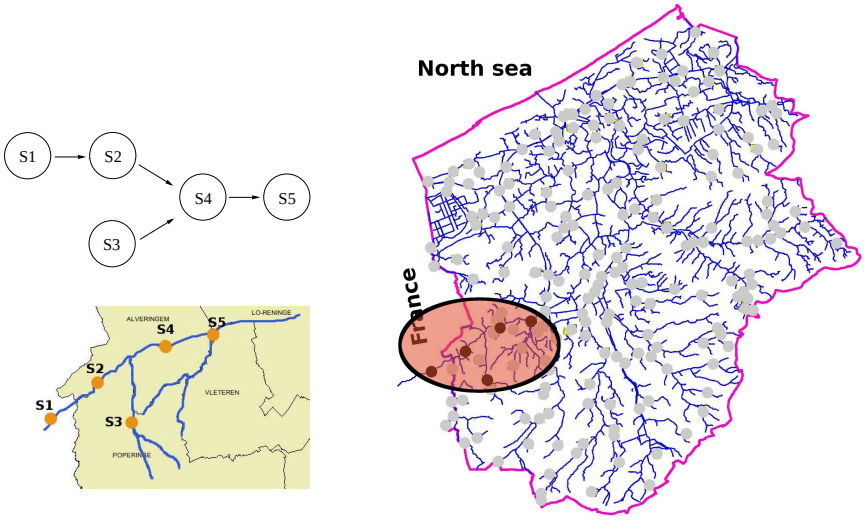

Figure 6.2: Top Left: Directed Acyclic Graph (DAG) of the sampling locations along the 2 river reaches. Bottom Left: Map of the river reaches considered in this case study. Locations S1, S2, S4 and S5 are located on the Yzer river while location S3 is located on a joining creek. Sampling location S1 is located in France. Right: Map of the part of the Yzer catchment located in Flanders, Belgium. The sampling locations are indicated by the dots. The area considered in this study is indicated with the ellipse and the considered sampling locations are indicated with black dots

polynomial regression smoother for which we use the Epanechnikov kernel. Local polynomial regression was introduced in Section 2.2.3. The choice of the kernel is not that important from a practical point of view (e.g. Fan and Gijbels, 1996). But, Fan and Gijbels (1996) showed that the Epanechnikov kernel has some nice asymptotical properties. The bandwidth was selected by a grid search using the AIC criterion (see e.g. Chapter 2). Next we introduce the models that were fitted to the data. Let μ denote the intercept at sampling location 5, and α_i the effect of the i^{th} sampling location relative to sampling location 5 (hence $\alpha_5 = 0$). The value of the regional trend at time t is denoted by $f(t)$, the γ_k are the parameters of the seasonal component modelled by Fourier terms, and the $(\alpha\gamma)_{ik}$ are parameters of the sampling location-season interactions. In contrast to the models in Chapter 5 no year-season interaction term could be used to enable the seasonal effect to change from one year to another. We would only include an interaction term in

Table 6.1: Mean models to assess the "regional" nonlinear trend in the nitrate concentration

Model	E (y_{it})	AIC
I	$\mu + \alpha_i + \gamma_1 \sin(2\pi t/12) + \gamma_2 \cos(2\pi t/12) + f(t)$	5102.0
II	$\mu + \alpha_i + \gamma_1 \sin(2\pi t/12) + \gamma_2 \cos(2\pi t/12) + f(t)$ $+(\alpha\gamma)_{i1} \sin(2\pi t/12) + (\alpha\gamma)_{i2} \cos(2\pi t/12)$	5096.4

the model if the model also contains the main effect, and the factor for year could not be included in the model since a main effect as it would interfere with the estimation of the nonlinear trend. Table 6.1 presents the models that were considered. The models are fitted by using the methods described in Section 6.3.

Model II has the lowest AIC and it is selected as the final model. The resulting OLS fit of the mean model is shown in Figure 6.3. The plot clearly shows the presence of seasonal variation and a decreasing trend from 1998 until the end of the time series. The parameter estimates of the mean model, $\hat{\boldsymbol{\beta}}$, are given in Table 6.2. The parameters of the dependence structure consist of the elements of the matrices $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{\Sigma_\eta}$ and $\boldsymbol{\Sigma_\epsilon}$. Their estimates are listed below (standard errors are shown between brackets).

$$\hat{\boldsymbol{A}} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0.77\ (0.52) & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1.05\ (0.42) & 0.04\ (0.07) & 0 & 0 \\ 0 & 0 & 0 & 0.39\ (0.29) & 0 \end{bmatrix},$$

$$\hat{\boldsymbol{B}} = \begin{bmatrix} 0.14\ (0.7) & 0 & 0 & 0 & 0 \\ 0 & 0.35\ (0.17) & 0 & 0 & 0 \\ 0 & 0 & 0.98\ (0.02) & 0 & 0 \\ 0 & 0 & 0 & -0.14\ (0.16) & 0 \\ 0 & 0 & 0 & 0 & 0.12\ (0.21) \end{bmatrix},$$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}} = \begin{bmatrix} 12\ (13) & & & & \\ 0 & 0.01\ (5.9) & & & \\ 0 & 0 & 1.11\ (1.0) & & \\ 0 & 0 & 0 & 14.2\ (14.9) & \\ 0 & 0 & 0 & 0 & 0.02\ (7.40) \end{bmatrix}$$

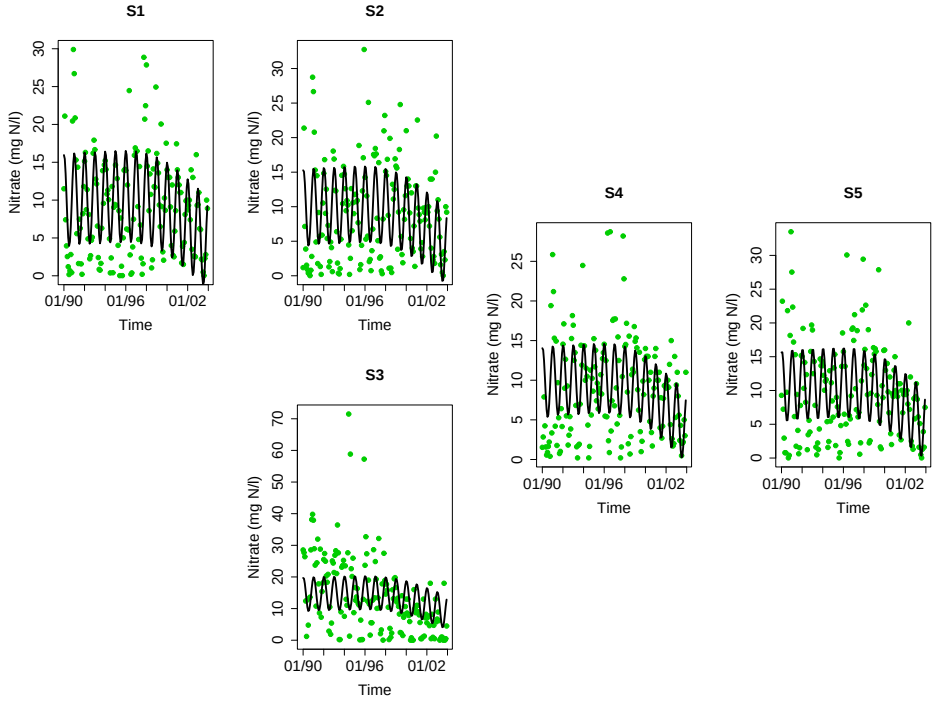

Figure 6.3: Model fit at five sampling locations of the river Yzer according to model II. Sampling locations S1, S2, S4, S5 are located on the main river and sampling location S3 is located on a tributary which drains into the Yzer between S2 and S4.

and

$$\hat{\boldsymbol{\Sigma}}_{\epsilon} = \begin{bmatrix} 12(13.7) & & & & \\ 8.0(6.9) & 21.6(6.2) & & & \\ 10.4(3.7) & 13.5(4.1) & 90.4(10.4) & & \\ 4.9(8.2) & 6.7(6.4) & 4.6(3.8) & 3.2(18.8) & \\ 17.0(4.7) & 18.8(4.3) & 13.6(4.3) & 8.3(10.4) & 26.9(11.6) \end{bmatrix}.$$

The model quality has to be checked by the use of an assessment of the standardised innovations. These innovations should be independent. This is assessed with a plot of the autocorrelation function (ACF). The ACF plot of the original series is given in Figure 6.4. The original nitrate observations are clearly correlated. The ACF of the standardised innovations are given in Figure 6.5. For these plots, we see that the model succeeds in reducing a considerable amount of the serial cor-

Table 6.2: Parameter estimates, standard errors and p-values for the linear part in the mean model (Model II)

Parameter	Estimate	Standard error	p-value
μ	9.85	0.20	0.00
α_4	-0.90	0.14	0.00
α_2	-0.79	0.14	0.00
α_1	-0.66	0.07	0.00
α_3	3.84	10.20	0.71
γ_2	3.70	0.38	0.00
γ_1	3.71	0.39	0.00
$(\alpha\gamma)_{4,2}$	-0.28	0.26	0.29
$(\alpha\gamma)_{2,2}$	0.50	0.25	0.04
$(\alpha\gamma)_{1,2}$	1.23	0.15	0.00
$(\alpha\gamma)_{3,2}$	0.36	1.18	0.76
$(\alpha\gamma)_{4,1}$	-0.92	0.26	0.00
$(\alpha\gamma)_{2,1}$	-0.04	0.25	0.87
$(\alpha\gamma)_{1,1}$	-0.18	0.15	0.23
$(\alpha\gamma)_{3,1}$	-0.23	1.20	0.85

Table 6.3: p-values for the Ljung-Box portmanteau tests of the autocorrelation coefficients of the standardised innovations at the first 5 lags

Lag	S1	S2	S3	S4	S5
1	0.78	0.87	0.63	0.90	0.11
2	0.95	0.97	0.10	0.96	0.27
3	0.20	0.99	0.17	0.74	0.37
4	0.06	1.00	0.14	0.46	0.52
5	0.04	0.99	0.05	0.52	0.58

relation present in the original series. A joint test of significance of the first i autocorrelation coefficients can be provided by the Ljung-Box portmanteau test (Ljung and Box, 1978). The p-values for the Ljung-Box portmanteau tests of the autocorrelation coefficients of the standardised innovations are presented in Table 6.3. In Table 6.3, the test is only significant for S1 at lag 5. From the ACF plots and the Ljung Box tests, we therefore conclude that the AR(1) structure seems to be adequate. The quality of the mean model is checked in a plot of the standardised innovations with respect to time. This graph is displayed in Figure 6.6. Friedman's

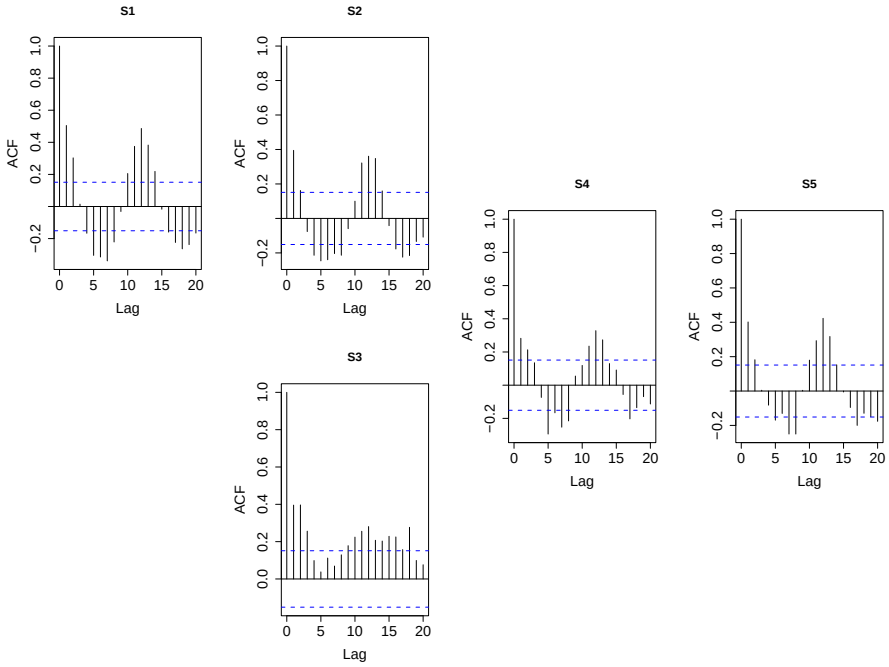

Figure 6.4: Autocorrelation plots of original nitrate series at the different sampling locations

supersmoother (Friedman, 1984) is added to each plot to check whether there is still a pattern present in the standardised innovations. From Figure 6.6 it can be seen that the smoothers stay close to zero. For S2, S3 and S4 the smoothers indicate deviations from zero at the boundaries.This was also observed in the case study in Section 5.4 and again these deviations are probably due to the combination of a boundary effect of the smoother, large nitrate values measured in the beginning of the time series and the Kalman filter which might have not reached steady state yet. Because no severe deviations are indicated by the smoother, the assumption of the existence of a regional trend seems acceptable. All processes were assumed to be Gaussian. The standardised innovations should therefore be distributed according to the standard normal distribution. Hence, most of standardised innovations are expected to lay approximately in the interval $[-2, 2]$. In Figure 6.6 it can be seen that at each sampling location a number of outliers are present. The normality of the standardised innovations is further assessed in Figure 6.7. Both the boxplot and the QQ-plot show a clear departure from normality. The boxplot indicates a consid-

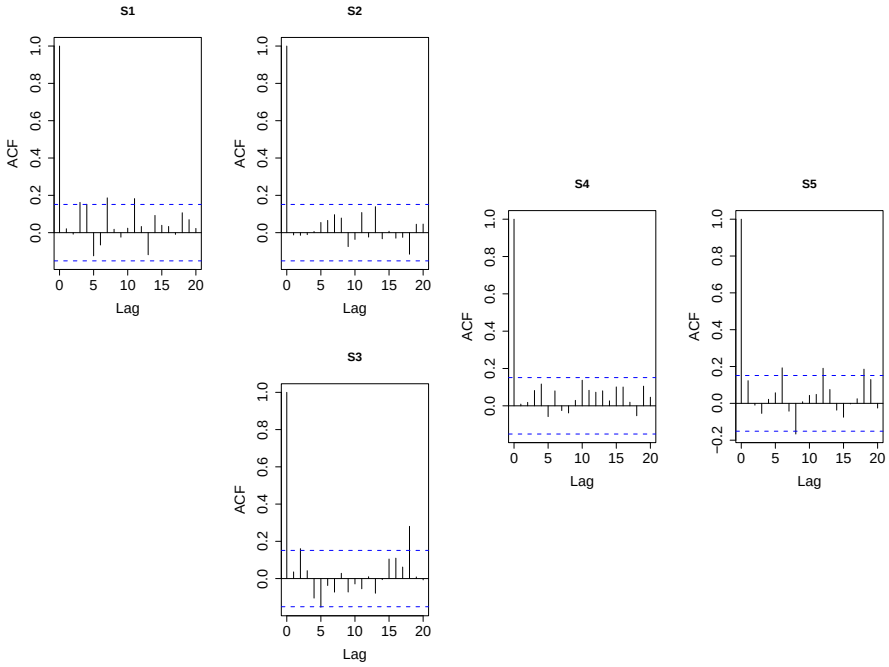

Figure 6.5: Autocorrelation plots for the standardised innovations of Model II

erable amount of outliers and the QQ-plot indicates that the distribution has larger tails than the normal distribution. On the other hand, from all plots it can be seen that the distribution of the standardised innovations is symmetric. For the research question, we need to infer on the mean model included in the observation equation. This is a deterministic component in the model and from Harvey (1989) we know that the asymptotic distribution of the estimators associated with the deterministic components are not affected when the Gaussianity assumption is dropped. Hence, the inference on the parameters of the mean model remains approximately valid.

Thanks to the additive model structure, the contribution of each predictor can be studied individually. This enables us to decompose the model into components that can be represented graphically. The trend and its derivative are shown in Figure 6.8, along with 95% asymptotic pointwise confidence intervals. A naive approach to assess the trend is to perform a t-test at each individual time instant. An equivalent result is obtained by assessing at which time instants zero is not contained in the 95% confidence intervals of the derivatives. In Figure 6.8 pointwise significant

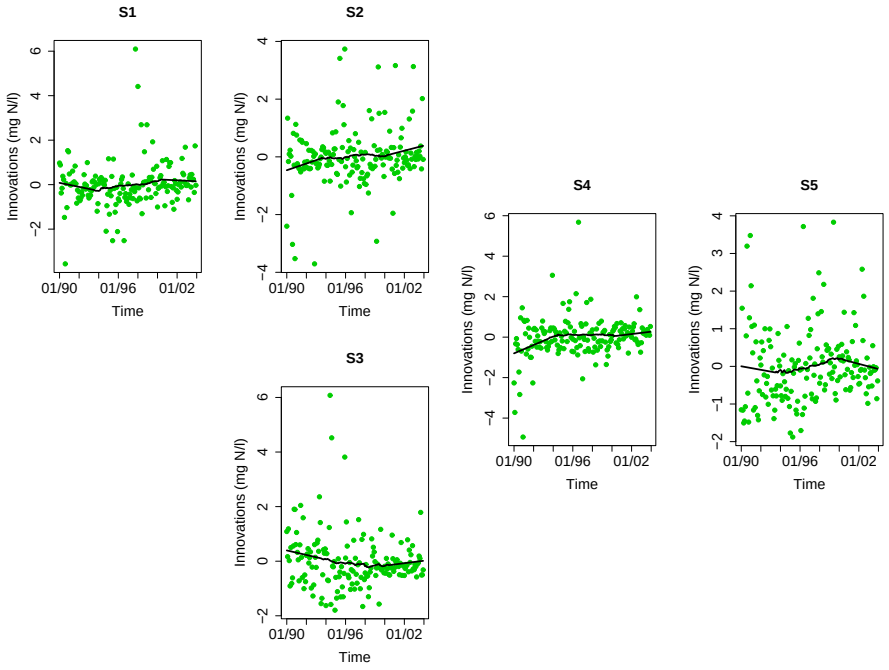

Figure 6.6: Plots of the standardised innovations of Model II, Friedman super-smoothers are added to the plots to assess the residual pattern

results ($\alpha = 0.05$) are indicated with a dot and it can be seen that a trend is present from January 1999 until July 2003.

For the test procedure to be formally valid, a multiplicity correction is needed to control the familywise Type I error at the α-level instead of controlling the Type I error of the individual tests. When the Holm procedure was applied to correct for multiplicity, no significant results were observed (results not shown). The Holm procedure however acts as if all tests are independent. But observations which are close in time are likely to have similar trends. Thus, in our setting, the trend tests are dependent, and this reduces the actual dimension of the multiplicity problem. Therefore we have proposed to use a modified maximum T approach that corrects for multiplicity and takes the dependences between the tests explicitly into account . The results of this approach are illustrated in Figure 6.9. Familywise significant first derivatives ($\alpha = 5\%$) are indicated with dots superimposed on the point estimates. The derivatives of the nonlinear trend are significantly different

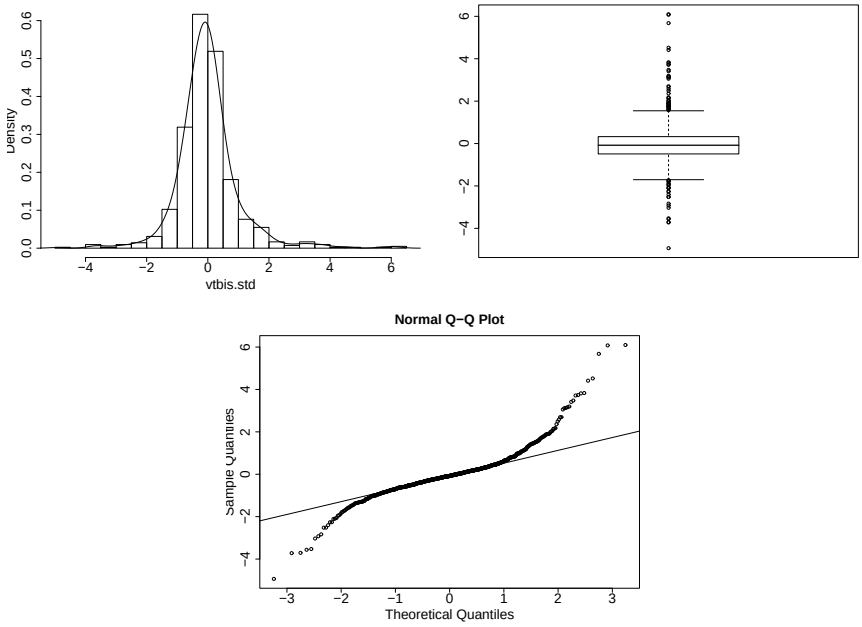

Figure 6.7: Histogram, boxplot and QQplot of the innovations of Model II

from 0 between September 1999 and January 2002. Since the estimates of the first derivatives of the nonlinear trend are negative, a significant decrease in nitrate concentration is concluded for this period. The fact that we exploited the dependences between the t-tests, clearly leads to a less conservative test procedure than classical corrections such as the Holms procedure. Compared to the naive approach, the nonlinear trend is not significant in 2002 and 2003. This is not surprising as the variance of the predictions based on smoothers is typically inflated in the boundary regions.

Although this data analysis methodology has no causal interpretation, it can be concluded that a decreasing trend in the nitrate concentration in the study region is established between the introduction of the first MAP and the second MAP. The trend remains significant until January 2002.

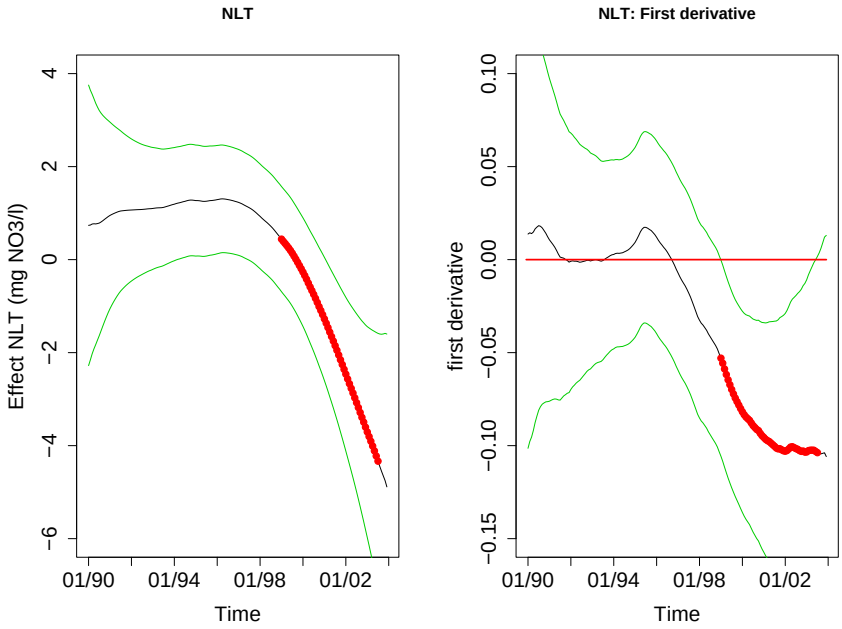

Figure 6.8: Evaluation of the common nonlinear trend (NLT) along the river Yzer. The estimated trend is presented in the left panel, and its first derivative is shown in the right panel. In both graphs, 95% pointwise confidence bands are depicted. Pointwise significant decreases are indicated with a dot superimposed on the point estimates

6.6 Discussion and conclusions

In this chapter a statistical methodology was developed for the detection of nonlinear trends in river monitoring network data. A spatio-temporal model was constructed to model the marginal mean and the dependence structure. According the specification of the marginal mean model, the trend can be studied at the level of individual sampling locations, or on a more regional scale.

In contrast with existing methodologies for (non)linear trend detection, our procedure takes the spatio-temporal dependence explicitly into account. As compared to ad hoc methods such as the methods based on the field significance (e.g. Burn and

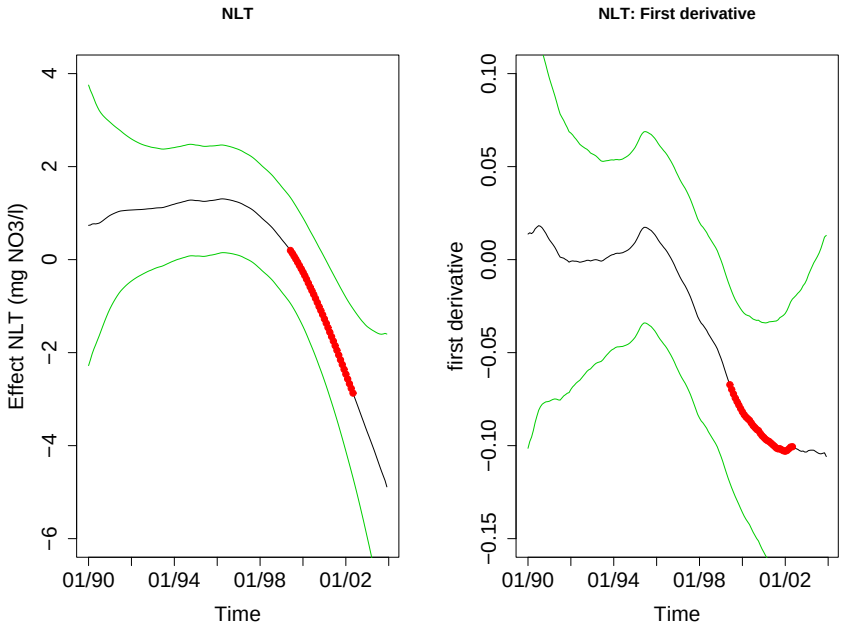

Figure 6.9: Evaluation of the common nonlinear trend (NLT) along the river Yzer. The estimated trend is presented in the left panel, and its first derivative is shown in the right panel. In both graphs, 95% pointwise confidence bands are depicted. Familywise significant decreases are indicated with a dot superimposed on the point estimate

Hag Elnur, 2002), our method provides statistical inference which is formally valid. For the detection of trends in water quality, the use of a nonparametric regression method is more flexible. Classical tests such as Mann Kendall tests for trend detection are not appropriate when sign changes occur in the trend. Our method also enables the detection of trends on a more local time scale. To verify at which time instants the nonlinear trend is beneficial, t-tests are performed at each time instant. Due to the specific dependence between these tests, classical multiplicity corrections are too conservative. We have adopted the free step-down resampling method Westfall and Young (1993) and sampled from an appropriate null distribution to take the dependences between the statistical tests into account.

The methodology has been illustrated in a case study where a significant decrease

in the nitrate concentration was detected in the study region between September 1999 and January 2002 ($\alpha = 0.05$), indicating a beneficial effect of the introduction of the manure action plans.

Chapter 7

Spatio-temporal modelling of river monitoring networks, a binary data approach

7.1 Introduction

The authorities of the member states of the European Union are responsible to develop a long term vision in order to comply with the environmental quality standards imposed by the European environmental legislation. Such standards are commonly expressed in terms of threshold levels. This provides a binary response to the decision maker. In case of nitrate, a value which is below the threshold indicates a good nitrate status, and a value above the threshold indicates that the nitrate

status is problematic. To evaluate and refine their strategy it is important to detect whether their actions have a beneficial effect. From the policy makers point of view it is relevant to assess the impact of management strategies on the violation frequency of water quality standards. This question can directly be assessed by transforming the observations into binary data using the water quality standard as a threshold. In this way the response variable is Bernoulli distributed so that trends in the compliance frequency can be modelled. A beneficial effect of such a transformation is that the statistical tests become distribution free in the sense that no distributional assumptions have to be made concerning the original distribution of the water quality variable. Such an approach is particularly useful when dealing with water quality indicators with a large fraction of censored observations such as for instance heavy metals and pesticides. Censoring of water quality data occurs due to concentrations which are below the detection limit of the measuring method. Although the transformation to binary data reduces the data complexity, the spatio-temporal dependence still remains.

To deal with non-normal data, a generalisation of the model framework used throughout this dissertation is needed. Before we introduce the generalised framework, we will start from the classical linear model to introduce the different components that we will need later on. Let y_{it} be an observation acquired on time t, $t = 1, \ldots, n$, at the i^{th} sampling location, $i = 1 \ldots p$ and let $\boldsymbol{x}_{it}$ be the $1 \times q$ vector of corresponding predictor values $\boldsymbol{x}_{it} = (x_{it,1}, \ldots, x_{it,q})$ that are measured simultaneously. Actually $\boldsymbol{x}_{it}$ is a row from a linear design matrix. Thus, if an intercept is to be included in the model, one of the elements of $\boldsymbol{x}_{it}$ should be set to 1. For the moment we will also assume the y_{it}'s to be i.i.d. normally distributed. The classical linear model can be written as,

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + \epsilon_{it}, \tag{7.1}$$

where the systematic part for the model is specified in terms of a number of parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)$ and can be written as $\mathrm{E}\,(y_{it}|\boldsymbol{x}_{it}) = \mu_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta}$. For the random part, we assume the residuals ϵ_{it} to be i.i.d. normally distributed with zero mean and constant variance σ^2, i.e. $\epsilon_{it} \sim N(0, \sigma^2)$. Thus, the y_{it}'s are normally distributed with mean μ_{it} and variance σ^2.

In many cases this model is not appropriate. An important case is the one in which the y_{it} and μ_{it} are bounded. For example, if the y_{it}'s represent count data, $y_{it} \geq 0$ and $\mu_{it} \geq 0$. In this chapter, y_{it} is considered to be binary. In particular we write $y_{it} = 1$ if the environmental threshold is violated and $y_{it} = 0$ if the water quality variable is below the threshold. Thus, the mean μ_{it} has to be in the interval $0 \leq \mu_{it} \leq 1$. The standard linear model is inadequate in these cases because

complicated and unnatural constraints on $\boldsymbol{\beta}$ would be required to make sure that μ_{it} stays in the range. McCullagh and Nelder (1989) give an extensive overview of *generalised linear models* that can be used for this purpose. To make the transition to generalised linear models more easy, we will rewrite Equation (7.1) to produce a three-part specification:

1. The *random component*: the y_{it}'s are independently normally distributed with mean μ_{it} and constant variance σ^2,

$$y_{it} \sim N(\mu_{it}, \sigma^2). \tag{7.2}$$

2. The *systematic component*: covariates $\boldsymbol{x}_{it}$ produce a *linear predictor* η_{it} given by

$$\eta_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta}. \tag{7.3}$$

3. The *link* between the random and systematic components:

$$\eta_{it} = \mu_{it}. \tag{7.4}$$

In doing so, we have introduced a new notation η_{it} for the linear predictor and a third component that specifies that μ_{it} and η_{it} are identical. We can also write the link more generally as

$$\eta_{it} = g(\mu_{it}), \tag{7.5}$$

where $g(.)$ is referred to as the *link function*. Classical linear models use a normal distribution for component 1 and the identity link function for component 3. Generalised linear models extend classical linear models by allowing a different distribution for component 1 and by using another monotonic differentiable function for the link function in component 3. Recall the constraints for count data $y_{it} \geq 0$ and $\mu_{it} \geq 0$ and for binary data $y_{it} = 1$ or $y_{it} = 0$ and $0 \leq \mu_{it} \leq 0$. For these cases, $g(.)$ will be used to transform the μ_{it} to a scale on which they are unconstrained. For example we may use $g(\mu_{it}) = \log(\mu_{it})$ if $\mu_{it} \geq 0$ or $g(\mu_{it}) = \text{logit}(\mu_{it}) = \log[\mu_{it}/(1 - \mu_{it})]$ if $0 \leq \mu_{it} \leq 1$. Other link functions are also possible, e.g. the probit link can be used for Bernoulli data instead of the logit link. The probit link is the inverse of the cumulative standard normal distribution function. Further, the distribution in component 1 becomes the Poisson distribution for count data and the Bernoulli distribution for binary observations. The usual restriction on component 1, is that this distribution should belong to the exponential family.

So far we have considered the observations to be i.i.d. Observations originating from a river monitoring network data, however, are not independent. Another extension is therefore needed to incorporate the dependence structure. A common extension to model dependent outcomes, is to include random terms in the linear predictor. Such models are than classified as *generalised linear mixed models* (GLMM's, e.g. Breslow and Clayton, 1993). It is often a reasonable approximation to assume that the random error terms are distributed according to a normal distribution. Although a full maximum likelihood analysis is possible, it usually involves irreducible high-dimensional integrals (Breslow and Clayton, 1993). Therefore a number of approximation methods have been developed to deal with GLMM's. Depending on the research question, different approaches are possible. When one is interested in the marginal mean, Marginal Quasi likelihood (MQL) can be used (Breslow and Clayton, 1993). In case the dependence structure can be assumed to have a block diagonal structure, the MQL can be optimised by the use of general estimation equations (e.g. Liang and Zeger, 1986, Zeger and Liang, 1986, Zeger et al., 1988 and Breslow and Clayton, 1993). If one is interested in the parameters of the mean model conditional on the random effects, penalised quasilikelihood (PQL) can be adopted (Breslow and Clayton, 1993). In this dissertation, we infer on the marginal mean. However, the approximations which are commonly made to apply MQL do not hold, e.g. the data at the sampling locations of a river network are not mutually independent and thus their dependence structure cannot be written as a block diagonal structure. Hence, MQL cannot be used directly. For the inference procedure to be formally valid we have therefore chosen to work within a full Bayesian framework. A short introduction to this statistical framework is given in Section 7.3.1.

In this chapter, a first onset is given towards the generalisation of the spatio-temporal models presented in Chapters 5 & 6. In particular, a logistic state space model for the probability of violating a threshold is presented. This model explicitly incorporates the dependence structure of the data. It uses a mean model to assess the impact of the introduction of a manure action plan (MAP) on the nitrate concentration and to correct for the seasonal variation. The formulation of the mean model allows the assessment to be done at the level of individual sampling locations or on a more regional scale. The dependence structure is introduced by the use of a latent variable S and temporal dependence is assumed to be behave as an AR(1) process. These assumptions have to be checked afterwards. Similar to the previous Chapters 5 & 6, the spatial dependence of the latent variable is assumed to be a branched unidirectional structure that can be represented as a Directed Acyclic Graph (DAG).

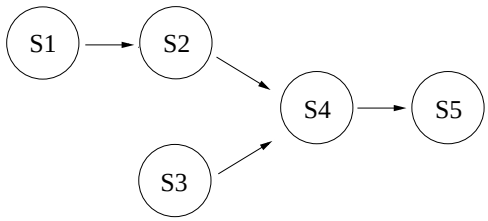

Figure 7.1: Directed Acyclic Graph (DAG) of five sampling locations along two joining river reaches

7.2 Spatio-temporal model

First the spatial dependence structure is derived in Section 7.2.1. In Section 7.2.2 this model is extended to include a temporal structure. Finally, the mean model is introduced in Section 7.2.3.

7.2.1 Spatial dependence structure

Let the $p \times 1$ vector $\boldsymbol{S} = (S_1, \ldots, S_p)^T$ denote a stationary spatial process, where S_i $(i = 1, ..., p)$ represents the response variable at sampling location i. The correlation structure of $\boldsymbol{S}$ is completely defined by the river monitoring network topology. This is illustrated in Figure 7.1 which shows 5 sampling locations along 2 joining river reaches. The direction of the flow is also indicated and it can also be interpreted as a Directed Acyclic Graph (DAG) (see e.g. Whittaker, 1990) in which the circles represent the graph's vertices associated with the corresponding S_i's. Missing edges or arrows indicate the conditional independences. Thus from Figure 7.1 we read $S_1 \perp\!\!\!\perp S_3$; $S_2 \perp\!\!\!\perp S_3$; $S_4 \perp\!\!\!\perp S_1 | S_2$; $S_5 \perp\!\!\!\perp S_1 | S_2$; $S_5 \perp\!\!\!\perp S_1 | S_4$; $S_5 \perp\!\!\!\perp S_2 | S_4$ and $S_5 \perp\!\!\!\perp S_3 | S_4$. The DAG implies zeroes in the variance-covariance matrix of $\boldsymbol{S}$. Thus it can equivalently be represented by a recursive system of equations (Wermuth, 1980),

$$\boldsymbol{S} = \boldsymbol{A}\boldsymbol{S} + \boldsymbol{\gamma}, \tag{7.6}$$

where the order of the elements of $\boldsymbol{S}$ can always be rearranged so that $\boldsymbol{A}$ is a lower triangular square matrix with zeroes at the diagonal, and $\boldsymbol{\gamma}$ is a multivariate zero-mean random vector with a diagonal variance-covariance matrix $\boldsymbol{\Sigma}_\gamma$. We further assume that $\boldsymbol{\gamma} \sim MVN(\boldsymbol{0}, \boldsymbol{\Sigma}_\gamma)$. For the DAG represented in Figure 7.1, $\boldsymbol{A}$

becomes

$$\boldsymbol{A} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ a_{21} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & a_{42} & a_{43} & 0 & 0 \\ 0 & 0 & 0 & a_{54} & 0 \end{bmatrix}$$

where a_{ij} models the dependence between sampling location S_i and S_j.

7.2.2 Spatio-temporal dependence structure

In a river monitoring network the data are gathered over time. Vector $\boldsymbol{S_t} = (S_{1t}, \ldots, S_{pt})^T$ now represents the observations at the sampling locations at time t with $t = 1, \ldots, n$. A Markovian structure is assumed for the temporal dependence. The quality of the temporal model has to be assessed through a residual analysis. To incorporate the temporal dependence structure, Equation (7.6) is extended to

$$\boldsymbol{S}_t = \boldsymbol{A}\boldsymbol{S}_t + \boldsymbol{B}\boldsymbol{S}_{t-1} + \boldsymbol{\eta}_t, \tag{7.7}$$

where $\boldsymbol{B}$ is a matrix containing the temporal autocorrelation coefficients (diagonal elements) and the spatio-temporal cross-correlation coefficients (off-diagonal elements), and $\boldsymbol{\eta}_t \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_\eta)$ with a diagonal variance-covariance matrix $\boldsymbol{\Sigma}_\eta$. Similar to matrix $\boldsymbol{A}$, we propose to only use cross-correlations between sampling locations which are directly connected according to the DAG structure. The off-diagonal elements of $\boldsymbol{B}$ are thus structured in a similar way as the elements of matrix $\boldsymbol{A}$. Hence $\boldsymbol{B}$ can be written as

$$\boldsymbol{B} = \begin{bmatrix} b_{11} & 0 & 0 & 0 & 0 \\ b_{21} & b_{22} & 0 & 0 & 0 \\ 0 & 0 & b_{33} & 0 & 0 \\ 0 & b_{42} & b_{43} & b_{44} & 0 \\ 0 & 0 & 0 & b_{54} & b_{55} \end{bmatrix}.$$

For $i \neq j$ the b_{ij} model the spatio-temporal dependence between S_{it} and S_{jt-1} and the b_{ii} model the temporal dependence between S_{it} and S_{it-1}.

Equation (7.7) can be reorganised so that the model can be written in its general state-space model representation,

$$\boldsymbol{S}_t = \boldsymbol{\Phi}\boldsymbol{S}_{t-1} + \boldsymbol{\delta}_t, \tag{7.8}$$

where $\boldsymbol{\Phi} = (\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{B}$ and $\boldsymbol{\delta}_t \sim MVN(\boldsymbol{0}, \boldsymbol{Q})$ with covariance matrix $\boldsymbol{Q} = (\boldsymbol{I}-\boldsymbol{A})^{-1}\boldsymbol{\Sigma}_\eta(\boldsymbol{I}-\boldsymbol{A})^{-T}$ and $t = 1, \ldots, n$. For the model to be completely defined, we assume $\boldsymbol{S}_0$ to be multivariate normally distributed, $\boldsymbol{S}_0 \sim MVN(\boldsymbol{0}, \boldsymbol{\Sigma}_{S_0})$.

Alternatively, the following notation can be used,

$$\boldsymbol{C}\boldsymbol{S}_N = \boldsymbol{\zeta}, \tag{7.9}$$

where $\boldsymbol{S}_N = (\boldsymbol{S}_1^T, \ldots, \boldsymbol{S}_n^T)^T$, $\boldsymbol{C}$ is a $pn \times pn$ square matrix constructed from the elements of the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{\zeta} \sim MVN(\boldsymbol{0}, \boldsymbol{\Sigma}_\zeta)$, where $\boldsymbol{\Sigma}_\zeta$ is a diagonal matrix built from the corresponding elements of $\boldsymbol{\Sigma}_\eta$. Hence, $\boldsymbol{S}_N$ is multivariate normally distributed with a zero mean and a covariance matrix $\boldsymbol{\Sigma}_{S_N}$ given by

$$\boldsymbol{\Sigma}_{S_N} = \boldsymbol{C}^{-1}\boldsymbol{\Sigma}_\zeta\boldsymbol{C}^{-T}. \tag{7.10}$$

7.2.3 Mean model and formulation of the GLMM

The latent process $\boldsymbol{S}_t$ cannot be observed. Instead a variable $\boldsymbol{y}_t$ is observed that indicates whether a certain water quality standard is violated or not. Hence, $\boldsymbol{y}_t$ gives a binary response and it is coded to be 1 in case of violation and 0 otherwise. The y_{it}'s $i = 1, \ldots, p$ and $t = 1, \ldots, t$ are believed to be independent conditional on a number of explanatory variables and on the latent spatio-temporal process S_{it}. Its conditional distribution is assumed to be Bernoulli. In the GLMM framework the model can be written as follows:

1. Random component: the y_{it} are assumed to be Bernoulli conditional on the predictors $\boldsymbol{x}_{it}$ and the latent spatio-temporal process S_{it}. Their conditional mean is given by
$$\mathrm{E}\left(y_{it}|S_{it}, \boldsymbol{x}_{it}\right) = \mu_{it}^c. \tag{7.11}$$

2. Systematic component: predictors $\boldsymbol{x}_{it}$ and the latent spatio-temporal process S_{it} produce the linear predictor ν_{it}^c given by
$$\nu_{it}^c = \boldsymbol{x}_{it}\boldsymbol{\beta}^c + S_{it} \tag{7.12}$$

3. Link between random and systematic components:
$$\nu_{it}^c = g(\mu_{it}^c) \tag{7.13}$$

4. The random effects are given by the spatio-temporal latent process $\boldsymbol{S}_N = (\boldsymbol{S}_1, \ldots, \boldsymbol{S}_n)$, which are multivariate normally distributed

$$\boldsymbol{S}_N \sim MVN(\boldsymbol{0}, \boldsymbol{\Sigma}_{S_N}) \tag{7.14}$$

When this model would be used for inference, the parameters $\boldsymbol{\beta}^c$ have an interpretation conditional on the latent process S. In an environmental context, however, we want to infer on the marginal mean. Via integration over the latent variable, every conditional model implies a marginal model (e.g. Heagerty and Zeger, 2000 and Griswold and Zeger, 2004),

$$\mu_{it}^m = \mathrm{E}\left(y_{it}\right) = \mathrm{E}_S(\mathrm{E}\left(y_{it}|S_{it}\right)) = \mathrm{E}_S(\mu_{it}^c). \tag{7.15}$$

This marginal mean μ_{it}^m is now further linked to a linear predictor $\nu_{it}^m = \boldsymbol{x}_{it}\boldsymbol{\beta}^m$ by $\nu_{it}^m = g(\mu_{it}^m)$, where the link function $g(.)$ is defined as before and $\boldsymbol{\beta}^m$ represents the parameter vector with the correct marginal interpretation. Fitting marginal models, however, usually involves the application of approximation methods such as the use of generalised estimation equations (e.g. Liang and Zeger, 1986, Zeger and Liang, 1986, Zeger et al., 1988 and Breslow and Clayton, 1993). The approximations which are commonly used, do not hold here because the variance covariance structure of the observations is not block diagonal. To enable a full likelihood based inference procedure for marginal models, Heagerty and Zeger (2000), Heagerty (2002) and Griswold and Zeger (2004) formulated a marginalised version of the GLMM model:

1. Random components: the marginal mean of the y_{it} conditional on the predictors $\boldsymbol{x}_{it}$ is given by

$$\mathrm{E}\left(y_{it}|\boldsymbol{x}_{it}\right) = \mu_{it}^m. \tag{7.16}$$

   The y_{it} are assumed to be Bernoulli conditional on the predictors $\boldsymbol{x}_{it}$ and the latent spatio-temporal process S_{it}.

$$\mathrm{E}\left(y_{it}|\boldsymbol{x}_{it}, S_{it}\right) = \mu_{it}^c. \tag{7.17}$$

2. Systematic components: the predictors $\boldsymbol{x}_{it}$ produce the linear predictor ν_{it}^m for the marginal component given by

$$\nu_{it}^m = \boldsymbol{x}_{it}\boldsymbol{\beta} \tag{7.18}$$

   The predictors $\boldsymbol{x}_{it}$ and the latent spatio-temporal process S_{it} produce the predictor ν_{it}^c for the conditional component given by

$$\nu_{it}^c = \Delta_{it} + S_{it} \tag{7.19}$$

where Δ_{ij} forms a mapping between the conditional and marginal model components.

3. Link between random and systematic components:

$$\nu_{it}^m = g(\mu_{it}^m) \tag{7.20}$$
$$\nu_{it}^c = g(\mu_{it}^c) \tag{7.21}$$

4. The random effects are given by the spatio-temporal latent process $\boldsymbol{S}_N = (\boldsymbol{S}_1, \ldots, \boldsymbol{S}_n)$, which are multivariate normally distributed

$$\boldsymbol{S}_N \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_{S_N}). \tag{7.22}$$

Let $h(.)$ be defined as the inverse of the link function $h(.) = g^{-1}(.)$. From Equation (7.15), it can be seen that Δ_{it} can be found as the solution to the integral

$$h(\nu_{it}^m) = \int_{\Re^p} h(\Delta_{it} + S_{it}) dP(\boldsymbol{S}_N), \tag{7.23}$$

where $P(\boldsymbol{S}_N)$ is the probability distribution of $\boldsymbol{S}_N$. When the probit link is used, Heagerty and Zeger (2000) and Griswold and Zeger (2004) have shown that

$$\Delta_{ij} = \sqrt{1 + S_{it}^2} \boldsymbol{x_{it}} \boldsymbol{\beta}^m. \tag{7.24}$$

Hence, they identified a conditional model structure that induces the marginal model of interest. Once this particular conditional model is known, the estimation of the desired marginal model only involves the estimation of this conditional model.

In this chapter, these GLMM's are estimated within the Bayesian framework. This statistical framework is briefly introduced in the next section.

7.3 Parameter estimation and Bayesian inference

First a very brief introduction to the Bayesian paradigm is given. The section then continues with some practical considerations on how to fit a Bayesian model by using Markov Chain Monte Carlo.

7.3.1 Introduction to Bayesian inference

Most of this section is taken from Gilks et al. (1996b). In the previous chapters we worked within the frequentistic framework where the observations are considered to be realisations of random variables and the model parameters are assumed to be fixed but unknown. In the Bayesian framework, however, no fundamental distinction is made between the observed random variables and the parameters of a statistical model: they are all considered as random quantities and they are also referred to as nodes. Let $\boldsymbol{D}$ denote the observed data, and $\boldsymbol{\theta}$ the model parameters. Then inference is provided by setting up a joint probability distribution $P(\boldsymbol{D}, \boldsymbol{\theta})$ over all random quantities. Let $P(\boldsymbol{\theta})$ denote the prior distribution on the model parameters. The set $\boldsymbol{\Theta}$ denotes the support of $\boldsymbol{\theta}$ and $P(\boldsymbol{D}|\boldsymbol{\theta})$ denotes the traditional likelihood function. Then the joint probability becomes

$$P(\boldsymbol{D}, \boldsymbol{\theta}) = P(\boldsymbol{D}|\boldsymbol{\theta})P(\boldsymbol{\theta}). \tag{7.25}$$

Once $\boldsymbol{D}$ is observed, Bayes theorem can be used to derive the distribution of $\boldsymbol{\theta}$ conditional on $\boldsymbol{D}$:

$$P(\boldsymbol{\theta}|\boldsymbol{D}) = \frac{P(\boldsymbol{D}, \boldsymbol{\theta})}{P(\boldsymbol{D})} = \frac{P(\boldsymbol{D}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{\int\limits_{\boldsymbol{\Theta}} P(\boldsymbol{D}|\boldsymbol{\theta})P(\boldsymbol{\theta})d\boldsymbol{\theta}}, \tag{7.26}$$

which is also referred to as the posterior distribution of $\boldsymbol{\theta}$. For inference, features such as moments, quantiles and credibility intervals of the posterior distribution can be used. A *credibility interval* is the Bayesian counterpart of a confidence interval in the frequentistic setting, however their interpretation is different. Bayesian inference treats parameters as random variables and therefore a 95% credibility interval on a certain parameter β means that 95% of the potential values of β will fall within the boundaries of the credibility interval.

In general, the statistic of interest is a function of $\boldsymbol{\theta}$. The posterior expectation of a function $f(\boldsymbol{\theta})$ is given by

$$\mathrm{E}\left(f(\boldsymbol{\theta})|\boldsymbol{D}\right) = \frac{\int\limits_{\boldsymbol{\Theta}} f(\boldsymbol{\theta})P(\boldsymbol{D}|\boldsymbol{\theta})P(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int\limits_{\boldsymbol{\Theta}} P(\boldsymbol{D}|\boldsymbol{\theta})P(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{7.27}$$

Analytical solutions of these integrations do often not exist. Numerical methods have therefore to be used. An example of such a technique is the use of Monte Carlo methods. A Monte Carlo algorithm evaluates $\mathrm{E}\left(f(\boldsymbol{\theta})|\boldsymbol{D}\right)$ by drawing samples $\boldsymbol{\theta}_k$, $k = 1, \ldots, m$ from $P(\boldsymbol{\theta}|\boldsymbol{D})$ and it approximates the expected value

E $(f(\boldsymbol{\theta})|\boldsymbol{D})$ by $1/m \sum_{k=1}^{m} f(\boldsymbol{\theta}_k)$. In general, it is not feasible to draw the samples independently. Fortunately, for the Monte Carlo approximation to hold, the $\boldsymbol{\theta}_k$ do not have to be independent as long as they are drawn from the support of $P(\boldsymbol{\theta}|\boldsymbol{D})$ in the correct proportions. This can be done by the use of a Markov chain by sampling the next state $\boldsymbol{\theta}_{k+1}$ from the conditional distribution $P(\boldsymbol{\theta}_{k+1}|\boldsymbol{\theta}_k, \boldsymbol{D})$, under the restriction that the Markov chain has $P(\boldsymbol{\theta}|\boldsymbol{D})$ as its stationary distribution. Such an approach is called *Markov Chain Monte Carlo* (MCMC). An introduction to Markov Chain Monte Carlo is beyond the scope of this dissertation and interested readers find a good introduction in Gilks et al. (1996a).

7.3.2 Fitting a model using MCMC

Most of this section is taken from Spiegelhalter et al. (1996). When one wants to use MCMC to fit a model, several steps are needed

1. Provide starting values of all unobserved quantities (parameters, latent variables and missing data)
2. Construct the full conditional distribution for each node
3. Draw k samples with the MCMC algorithm
4. Monitor the output to establish the total run length and the length of the *burn-in* number, which is the number of iterations needed before the Markov Chain converged to the stationary posterior distribution
5. Repeat steps 3 - 4 until the total run length has been reached
6. Calculate summary statistics of the quantities of interest for inference about the true values of the parameters
7. Assess the quality of the model

In principle the initialisation in step 1 is not that important since the chain must be run long enough “to forget” its starting values. However, extreme starting values can lead to a very long burn-in, or can make the sampler to fail to converge to the main support of the posterior distribution.

Step 2 can be carried out analytically or by dedicated the model in specific software, such as e.g. BUGS (http://mathstat.helsinki.fi/openbugs/) or JAGS (http://www-fis.iarc.fr/ martyn/software/jags/).

In step 3 the output of the MCMC sampler should be assessed to check for mixing and convergence. This can be done by plotting the evolution of the MCMC chain for each of the parameters. When parallel chains are used they overlap when convergence is reached. In case parallel chains are simulated, the Gelman and Rubin (1992) statistic (GR-statistic) can also be used for this purpose (Gelman, 1996). For each parameter, these chains can be used to provide a pooled estimate of its variance. The GR-statistic estimates the potential scale reduction in the pooled estimate of variance which could be reached if the chain would be continued until infinity. As the simulation continues, this estimate becomes closer to one, indicating that the chains are overlapping. The GR-statistic is implemented in the CODA package of R (Plummer et al., 2004). This package provides a point estimate and a 97.5% percentile for the GR-statistic. If the point estimate and the 97.5% points are near to 1, this indicates that a reasonable convergence is reached for the assessed parameter.

To assess the quality of fit of a binary response regression model, an analysis of the residuals $r_{it} = y_{it} - \mu_{it}$ is suggested by Albert and Chib (1993). In a Bayesian analysis, they have a continuous posterior distribution which can give information about outliers (Albert and Chib, 1995). The residuals can be obtained at each iteration. If the posterior distribution of μ_{it} is in conflict with the observed value of y_{it}, then the posterior distribution of r_{it} will be concentrated towards extreme values (Albert and Chib, 1995). For Bernoulli distributed data, the support of r_{it} is in the interval $[y_{it} - 1, y_{it}]$. Hence, an observation $y_{it} = 0$ is unusual if the posterior distribution of r_{it} is located close to the value -1, and an observation $y_{it} = 1$ is considered as an outlier if the posterior of r_{it} is concentrated towards the endpoint 1.

7.4 Case study

A dominant problem in Flemish water bodies is the eutrophication due to nutrient pollution. A considerable nutrient load originates from agricultural activities. One of the major actions to restrict the nutrient pollution from agriculture was the introduction of two Manure Action Plans (MAP's)(Vlaams Parlement, 1995, 1999).

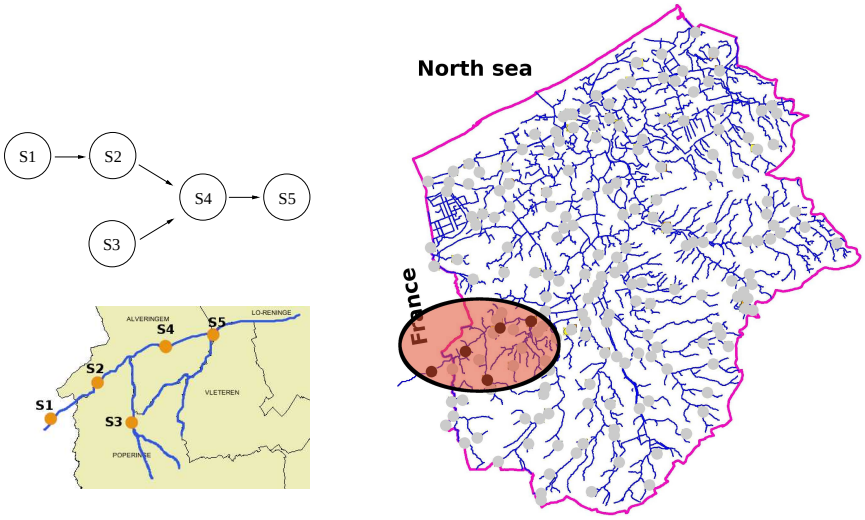

Figure 7.2: Top Left: Directed Acyclic Graph (DAG) of the sampling locations. Bottom Left: Map of the river reaches considered in this case study. Locations S1, S2, S4 and S5 are located on the Yzer river while location S3 is located on a joining creek. Right: Map of the part of the Yzer catchment located in Flanders, Belgium. The sampling locations are indicated by the dots. The area considered in this study is indicated with the ellipse and the black dots are the sampling locations included in this study

Such a MAP restricts the amount of fertilisers that can be used by farmers in areas which are susceptible to eutrophication. The first MAP (MAPI) was introduced in 1996 (Vlaams Parlement, 1995) and after an evaluation a new and more restrictive MAP (MAPII) was implemented in 2000 (Vlaams Parlement, 1999). The aim of this case study is to assess whether the introduction of these MAP's had an effect on the violation frequency of the nitrate standard of 11.3 mg N/l.

The data of 5 sampling locations of the physico-chemical monitoring network of the Flemish surface waters are used. They are located along 2 joining reaches in the Yzer catchment. Their DAG and location in the catchment is indicated on the map in Figure 7.2. Sampling locations S1, S2, S4 and S5 are located on the Yzer while sampling location S3 is located on a joining creek. Every sampling location is monitored on a monthly basis. Data between 1990 and 2003 are available. Hence the number of time instants at which a sample was taken is $n = 168$ and the entire

dataset consists of 840 observations in total.

The observations are taken at time intervals that are much larger than the timescale of the water flow. Therefore we can assume the matrix $\boldsymbol{B}$, used to describe the temporal correlation, to be diagonal. Hence, we only model the temporal auto-correlations for a particular state S_{it} at time t and not the spatio-temporal cross-correlations between S_{it} and its parents in the DAG $\boldsymbol{S}_{t-1}^{[a_i]}$ at time $t-1$. This leads to the reduction of the parent set $[b_i]$ to $[b_i] = i$, containing only the current sampling location. The nitrate series are transformed into a binary response by the use of the nitrate threshold of 11.3 mg N/l. In particular the response is 1 if the nitrate concentration is above the threshold and zero when the nitrate concentration is below the threshold. Seasonal variation is typically present in water quality data and the model has to account for it.

A linear model is used to assess the impact of the introduction of MAPI and MAPII on the trend in the violation frequency of this nitrate standard. The model also has to account for seasonal variation. The presence of seasonal variation in the nitrate series was clearly illustrated in Figure 1.6 where nitrate data of all years was plotted in function of the day of the year. A common approach to deal with this variation is to include sinusoidal functions of fixed periods to describe the seasonal cycle within a year (e.g. Hirst, 1998, Cai and Tiwari, 2000, McMullan et al., 2003 and McMullan, 2004). A function which is often used for this purpose is $\alpha\cos(2\pi(t/P)+\theta)$, where P is the period which is taken to be one year, α is the amplitude of the seasonal trend and θ is a parameter to allow for a phase shift. Hence, α and θ have to be estimated. This function, however, is nonlinear in the parameter θ because the parameter is appears within the cosine function. However, it can be expressed in a linear form by using standard trigonometric expansion of the cosine term. This is the parameterisation of our choice and therefore we use Fourier basis functions to model the seasonal effect. They have a period of one year $(\gamma_1\sin(2\pi t/12)+\gamma_2\cos(2\pi t/12))$. To answer the research question, the following model is considered for the linear predictor corresponding to the marginal mean:

$$g(\mu_{it}^m) = \nu_{it}^m = \alpha_0 + \alpha_i + \beta_1 t + \beta_2 t_{MAPI} + \beta_3 t_{MAPII} + \gamma_1 \sin(\frac{2\pi t}{12}) + \gamma_2 \cos(\frac{2\pi t}{12}) \quad (7.28)$$

where $t = 1 \ldots n$, $g(.)$ is the probit link, α_0 is the effect of sampling location 5, and α_i is the effect for the i^{th} sampling location relative to sampling location 5 (hence $\alpha_5 = 0$), β_1 is the effect of the long term trend, β_2 is the trend change due to the introduction of the first MAP, t_{MAPI} indicates the time since the introduction of the

first MAP where $t_{MAPI} = 0$ for $t \leq 72$ and $t_{MAPI} = t - 72$ for $t > 72$, β_3 is the trend change due to the introduction of the second MAP, t_{MAPII} is the time since the introduction of MAP II and $t_{MAPII} = 0$ for $t \leq 120$ and $t_{MAPII} = t - 120$ for $t > 120$, and γ_1 and γ_1 are the parameters for the seasonal component modelled by the Fourier terms. The formulation of the mean model thus enables the trend to change at 1996 and 2000 when MAP I and MAP II were implemented, respectively. Note that the parameters β_2 and β_3 do not depend on the sampling location. This enables inference on a regional scale, but this restrictive model assumption must be assessed by using diagnostics on the fitted model.

To estimate the marginal model, we need to identify the conditional structure that induces the marginal model of interest. Let us first rewrite the marginal linear predictor as $\nu_{it}^m = \boldsymbol{x}_{it}\boldsymbol{\beta^m}$. In the case study, the probit link is used. From Equation (7.24) we know that the function Δ_{it} that connects the marginal model part to the conditional model part then becomes $\Delta_{ij} = \sqrt{1 + S_{it}^2}\boldsymbol{x_{it}}\boldsymbol{\beta}^m$ (Griswold and Zeger, 2004). From the model formulation (7.17)-(7.22) it can be deduced that the following GLMM has to be implemented to obtain the posterior distributions of the parameters $\boldsymbol{\beta}^m$ of the marginal model,

$$\mathrm{E}\left(y_{it}|\boldsymbol{x}_{it}, S_{it}\right) = \mu_{it}^c \tag{7.29}$$

$$y_{it}|S_{it}, \boldsymbol{x}_{it} \sim \text{Bernoulli}(\mu_{it}^c) \tag{7.30}$$

$$\nu_{it}^c = (\sqrt{1 + S_{it}^2})\boldsymbol{x_{it}}\boldsymbol{\beta}^m + S_{it} \tag{7.31}$$

$$\nu_{it}^c = g(\mu_{it}^c) \tag{7.32}$$

$$\boldsymbol{S}_N \sim MVN(\boldsymbol{0}, \boldsymbol{\Sigma}_{S_N}). \tag{7.33}$$

This GLMM was implemented in the JAGS software. Uniform priors were used for all parameters. Their supports are given in Table 7.1. The specification of the prior distributions on the parameters of the mean model enables the contribution of each term in the mean model to be in the interval $[-8, 8]$ on the probit scale. Two parallel chains were used in the MCMC. In the first chain the parameters of the mean model and the latent variable were set 0, the spatio-temporal in $\boldsymbol{A}$ and $\boldsymbol{B}$ were set 0.5 and the variances $\sigma_{\eta,ii}^2$ were set 1. The second chain was initialised by (1) setting the parameters of the mean model at the estimates obtained by a GLM-fit, (2) using the values obtained in the case study of Chapter 6 to initialise the values of latent variable and the spatio-temporal dependence structure.

30000 iterations were used as burn in and another 120000 iterations were used to approximate the posterior distributions of the parameters. Diagnostic plots to

Table 7.1: Support of the uniform priors on the model parameters

Parameter	Supports
α_i's	$[-8, 8]$
β_1	$[-0.05, 0.05]$
β_2	$[-0.08, 0.08]$
β_3	$[-0.16, 0.16]$
γ_1 and γ_2	$[-8, 8]$
Precisions $1/\sigma^2_{\eta,ii}$	$[0.005, 1]$
Spatial parameters in $\boldsymbol{A}$	$[-0.99, 0.99]$
Temporal parameters in matrix $\boldsymbol{B}$	$[-0.99, 0.99]$

assess the model quality and plots to assess the convergence of the MCMC algorithm can be found in the Appendix of this chapter. For all parameters both chains are shown to be clearly overlapping. The GR-statistic was used to check whether both chains had converged. The point estimates and the 97.5 percentiles of the test statistics are given in Table 7.2, which shows that they are all close to 1. This indicates that both chains converged (Gelman, 1996).

Summary statistics for all parameters in the model are given in Table 7.3. For each parameter a 95% credibility interval is given. There is strong evidence for an effect if zero is not included in the credibility interval. Note that there is not much evidence in favour of temporal correlation (parameters b_{ij}) and that there is a strong evidence in favour of a positive spatial correlation between sampling locations in the main river (parameters a_{ij}). The evolution of the marginal mean of the violation probability and corresponding 95% credibility intervals are shown in Figure 7.3. The plot indicates a seasonal pattern and it also seems that the probability of violation is decreasing in the most recent years. Since the credibility intervals of the parameters γ_1 and γ_2 are above zero, these parameters are very likely to be positive. The contribution of the seasonal effect to ν^m_{it} is represented in Figure 7.4. The plot is obtained by using the posterior means of the parameters γ_1 and γ_2. The contribution of the seasonal effect is positive in winter and negative in summer indicating that there is a higher probability of violation during the winter period. This could be expected because the run-off of the soluble nitrate is typically much higher during the wet winter period.

To infer on the effect of both MAP's, the 95% credibility intervals of β_2 and β_3 have to be assessed, they are $[-0.03, 0.01]$ and $[-0.06, -0.004]$, respectively. Hence,

Table 7.2: GR-statistics of the parameters

Parameter	Point est.	97.5% quantile
α_1	1.02	1.09
α_2	1.00	1.00
α_3	1.00	1.00
α_4	1.00	1.00
α_4	1.00	1.00
β_1	1.02	1.09
β_2	1.02	1.03
β_3	1.02	1.06
γ_1	1.00	1.02
γ_2	1.01	1.07
b_{11}	1.01	1.03
b_{22}	1.00	1.01
b_{33}	1.00	1.02
b_{44}	1.00	1.01
b_{55}	1.00	1.03
a_{21}	1.02	1.08
a_{42}	1.04	1.12
a_{43}	1.00	1.00
a_{54}	1.01	1.02
$\sigma_{\eta,11}$	1.05	1.13
$\sigma_{\eta,22}$	1.16	1.37
$\sigma_{\eta,33}$	1.07	1.13
$\sigma_{\eta,44}$	1.01	1.02
$\sigma_{\eta,55}$	1.00	1.00

Table 7.3: Posterior means and 95% credibility intervals of the parameters

Parameter	Estimate	2.5% percentile	97.5% percentile
α_0	-0.28	-0.76	-0.23
α_0	-0.20	-0.40	-0.01
α_0	-0.14	-0.35	0.07
α_0	-0.02	-0.24	0.20
α_0	0.46	0.13	0.80
β_1	0.003	-0.01	0.01
β_2	-0.007	-0.03	0.01
β_3	-0.032	-0.06	-0.004
γ_1	0.67	0.44	0.90
γ_2	0.53	0.30	0.75
b_{11}	0.13	-0.28	0.48
b_{22}	0.08	-0.13	0.29
b_{33}	0.38	-0.06	0.68
b_{44}	0.02	-0.22	0.24
b_{55}	0.10	-0.29	0.35
a_{21}	0.79	0.41	0.98
a_{42}	0.82	0.50	0.98
a_{43}	0.46	-0.13	0.95
a_{54}	0.86	0.61	0.99
$\sigma_{\eta,11}$	5.68	3.19	1.44
$\sigma_{\eta,22}$	1.44	1.01	2.8
$\sigma_{\eta,33}$	1.49	1.01	3.27
$\sigma_{\eta,44}$	1.43	1.01	2.44
$\sigma_{\eta,55}$	1.22	1.01	1.86

only very little evidence is supporting a trend change due to the introduction of MAPI while we may conclude that there is much evidence in favour of a trend change after the introduction of MAPII. The point estimate of β_3 indicates that the magnitude of the trend decreases after the introduction of MAPII. In order to infer on size of the trend after the introduction of MAPII, a credibility interval the sum of $\beta_1 + \beta_2 + \beta_3$ is needed. The posterior mean of this sum is -0.037 and the corresponding credibility interval is [-0.057,-0.017]. Hence after the introduction of MAPII, a decreasing trend in the violation probability is established.

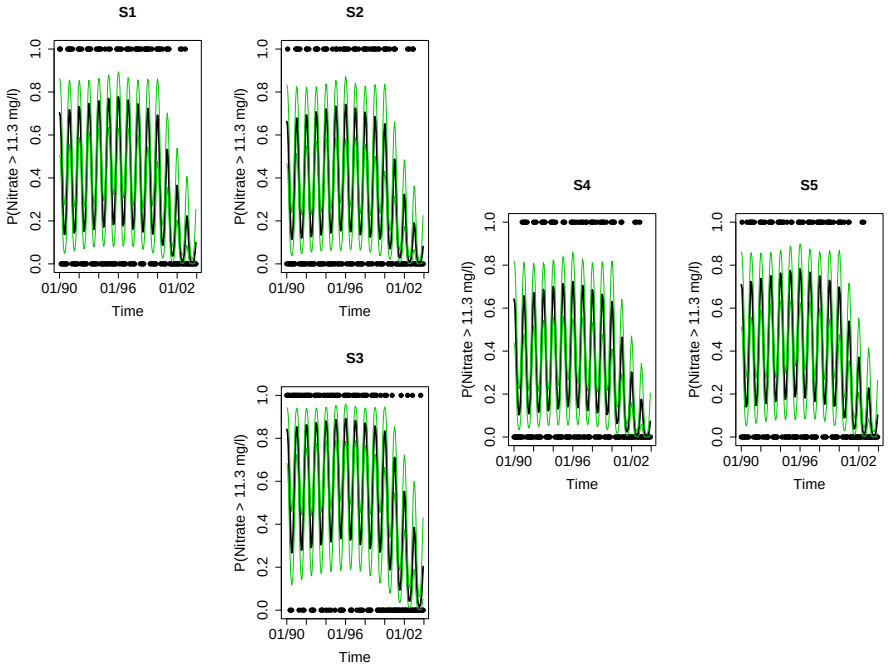

Figure 7.3: Evolution of the violation frequency of the nitrate standard of 11.3 mg N/l at five sampling locations of the river Yzer. The black line indicates the posterior mean probability to violate the standard, and the grey lines are the 95% credibility bands

7.5 Conclusions

In this chapter an extension of the spatio-temporal model for river monitoring networks is proposed for non-normal data. In particular, Bernoulli distributed observations originating from transforming the data using an environmental threshold were considered. This approach can be further adapted as long as the conditional distribution of the data is a member of the exponential family, by using the appropriate link function. The spatial dependence structure was restricted to a structure that was induced by river topology. The temporal dependence structure was assumed to be an AR(1) process. The temporal dependence can be extended towards an AR(p), process including the states at earlier time instants in Equation (7.8).

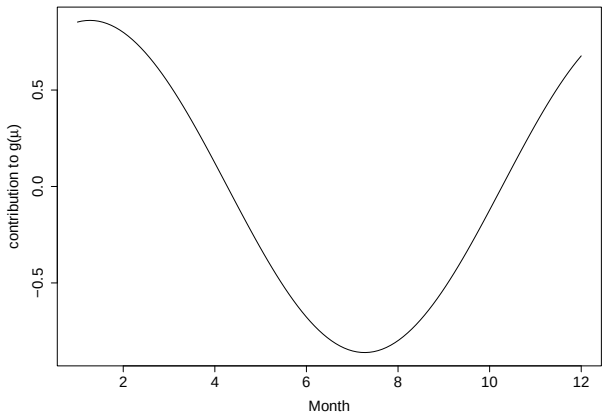

Figure 7.4: Contribution of the seasonal effect to ν_{it}^m obtained by using the posterior means of the distributions of γ_1 and γ_2

We think that our approach is also well suited to deal with water quality variables that consists of a large fraction of censored data. To reduce the loss of information due to the transformation into a binary response, the resolution could be refined by introducing a transformation of the continuous variable into a multinomial response.

The methodology was illustrated on a small case study on the river Yzer, Belgium. It consists of an assessment of the probability to violate the nitrate standard of 11.3 mg N/l. In the study region a strong seasonal pattern was present in the violation probability. This probability was larger during the wet winter period than in the dry summer period. There is strong evidence in favour of a trend change which is associated with the introduction of the second manure action plan. In particular, a decreasing trend in the probability to violate the standard is established in the study region after the introduction of the second manure action plan in 2000. There is also much evidence in favour of the presence of a positive spatial correlation between subsequent sampling locations in the main river. However, a temporal dependence was not likely to be strong.

7.6 Appendix

In Figure 7.5 diagnostic plots are given for the residuals $r_{it} = y_{it} - \mu_{it}^c$ obtained using MCMC. In the Bayesian framework, the residuals have a continuous posterior distribution and they can give information about outliers (Albert and Chib, 1995). If the posterior distribution of μ_{it} is in conflict with the observed value of y_{it}, then the posterior distribution of r_{it} will be concentrated towards extreme values (Albert and Chib, 1995). For Bernoulli distributed data, the support of r_{it} is in the interval $[y_{it} - 1, y_{it}]$. Hence, an observation $y_{it} = 0$ is unusual if the posterior distribution of r_{it} is located close the value -1, and an observation $y_{it} = 1$ is considered as an outlier if the posterior of r_{it} is concentrated towards the endpoint 1. On each time instant, the residual distribution is represented using boxplots. The box of most boxplots start close to zero. Some boxplots are entirely shifted to the endpoints. This indicates that there may be outliers present. In particular some outliers seem to be present in the middle of the time series for sampling locations S4 and at the end of the time series for S3. This can indicate that a mean model which considers a separate trend in S3 and/or S4 could be more appropriate.

In Figure 7.6-7.15 the evolution of the two MCMC chains are given for the parameters of the mean model. In all figures the chains are overlapping, indicating that they converged. Along with the evolution of the MCMC chains a plot of the posterior distribution of the parameter is given as well.

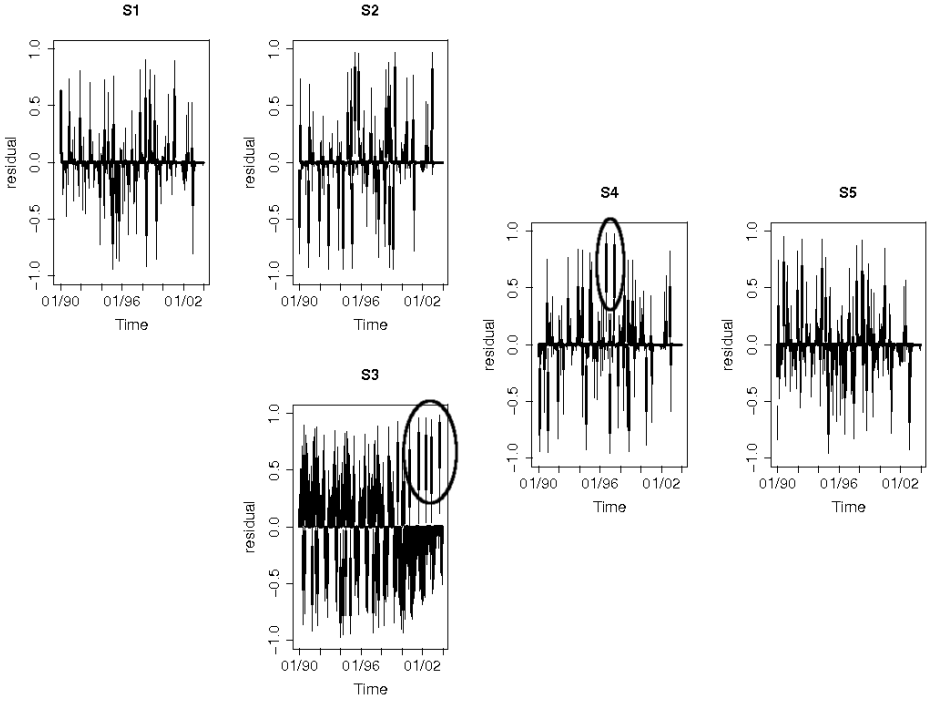

Figure 7.5: Diagnostic plots from the residuals

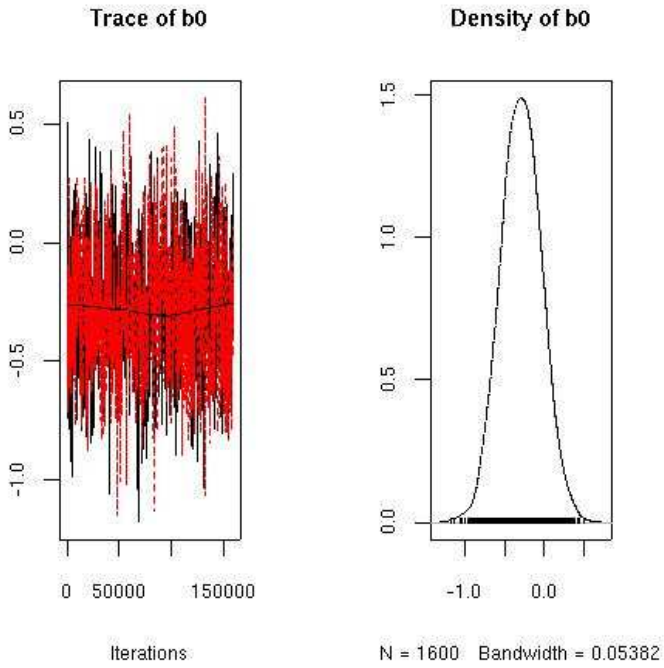

Figure 7.6. Evolution of the MCMC chains for the parameters α_0

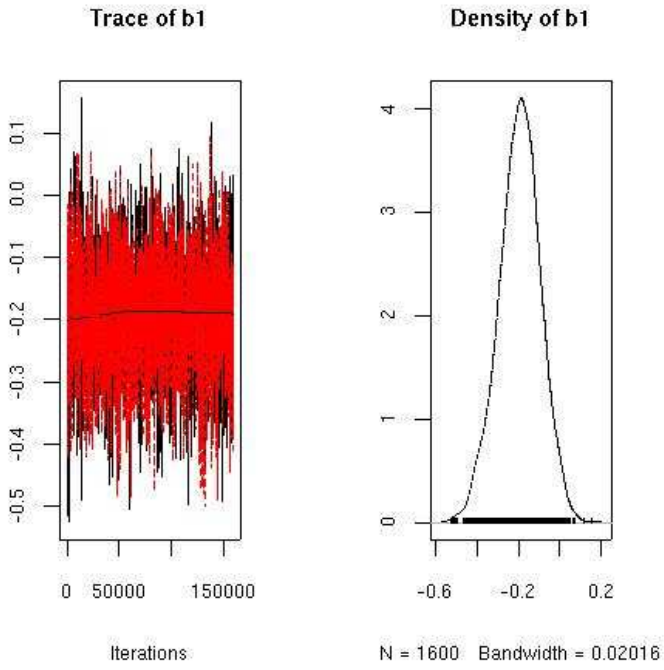

Figure 7.7. Evolution of the MCMC chains for the parameters α_1

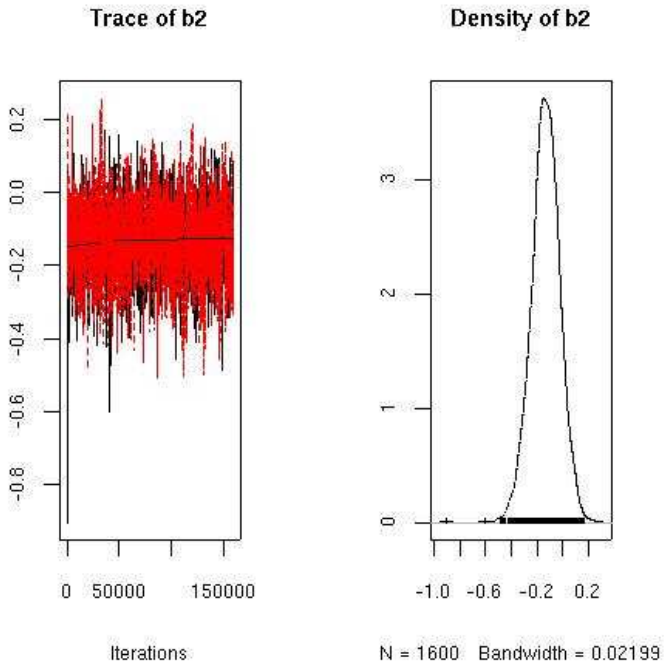

Figure 7.8. Evolution of the MCMC chains for the parameters α_2

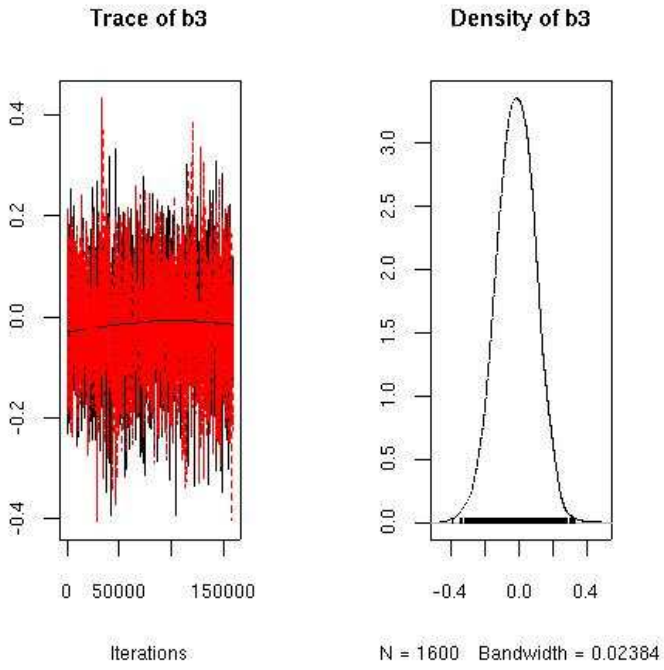

Figure 7.9. Evolution of the MCMC chains for the parameters α_3

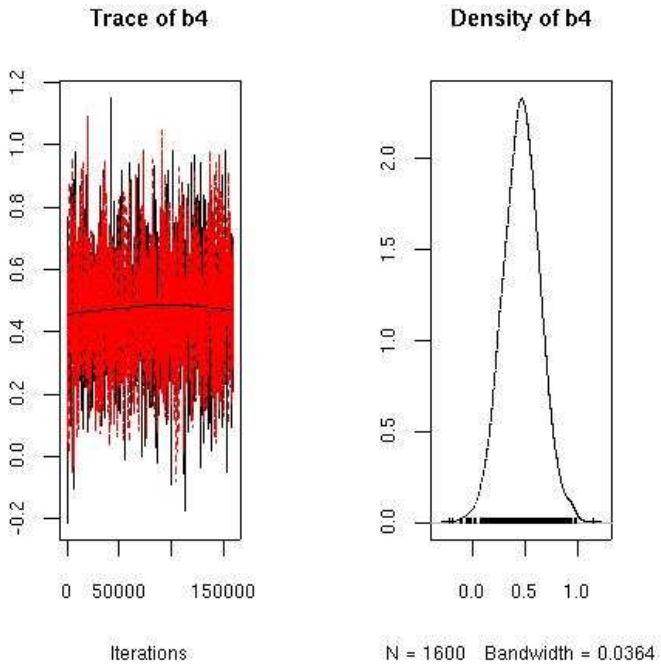

Figure 7.10. Evolution of the MCMC chains for the parameters α_4

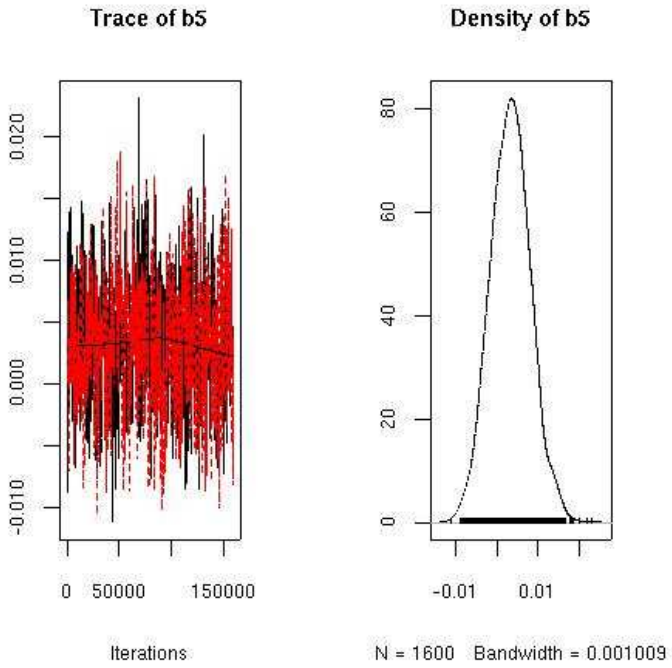

Figure 7.11. Evolution of the MCMC chains for the parameters β_1

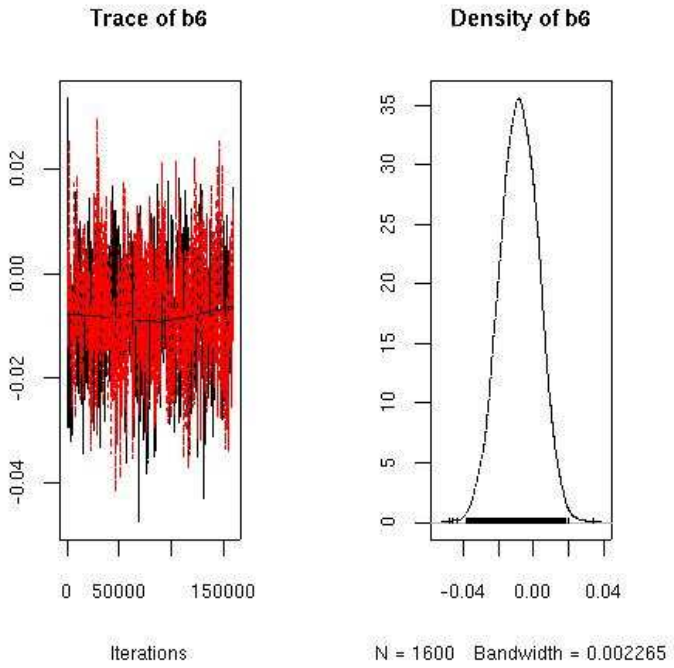

Figure 7.12. Evolution of the MCMC chains for the parameters β_2

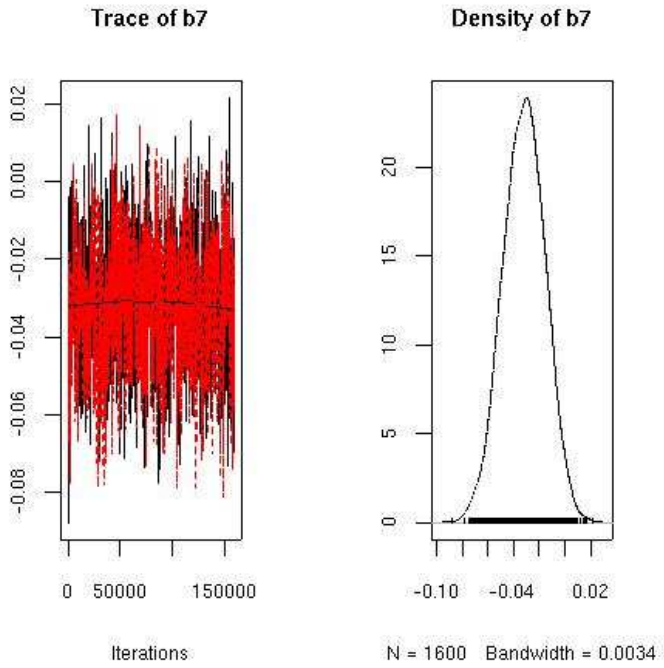

Figure 7.13. Evolution of the MCMC chains for the parameters β_3

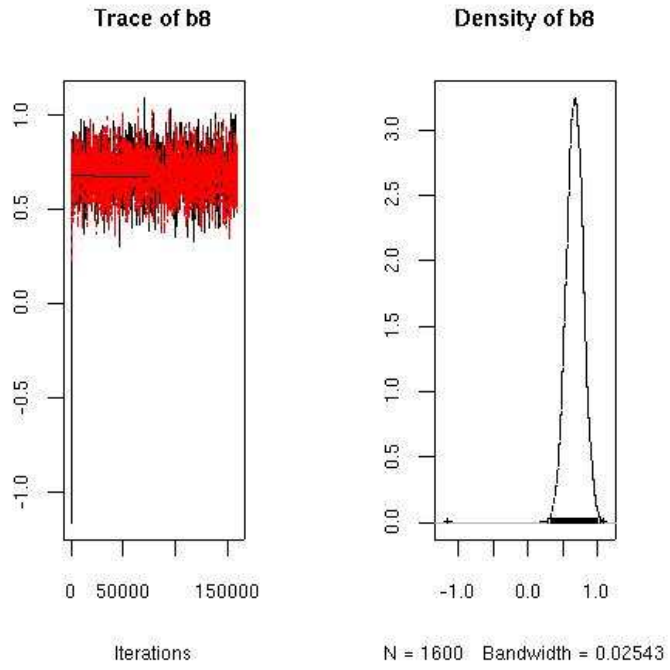

Figure 7.14. Evolution of the MCMC chains for the parameters γ_1

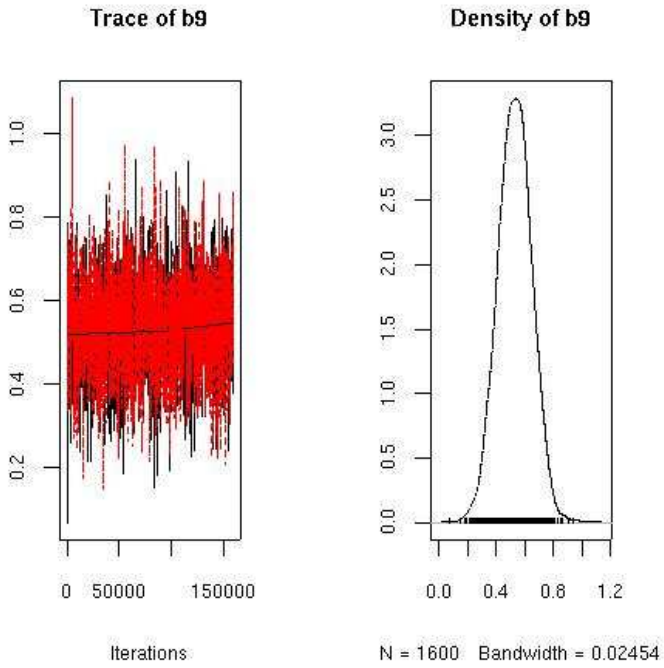

Figure 7.15. Evolution of the MCMC chains for the parameters γ_2

Chapter 8

Discussion, conclusions and future research perspectives

Our research was initially triggered by the Flemish environmental agency (VMM). Back in 2000 they introduced us to the large amount of data they collected through their monitoring networks. At that time the VMM's data validation procedure was entirely based on human experts. Due to the large amount of data a clear need was felt for an automated procedure to assist the experts with this validation process. Since the development of water quality monitoring networks is one of the key actions of the Water Framework Directive (WFD)(EC, 2000), data validation is clearly a problem that involves all European water authorities. After an exploratory analysis it quickly became clear that the considerable amount of missing data, the irregular sampling frequency and the nonlinear patterns and relationships present in the water quality variables required flexible models that are too a large extent driven by the data. Within this perspective, additive models were explored. While

examining the literature in context of the data validation problem, a second research opportunity became clear. At that time, the assessment of water quality data was mainly performed at the level of individual sampling locations and few ad hoc methods existed to infer on the water quality on a more regional scale. However, in order to reach the goals of the WFD, environmental agencies need statistical tools to infer on the evolution of the water status on a larger spatial scale. Therefore, spatio-temporal models for river networks have to be developed. The data validation problem and spatio-temporal modelling became the two major themes of this dissertation and the conclusions will be structured accordingly. In Section 8.1 the conclusions and perspectives on the validation procedure are given, while Section 8.2 deals with spatio-temporal modelling.

8.1 Statistical data validation

Quality assurance is specifically mentioned as an important activity in the WFD guidance document on monitoring (EC, 2003; Højberg et al., 2007). Hence data validation is an important activity in order to construct high quality environmental databases. In the next section our contribution to this problem is given.

8.1.1 Major contributions

In this dissertation a method for the validation of river water quality data is proposed. Based on the historical data an additive model is fitted. The model is then used to construct a prediction interval for a future observation. When the new observation is located within the interval the new observation is declared “valid”, otherwise it should be passed on to an expert for further evaluation.

The additive models were clearly able to catch the cyclic pattern present in the data, and they could model the nonlinear behaviour and relationships typically associated with river water quality data. As an interesting feature, the observed associations between the response and the predictors were found to respect known physical and biochemical relationships. Since the model selection is carried out at each time step, the models succeed to adapt to changes in the processes of the underlying river. The models were also capable to capture most of the serial correlation that was present in the monthly observations of the monitoring network.

Different prediction intervals were considered, analytical PI's (aPI), percentile based bootstrap intervals (pbPI) and the studentised prediction error based bootstrap PI's (sbPI). The coverages of the 95% sbPI's have been shown to be better and more robust than the other intervals. The sbPI's was also shown to be adequate to detect suspicious observations.

When the semi-automatic procedure is applied in practice, it should be used in an alternating fashion. One by one, each of the monitored variables should be chosen to be the response that has to be validated, while the remaining variables are used as potential predictors. This allows our procedure to detect suspicious observations located at the edges as well as observations laying in the centre of the univariate distribution of the validated response variables. In conclusion, our method combines the interesting features of classical multivariate outlier detection tools without having to impose linear relationships typically associated with these methods.

An ICT-tool based on this methodology is currently implemented at the VMM. It is used to validate all incoming measurements of their physico-chemical monitoring network on a day-to-day basis.

8.1.2 Future perspectives

For river monitoring networks that are sampled at a higher frequency, our method will no longer deal correctly with the serial correlation that is present in the data. The presented validation procedure should be adapted so as to account for serially correlated observations. One could try to model the serial correlation explicitly, e.g. by including AR terms in the additive model. Another possibility is the use of moving block bootstrap techniques to resample the time series by the use of independent blocks that capture the real-world dependence structure (Brumback et al., 2000). A challenge to both methods is that they should be adapted before they are able to work with missing observations and observations that are acquired at irregular time steps. The selection of the smoothing parameters by the use of traditional techniques such as cross validation are also known to be problematic in the presence of correlated errors (e.g. Hart, 1991). There is thus a big challenge related to the development of automatic validation procedures for environmental data obtained at higher, possibly irregular frequencies.

Another interesting problem that was not addressed in this study is the presence of

censored observations. In environmental time series, censored observations often occur due to the detection limits of the measuring methods. An additional problem is that these detection limits change over time due to technical improvements, the use of other protocols and/or the agencies that are contracting other laboratories for the analysis of their samples. Further research is needed to enable the data validation procedure to deal with censored observations in a proper way.

8.1.3 Conclusion from the case study

Our method was applied to the raw data of the Yzer basin. It detected unexpectedly high nitrate concentrations in the beginning of 2004. The diagnostic plots that were constructed indicated that the rejection of the nitrate data was related to the trend. After consulting the literature, this event could be explained by a dry summer in 2003 that was followed by an extremely wet period during the first months of 2004. During the dry summer and the autumn large amounts of nitrate had accumulated in the soils and the nitrate was washed to the receiving water during the subsequent extremely wet winter period.

8.2 Spatio-temporal models for river networks

To infer on the water status on a more spatial scale, spatio-temporal models are needed. Recently, river network modelling has entered the spatio-temporal arena (Gardner et al., 2003; Monestiez et al., 2005; Cressie et al., 2006; Ver Hoef et al., 2006). With respect to the spatial dependence structure an important distinction has to be made with the classical spatial structures. Due to the directional water flow within the river reaches, a causal interpretation can be given to the correlations. However, in contrast to time, rivers can join or split. This implies a more general branched unidirectional structure.

8.2.1 Major contributions

The few existing contributions in literature focus on spatial prediction on a river network. We do not aim to perform predictions at intermediate locations that are

not sampled. We want to perform an assessment at the sampling locations, for which we also have to take the spatio-temporal dependence structure into account to assure valid statistical inference. From this perspective, we proposed a spatio-temporal state-space model for river monitoring networks where the spatial dependence structure of the state variable is directly deduced from the river topology, and the temporal dependence structure is modelled by an AR(1) process. The state variable is embedded into an observation model that contains a model for the mean and accounts for cross-correlation between sampling locations that are not connected according to the river architecture. A marginal mean model is used to answer the research questions. The methodology is shown to be very flexible and according to the specification of the mean model, an assessment is possible on the level of individual sampling locations as well as on a more regional scale.

We proposed an expectation-conditional-maximisation (ECM) algorithm for the parameter estimation of spatio-temporal models with a parametric mean model. It makes use of the Kalman filter and smoother recursions, and uses generalised least squares for the estimation of the parameters of the mean model.

To assess nonlinear trends, the parametric mean model was replaced by a semi-parametric model. The estimation procedure however had to be adjusted to limit the computational burden. Therefore ordinary least squares was proposed to estimate the mean model. This also provides unbiased estimators for the parameters of the mean model. However, the reduction of the computational complexity does not come for free: the estimators are asymptotically less efficient. The residuals from the OLS fit are subsequently used for the estimation of the dependence structure. Only some minor adjustments were needed to use the ECM algorithm that was obtained for fully parametric models. In contrast with existing methodologies for (non)linear trend detection, our procedure takes the spatio-temporal dependence explicitly into account. Moreover, our method enables the detection of trends on a more local time scale. To verify at which time instants the nonlinear trend is significant, tests on its first derivative are performed at each time instant. Classical methods for multiple hypothesis testing were too conservative because they cannot incorporate the specific dependence between the tests. We have adopted the free step-down resampling method Westfall and Young (1993) and sampled from an appropriate null distribution to take the dependences between the statistical tests into account. This procedure explicitly takes the dependences between the statistical tests into account.

Finally, an extension is proposed to deal with non-normal data. Marginalised gen-

eralised linear mixed models were used to incorporate the dependence structure and to enable inference on the marginal mean. The method is derived in detail for binary data. The binary response was obtained by transforming the continuous data using the environmental threshold. According to us the transformation of the water quality data into binary data seems to be particularly suited to deal with water quality variables that consist of a large fraction of censored data. The spatial dependence structure was restricted to the structure that was induced by the river topology and the temporal dependence structure was assumed to be an AR(1) process.

8.2.2 Future perspectives

Spatio-temporal modelling in rivers is still in its initial stage. Hence, a lot of unexplored opportunities are waiting to be tackled by researchers. The development of more realistic correlation structures is one of the topics which should be addressed. Some interesting issues are

- When rivers enter tidal areas, the spatial dependence will become bidirectional.

- In our model, the temporal correlation structure is restricted to an AR(1) process. For water quality data sampled at time intervals of one month, this seemed to be the right model. For higher sampling frequencies more complex temporal structures will be needed. The methodology, however, can be easily extended to use more general ARMA structures. Harvey (1989) for instance, showed how AR(p) processes can be handled by the Kalman filter. In case of an AR(2) process the state variable $\boldsymbol{S}_t = (S_{1t}, ..., S_{pt})^T$ has to be replaced by a vector $(S_{1t}, ..., S_{pt}, S_{1t-1}, ..., S_{pt-1})^T$. This leads to a reformulation of the observation model and the Kalman filter equations.

- The spatial variance-covariance matrix of the observation model $\boldsymbol{\Sigma}_\epsilon$ enables cross-correlations between sampling locations that were not connected by the river. In this work we used a saturated parameterisation for $\boldsymbol{\Sigma}_\epsilon$. However, for large monitoring networks too many parameters are involved and to reduce the complexity $\boldsymbol{\Sigma}_\epsilon$ should be further parameterised. Due to the estimation orthogonality in the first CM step, this will only alter the update Equation (5.35).

The presented ECM algorithm cannot handle missing data. Therefore future extensions of the ECM are needed. Both the ECM algorithm and the Kalman filter are in principle well suited to deal with missing data. In particular, the E-step of the EM algorithm should be modified to provide sufficient statistics for both the latent variable as for the missing observations.

A big challenge is the development of methods to deal with non-normal data that are acquired on a river network. We have given a first impulse on how to treat binary data. To reduce the loss of information due to the transformation into a binary response, the resolution could be refined by introducing a transformation of the continuous variable into a multinomial response. For other types of non-normal data, the approach can be further adapted as long as the conditional distribution of the data is a member of the exponential family. By using an appropriate link function and an appropriate mapping function between the marginal model and the conditional model, an appropriate generalised linear mixed model can be formulated for the estimation of the parameters of the marginal mean model. The use of more realistic correlation structures is also an issue that should be addressed in this setting.

8.2.3 Conclusions on the study region

The methodology was applied on a case study at five sampling locations of the river Yzer. The augmented data had to be used because our estimation algorithms are currently not designed to deal with missing data. In the sampling period, a first manure action plan (MAP) was introduced in 1996 (Vlaams Parlement, 1995) and a second and more restrictive MAP was established in 2000 (Vlaams Parlement, 1999). Both MAP's aim to reduce the nutrient pollution originating from agricultural activities (Vlaams Parlement, 1995, 1999). Depending on the formulation of the mean model inference is possible on a regional scale, on the level of a river reach or on the level of individual sampling locations.

In a first case study the annual average of the nitrate concentration in 2003 is shown to be very significantly lower than the general mean ($p < 0.01$). Moreover, in the main river, the mean nitrate concentration of 2003 was also significantly lower than the mean of 2001 and 2002 ($p = 0.03$).

In the second case study, the spatio-temporal model was used to estimate a non-linear trend. A significant decrease in the nitrate concentration is established in

the study region between the introduction of the first MAP and the second MAP ($\alpha = 0.05$). The trend remains significant until January 2002.

Finally an assessment was done on the violation frequency of the nitrate standard of 11.3 mg N/l. In the study region a strong seasonal pattern was present in the violation probability. The probability to violate the standard was larger during the wet winter period than in the dry summer period. There was also strong evidence in favour of a trend change after the introduction of the second manure action plan. In particular, a decreasing trend in the probability to violate the standard is detected in the study region after the introduction of MAPII.

Although the data analysis has no causal interpretation, the results of the case studies give a strong indication that the introduction of the manure action plans had a beneficial effect on the nitrate status in the study region.

Bibliography

Akaike, H. (1973). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19:716–723.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679.

Albert, J. H. and Chib, S. (1995). Bayesian residual analysis for binary response regression models. *Biometrika*, 82(4):747–759.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185.

Alwan, L. C. (1992). Effects of autocorrelation on control chart performance. *Communications in Statistics - Theory and Methods*, 21(4):1025–1049.

Anonymous (2005). *Water- & waterbodemkwaliteit - Lozingen in het water - Evaluatie saneringsinfrastructuur 2004*. Vlaamse Milieumaatschappij, Aalst.

Ansley, C. and Kohn, R. (1985). Estimation, filtering and smoothing in state space models with incompletely specified initial conditions. *Annals of Statistics*, 13:1286–1316.

Bengtsson, T. and Cavanaugh, J. E. (2006). An improved Akaike information criterion for state-space model selection. *Computational Statistics and Data Analysis*, 50:2635–2654.

Bilonick, R. A. (1983). Risk qualified maps of hydrogen ion concentration for the New York state area 1966-1978. *Atmospheric Environment*, 17:2513–2524.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25.

Brumback, B. A., Neas, L. M., Ryan, L. M., Schwartz, J. D., Stark, P. C., and Burge, H. A. (2000). Transitional regression models, with application to environmental time series. *Journal of the American Statistical Association*, 95:16–27.

Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, 17(2):453–510.

Burn, D. H. and Hag Elnur, M. A. (2002). Detection of hydrologic trends and variability. *Journal of Hydrology*, 255:107–122.

Cai, Z. and Tiwari, R. C. (2000). Application of a local linear autoregressive model to BOD time series. *Environmetrics*, 11:341–350.

Casella, G. and Berger, R. L. (2002). *Statistical inference*. Duxbury Press, Pacific Grove, second edition.

Chatfield, C. (1993). Calculating interval forecasts. *Journal of Business and Economic Statistics*, 11(2):121–134.

Clements, M. P. and Taylor, N. (2001). Bootstrapping prediction intervals for autoregressive models. *International Journal of Forecasting*, 17(2):247–267.

Cleveland, R. B., Cleveland, W. S., McRae, J., and Terpenning, I. (1990). STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6:3–73.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836.

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596–610.

Cleveland, W. S. and Grosse, E. (1991). Computational methods for local regression. *Statistics and Computing*, 1:47–62.

Cressie, N., Frey, J., Harch, B., and Smith, M. (2006). Spatial prediction on a river network. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(2):127–150.

Cressie, N. and Majure, J. (1997). Spatio-temporal statistical modeling of livestock waste in streams. *Journal of Agricultural, Biological, and Environmental Statistics*, 2:24–47.

Davison, A. and Hinkley, D. (1997). *Bootstrap Methods and their Applications*. Cambridge Series on Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, first edition.

De Rycke, A., Devos, K., and Decleer, K. (2001). *Verkennende ecologische gebiedsvisie voor de IJzervallei. Rapport Instituut voor Natuurbehoud*. Instituut voor Natuurbehoud, Brussel.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Digalakis, V., Rohlicek, J., and Ostendorf, M. (1993). ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition. *IEEE Transactions Speech and Audio Processing*, 1:431–442.

Diggle, P., Liang, K., and Zeger, S. (1994). *Analysis of longitudinal data*. Clarendon Press, Oxford.

Dominici, F., McDermott, A., Zeger, S. L., and Samet, J. M. (2002). On the use of generalized additive models in time-series studies of air pollution. *American Journal of Epidemiology*, 156(3):193–203.

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

EC (2000). Directive 2000/60/EC of the European Parliament and of the Council of October 23 2000 establishing a framework for community action in the field of water policy. *Official Journal of the European Communities*, pages L327/1–L327/72, 22.12.2000.

EC (2003). Water framework directive, common implementation strategy, working group 2.7. monitoring. Available on http://forum.europa.eu.int/Public/irc/env/wfd/library.

Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Monographs on Statistics and Applied Probability. Chapman & Hall, New York, first edition.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121.

Eubank, R. (2006). *A Kalman filter primer*. Chapman & Hall, New York, first edition.

Eubank, R. and Speckman, P. (1993). Confidence bands in nonparametric regression. *Journal of the American Statistical Association*, 88(424):1287–1301.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, 87(420):998–1004.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiency. *The Annals of Statistics*, 21(1):196–216.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*. Monographs on Statistics and Applied Probability. Chapman & Hall, New York, first edition.

Friedman, J. H. (1984). A variable span scatterplot smoother. Laboratory for computational statistics. *Stanford University Technical Report No. 5*.

Gardner, B., Sullivan, P., and Jr, A. L. (2003). Predicting stream temperatures: geostatistical model comparison using alternative distance metrics. *Canadian Journal of Fisheries and Aquatic Sciences*, 60:344–351.

Gasser, T. and Müller, H. G. (1979). *Smoothing techniques for curve estimation, T. Gasser and M. Rosenblatt (eds)*, pages pp. 23–68. Springer-Verlag, Heidelberg.

Gelman, A. (1996). *Markov Chain Monte Carlo in practice, W.R. Gilks, S. Richardson and D. J. Spiegelhalter (eds)*, chapter Inference and monitoring convergence, pages 131–144. Chapman & Hall, New York, first edition.

Giannitrapani, M., Bowman, A. W., and Scott, E. M. (2005). Additive models for correlated data with applications to air pollution monitoring. *Technical Report, Department of Statistics, University of Glasgow*.

Gilks, W., Richardson, S., and Spiegelhalter, D. J., editors (1996a). *Markov Chain Monte Carlo in practice*. Chapman & Hall, New York, first edition.

Gilks, W., Richardson, S., and Spiegelhalter, D. J. (1996b). *Markov Chain Monte Carlo in practice, W.R. Gilks, S. Richardson and D. J. Spiegelhalter (eds)*, chapter Introducing Markov Chain Monte Carlo, pages 1–20. Chapman & Hall, New York, first edition.

Griswold, M. E. and Zeger, S. L. (2004). On marginalized multilevel models and their computation. *Johns Hopkins University, Dept. of Biostatistics Working Papers*, Paper:99.

Hart, J. D. (1991). Kernel regression estimation with time series errors. *Journal of the Royal Statistical Society, Series B*, 53(1):173–187.

Hart, J. D. (1997). *Nonparametric smoothing and lack-of-fit tests*. Springer series in statistics. Springer-Verlag, New York, first edition.

Harvey, A. C. (1989). *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press, Cambridge, first edition.

Hasti, T. and Loader, C. (1993). Local regression: automatic kernel carpentry. *Statistical Science*, 8(2):120–129.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*. Monographs on Statistics and Applied Probability. Chapman & Hall, New York, first edition.

Hastie, T. J., Tibshirani, R. J., and Friedman, J. H. (2001). *The elements of statistical learning*. Springer series in statistics. Springer-Verlag, New York, first edition.

Heagerty, P. J. (2002). Marginalized transition models and likelihood inference for longitudinal categorical data. *Biometrics*, 58:342–351.

Heagerty, P. J. and Zeger, S. L. (2000). Marginalized multilevel models and likelihood inference. *Statistical Science*, 15(1):1–26.

Hirsch, R. M., Slack, J. R., and Smith, R. A. (1982). Techniques of trend analysis for monthly water quality data. *Water Resources Research*, 18:107–121.

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

Huang, H. and Cressie, N. (1996). Spatio-temporal prediction of snow water equivalent using the Kalman filter. *Computational Statistics & Data Analysis*, 22(2):159–175.

Højberg, A. L., Refsgaard, J. C., van Geer, F., Jørgensen, L. F., and Zsuffa, I. (2007). Use of models to support the monitoring requirements in the water framework directive. *Water Resources Management*, Published 0nline.

Kauermann, G. and Opsomer, J. (2004). Generalized cross-validation for bandwidth selection and backfitting estimates in generalized additive models. *Journal of Computational and Graphical Statistics*, 13(1):66–89.

Kim, J. H. (1999). Asymptotic and bootstrap prediction regions for vector autoregression. *International Journal of Forecasting*, 15(4):393–403.

Kim, J. H. (2004). Bootstrap prediction intervals for autoregression using asymptotically mean-unbiased estimators. *International Journal of Forecasting*, 20(1):85–97.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22.

Ljung, G. M. and Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65:553–564.

Loader, C. (1999a). Bandwidth selection: Classical or plug-in? *The Annals of Statistics*, 27(2):415–438.

Loader, C. (1999b). *Local regression and likelihood*. Statistics and Computing. Springer-Verlag, New York, first edition.

Mammen, E. and Park, B. U. (2005). Bandwidth selection for smooth backfitting in additive models. *The Annals of Statistics*, 33(3):1260–1294.

Mardia, K. V. and Marshall, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71(1):135–146.

Maruyama, G. M. (1997). *Basics of structural equation modeling*. Sage Publications, Inc., Thousand Oaks.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. Monographs on Statistics and Applied Probability. Chapman & Hall, New York, second edition.

McLachlan, G. and Krishnan, T. (1997). *The EM algorithm and extensions*. Wiley, New York.

McMullan, A. (2004). *Non-linear and nonparametric modelling of seasonal environmental data, Ph.D. thesis*. University of Glasgow.

McMullan, A., Bowman, A. W., and Scott, E. (2003). Non-linear and nonparametric modelling of seasonal environmental data. *Computational Statistics*, 18:167–183.

McWilliams, T. P. (1990). A distribution-free test for symmetry based on a runs statistic. *Journal of the American Statistical Association*, 85(412):1130–1133.

Meng, X. L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, 80:267–278.

Monestiez, P., Bailly, J., Lagacherie, P., and Voltz, M. (2005). Geostatistical modelling of spatial processes on directed trees: Application to fluvisol extent. *Geoderma*, 128(3-4):179–191.

Montgomery, D. (2005). *Introduction to Statistical Quality Control*. John Wiley and Sons, Inc, New York, fifth edition.

Montgomery, D. and Mastrangelo, C. M. (1991). Some statistical process control methods for autocorrelated data. *Journal of Quality Technology*, 23(3):179–193.

Nadaraya, E. (1964). On estimating regression. *Theory of Probability and its applications*, 10:186–190.

Opsomer, J. (2000). Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis*, 74:166–179.

Opsomer, J. and Ruppert, D. (1998). A fully automated bandwidth selection method for fitting additive models. *Journal of the American Statistical Association*, 93(422):605–619.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41(1):100–115.

Parr, T., Sier, A., Battarbee, R., Mackay, A., and Burgess, J. (2003). Detecting environmental change: science and society-perspectives on long-term research and monitoring in the 21st century. *Science of the Total Environment*, 310(1-8).

Penny, K. (1996). Appropriate critical values when testing for a single multivariate outlier by using the Mahalanobis distance. *Applied Statistics Journal of the Royal Statistical Society Series C*, 45(1):73–81.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2004). *coda: Output analysis and diagnostics for MCMC*. R package version 0.9-1.

Pourahmadi, M. (2001). *Foundations of time series analysis and prediction theory*. Wiley Series in Probability and Statistics. John Wiley and Sons, Inc, New York, first edition.

Prucha, I. (1984). On the asymptotic efficiency of feasible Aitken estimators for seemingly unrelated regression models with error components. *Econometrica*, 52(1):203–308.

Qian, S., Borsuk, M. E., and Stow, C. A. (2000). Seasonal and long-term nutrient trend decomposition along a spatial gradient in the Neuse river watershed. *Environmental Science & Technology*, 34(21):4474–4482.

Reynolds, M. R. and Lu, C.-W. (1997). Control charts for monitoring processes with autocorrelated data. *Nonlinear Analysis, Theory, Methods & Applications*, 30(7):4059–4067.

Roberts, S. W. (1959). Control charts based on geometric moving averages. *Technometrics*, 1(3):239–250.

Rouhani, S. and Wackernagel, H. (1990). Multivariate geostatistical approach to space-time data analysis. *Water Resources Research*, 26:585–591.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, 90(432):1257–1270.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, 46:561–84.

Shewart, W. A. (1931). *Economic Control of Quality of Manufactured Product*. Van Nostrand Reinhold Co., New York.

Shin, D. W. and Oh, M. S. (2002). Asymptotic efficiency of the ordinary least squares estimator for regressions with unstable regressors. *Economic Theory*, 18:1121–1138.

Shumway, R. H. and Stoffer, D. S. (1982). An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis*, 3:253–264.

Shumway, R. H. and Stoffer, D. S. (2006). *Time series analysis and its applications*. Springer Texts in Statistics. Springer-Verlag, New York, second edition.

Spiegelhalter, D. J., Best, N. G., Gilks, W. R., and Inskip, H. (1996). *Markov Chain Monte Carlo in practice, W.R. Gilks, S. Richardson and D. J. Spiegelhalter (eds)*, chapter Hepatitis B: a case stuy in MCMC methods, pages 21–45. Chapman & Hall, New York, first edition.

Stålnacke, P., Grimvall, A., Sundblad, K., and Wilander, A. (1999). Trends in nitrogen transport in Swedish rivers. *Environmental Monitoring and Assessment*, 59(1):47–72.

Trapletti, A. (2004). *tseries: Time series analysis and computational finance*. R package version 0.9-24.

Van Belle, G. and Hughes, J. P. (1984). Nonparametric tests for trend in water quality. *Water Resources Research*, 20(1):127–136.

Venkatasubramanian, V., Rengaswamy, R., Kavuri, S. N., and Yin, K. (2003). A review of process fault detection and diagnosis part III: Process history based methods. *Computers and Chemical Engineering*, 27(3):327–346.

Ver Hoef, J., Peterson, E., and Theobald, D. (2006). Spatial statistical models that use flow and stream distance. *Environmental and Ecological Statistics*, 13:449–464.

Vlaams Parlement (1995). Decreet van 20 december 1995 tot wijziging van het decreet van 23 januari 1991 inzake de bescherming van het leefmilieu tegen de verontreiniging door meststoffen. *Belgisch Staatsblad*, 30/12/1995(249 (Ed. 5)):36069.

Vlaams Parlement (1999). Decreet van 11 mei 1999 tot wijziging van het decreet van 23 januari 1991 inzake de bescherming van het leefmilieu tegen de verontreiniging door meststoffen en tot wijziging van het decreet van 28 juni 1985 betreffende de milieuvergunning. *Belgisch Staatsblad*, 20/08/1999(Ed. 1):30967.

Wardell, D. G., Moskowitz, H., and Plante, R. D. (1992). Control charts in the presence of data correlation. *Management Science*, 38(8):1084–1105.

Watson, G. (1964). Smooth regression analysis. *Sankhya series A*, 26:359–372.

Wecker, W. E. and Ansley, C. F. (2002). The signal extraction approach to nonlinear regression and spline smoothing. *Journal of the American Statistical Association*, 78(381):81–89.

Wermuth, N. (1980). Linear recursive equations, covariance selection, and path analysis. *Journal of the American Statistical Association*, 75(372):963–972.

Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. John Wiley and Sons, New York.

Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. John Wiley & Sons Ltd, Chichester, first edition.

Wikle, C. and Cressie, N. (1999). A dimension-reduced approach to space-time Kalman filtering. *Biometrika*, 86(4):815–829.

Wood, S. and Augustin, N. (2002). GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling*, 157:157–177.

Xu, K. and Wikle, C. K. (2005). Estimation of parameterized spatio-temporal dynamic models. In review (available at http://www.stat.missouri.edu/∼wikle/xu_wikle_070705.pdf).

Yoo, C. K., Lee, I. B., and Vanrolleghem, P. (2004). Application of multiway ICA for on-line process monitoring of a sequencing batch reactor. *Water Research*, 38(7):1715–1732.

Yoo, C. K., Villez, K., Lee, I. B., Rosen, C., and Vanrolleghem, P. (2007). Multi-model statistical process monitoring and diagnosis of a sequencing batch reactor. *Biotechnology and Bioengineering*, 96(4):687–701.

Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42:121–130.

Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, 44:1049–1060.

Summary

The European Water Framework Directive (WFD)(EC, 2000) is one of the driving forces in environmental policy in the European Union. The WFD's overall environmental objective is the achievement of 'good status' for all of Europe's surface- and ground waters within a 15-year period. Its implementation is a big challenge for the European environmental managers. One of the key actions of the WFD is the design of operational monitoring programmes. Thus, large amounts of water quality data are being collected, processed and stored throughout Europe. Due to the large amount of the data and their complex nature, statistical modelling has become an essential tool to extract reliable information from these observations. Because high quality data is essential for an adequate management of the water resources, data validation procedures are required to build consistent databases. Once the environmental agencies have a consistent database at their disposal, the data should be used to assess the evolution of the water status and to evaluate the impact of their management strategies. Such an assessment should be possible at the level of individual sampling locations as well as on a more regional scale. Due to the spatio-temporal dependence structure of monitoring network data, spatio-temporal models are needed for a correct statistical assessment. For river monitoring networks the development of spatio-temporal models has just begun. The data validation problem and the development of spatio-temporal models for river network data are the two major themes of this dissertation.

In the first part the data validation problem is addressed. Like other environmental data, water quality data have a complex nature. They contain a considerable amount of noise due to their natural variability and the measurement error. They may contain missing values, are often non-normally distributed and are commonly

gathered at irregular time instants. Moreover, they possess cyclic variations and contain nonlinear trends. With this respect, additive models are explored to model water quality data. These models are then used to design an automatic validation procedure for new observations that are acquired with a river monitoring network. Based on historical data, additive models are fitted to predict new observations and to construct prediction intervals (PI's). A new observation is declared valid if it is located within the interval. Several methods were developed to derive such PI's and the PI that was based on bootstrapping studentised prediction errors was shown to be most accurate and most robust to deviations from normality. The coverage of these prediction intervals and their power to detect anomalous data are successfully established in a simulation study. The method is illustrated on two case studies in which the method detected abnormal nitrate concentrations in the water body provoked by a dry summer which was followed by an extremely wet winter period. Currently, the Flemish environmental agency is also using our method for the validation of new observations from their physico-chemical monitoring network.

In the second part a spatio-temporal model is developed for river monitoring network data. The aim was to enable valid statistical inference based on the data that is observed at the sampling locations. Therefore the observations of the monitoring network at a certain time instant can be considered as the realisation of a finite-dimensional multivariate random variable with each dimension corresponding to each of the p sampling locations. This enables us to write the model as a p-dimensional state-space model. The state variable is defined by a Directed Acyclic Graph (DAG) that is derived from the river network topology. In reality the dependence structure based on the DAG may be obscured by environmental factors such as rainfall and climatological conditions in general. This is taken into account by embedding the state variable into an observation model. Initially, the observation model was extended with a linear model for the mean. The specification of the mean model allows the assessment of different research questions. An efficient expectation-conditional-maximisation (ECM) algorithm is proposed for parameter estimation, using the Kalman filter and smoother in both E- and CM-steps. However, many environmental processes are characterised by a nonlinear trend. To allow the estimation of such a nonlinear trend, we replaced the parametric mean model by a semiparametric model that used a smoother for the estimation of the trend component. This procedure also allows to test for trends on a smaller time scale. To detect if the local trend is significant, tests on the first derivative of the nonlinear trend are performed at each time step. This results, however, in a large number of simultaneous tests. Multiplicity is thus another problem which had to be addressed. Many environmental processes are also non-Gaussian. To

handle such data, a generalisation of our spatio-temporal model is needed. A first attempt is presented that can handle binary data. Environmental compliance is often based on threshold levels, providing a binary response to the decision maker. We made use of generalised linear mixed models (GLMM) to model such a binary response. Again, the spatio-temporal dependence structure is introduced by using a latent state variable. In a GLMM the parameters of the mean model have a conditional interpretation. In an environmental context, however, we want to infer on the marginal mean. Therefore the marginalised version of the GLMM of Heagerty and Zeger (2000) is used. They introduced a mapping function between the conditional and marginal model components to identify a conditional model structure that allows immediate estimation of the marginal mean parameters.

The spatio-temporal models are applied on a case study of five sampling locations of the river Yzer. In the sampling period, a first manure action plan (MAP) was introduced in 1996 (Vlaams Parlement, 1995) and a second and more restrictive MAP was established in 2000 (Vlaams Parlement, 1999). Both MAP's aim to reduce the nutrient pollution originating form agricultural activities. Our modelling procedure was shown to be very flexible. Depending on the formulation of the mean model, inference is possible on a regional scale, on the level of a river reach or on the level of individual sampling locations. In a first case study the annual average of the nitrate concentration in 2003 is shown to be very significantly lower than the general mean ($p < 0.01$). Moreover, in the main river, the mean nitrate concentration of 2003 was also significantly lower than the mean of the two most recent years ($p = 0.03$). In the second case study, the spatio-temporal model was used to estimate a nonlinear trend. A significant decrease in the nitrate concentration is established in the study region between the introduction of the first MAP and the second MAP ($\alpha = 0.05$). The trend remains significant until January 2002. Both case studies indicated a strong seasonal variation with lower nitrate values in summer and higher contributions in winter. Finally an assessment was done of the violation of the nitrate standard of 11.3 mg-N/l. In the study region a strong seasonal pattern was present in the violation probability. The probability to violate the standard was larger during the wet winter period than in the dry summer period. There was a strong evidence in favour of the presence of a trend change after the introduction of the second manure action plan. The trend change was large enough to establish a decreasing trend in the violation probability of the standard in the study region. Although the data analysis has no causal interpretation, the results of the case studies give a strong indication that the introduction of the manure action plans had a beneficial effect on the nitrate status in the study region.

Samenvatting

De Europese kaderrichtlijn water (KRLW)(EC, 2000) heeft verregaande gevolgen voor het waterbeleid in de Europese lidstaten. De algemene doelstelling van de richtlijn is een goede toestand voor oppervlaktewater en grondwater te bereiken tegen eind 2015. De uitbouw van meetnetten is één van de kernactiviteiten die door de richtlijn wordt beoogd. Hierdoor worden in Europa grote hoeveelheden waterkwaliteitsdata bemonsterd en opgeslagen. De grote hoeveelheid van gegevens en de complexiteit van milieukundige data vereisen het gebruik van modellen voor een doeltreffende analyse van de waterstatus. Voor de uitbouw van een goed waterbeleid is het essentieel om te beschikken over data van hoge kwaliteit. Daarom zijn efficiënte methoden voor het valideren van de meetgegevens vereist. Uiteraard dienen de gegevens die beschikbaar zijn na de validatie te worden geanalyseerd. Het opvolgen van de evolutie van de waterstatus en het evalueren van de impact van de reeds getroffen maatregelen zijn cruciaal voor de verdere uitbouw en verfijning van een langetermijnvisie met het oog op het behalen van de algemene doelstelling van de KRLW. Aangezien de KRLW de waterproblematiek integraal benadert op stroomgebiedniveau is het wenselijk om de gevens niet enkel op meetpuntniveau te analyseren maar tevens op een subbekken- en bekkenniveau. De ruimtelijke en temporele afhankelijkheid van de waterkwaliteitsdata vereisen het gebruik van spatio-temporele statistische modellen voor het uitvoeren van een correcte statistische analyse. Voor riviernetwerken staat de ontwikkeling van dergelijke spatio-temporele modellen nog in de kinderschoenen. Om aan deze noden tegemoed te komen zijn de ontwikkeling van methoden voor datavalidatie en de ontwikkeling van spatio-temporele modellen voor riviernetwerken de twee kernthema's van dit doctoraatsonderzoek.

Het eerste deel van dit onderzoek spitst zich toe op de ontwikkeling van een semi-automatische methode voor de validatie van waterkwaliteitsdata. Waterkwaliteitsdata worden gekarakteriseerd door ondermeer een grote variabiliteit, meetruis, cyclische variatie, niet-lineare trends en ontbrekende waarnemingen. Daarnaast is de bemonsteringsfrequentie dikwijls onregelmatig. Dit zorgt ervoor dat flexibele modellen zijn vereist. Daarom wordt geopteerd voor het gebruik van additieve modellen voor de ontwikkeling van de datavalidatiemethode. Aan de hand van de historische data worden deze modellen gefit. Vervolgens worden ze gebruikt voor het voorspellen van nieuwe metingen en voor het construeren van verwachtingsintervallen. Wanneer een nieuwe meting in het verwachtingsinterval ligt, wordt ze aanvaard. Als dit niet het geval is, moet ze verder worden onderzocht door experten. Op deze manier kunnen de experten zich toespitsen op de analyse van metingen die een potentiële afwijking bevatten. Voor het opstellen van de verwachtingsintervallen worden verschillende methoden gebruikt. De methode waarbij gebruik gemaakt wordt van het bootstrappen van predictiefouten bleek het meest accuraat te zijn. Tevens zijn deze intervallen meer robuust voor afwijkingen van normaliteit. Uit een simulatiestudie blijken deze intervallen over een hoge kracht te beschikken om mogelijke afwijkende observaties te detecteren. De methode wordt geïllustreerd aan de hand van een gevallenstudie waarbij alle nitraatmetingen van 2003 en 2004 worden gevalideerd in het IJzerbekken. De methode detecteert hierbij een grote hoeveelheid afwijkende metingen in het begin van 2004. Uit de literatuur bleek dat deze hoge metingen toe te wijzen zijn aan de droge zomer van 2003 die werd gevolgd door een extreem nat voorjaar in 2004. Onze methode voor datavalidatie wordt momenteel gebruikt door de Vlaamse milieumaatschappij (VMM) voor het valideren van de meetgegevens afkomstig van het fysico-chemisch meetnet voor oppervlaktewater.

Het tweede deel van dit doctoraatsonderzoek focust zich op de ontwikkeling van spatio-temporele modellen voor de analyse van riviernetwerken. Hierbij maken we gebruik van toestandsmodellen. De afhankelijkheidsstructuur wordt gemodelleerd aan de hand van een latent proces. Voor de temporele afhankelijkheidsstructuur wordt een AR(1) proces verondersteld. De spatiale afhankelijkheidsstructuur wordt afgeleid van de riviertopologie. Een dergelijke afhankelijkheidsstructuur is echter nogal restrictief. Meetpunten die dichtbij elkaar liggen maar niet rechtstreeks verbonden zijn met elkaar door de rivier zullen in de realiteit tevens gecorreleerd zijn. Dit kan bijvoorbeeld door het voorkomen van gelijkaardige klimatologische condities. Daarom wordt het latent proces opgenomen in een observatiemodel die wel correlatie toelaat tussen meetpunten die niet verbonden zijn door de rivier. In het observatiemodel wordt tevens een lineair model opgenomen

voor het gemiddelde. Naargelang de specificatie van dit model is het toetsen van verschillende onderzoeksvragen mogelijk. Naast de ontwikkeling van het model, hebben we tevens een schattingsalgoritme ontwikkeld dat gebruik maakt van de Kalman filter en smoother. Milieukundige processen worden echter vaak gekarakteriseerd door niet-lineaire trends. Om een analyse van dergelijke trends mogelijk te maken hebben we het lineaire model voor het gemiddelde vervangen door een semi-parametrisch model. Het semi-parametrische model wordt uitgerust met een smoother voor het schatten van de trend. Dat maakt tevens een analyse van kortetermijntrends mogelijk. De analyse bestaat erin om op elk tijdstip na te gaan of de eerste afgeleide van de niet-lineaire trend significant is. Dat impliceert echter dat een groot aantal testen simultaan moeten worden uitgevoerd. Daarom is een aangepaste methode voor multipliciteitscorrectie geïmplementeerd. Tenslotte hebben we een eerste aanzet gegeven voor het uitbreiden van onze modellen voor niet-normale meetgegevens. Hierbij hebben we ons toegespitst op binaire data. Milieukundige reglementeringen maken dikwijls gebruik van normen die niet mogen worden overschreden. De norm zorgt dus voor een binaire uitkomst voor de beleidsmaker. Opnieuw wordt een latent proces gebruikt voor de modellering van de afhankelijkheidsstructuur. Aan de hand van veralgemeende lineaire gemengde modellen (GLMM) hebben we de binaire uitkomstvariable gemodelleerd. De parameters van een GLMM hebben echter een conditionele betekenis. Ze geven de verandering weer, gegeven een bepaalde waarde voor het latent proces. In milieukundige toepassingen is het echter interessanter om gebruik te maken van marginale modellen. Daarom hebben we gebruik gemaakt van de modelstructuur die door Heagerty and Zeger (2000) werd geïntroduceerd. Aan de hand van het verband tussen conditionele en marginale modellen, kan een specifieke GLMM structuur worden geïdentificeerd die het toelaat om de parameters van het marginaal model rechtstreeks te schatten.

De spatio-temporele aanpak wordt geïllustreerd aan de hand van drie gevallenstudies waarin 5 meetpunten van het IJzerbekken worden beschouwd. Tijdens de bemonsteringsperiode werden mestactieplan I (MAPI)(Vlaams Parlement, 1995) en mestactieplan II (MAPII)(Vlaams Parlement, 1999) van kracht. MAPI werd geïntroduceerd op 1 januari 1996, en MAPII op 1 januari 2000. In de jaarrapporten maakt de VMM gebruik van jaarlijkse gemiddelden. In dit kader wordt het spatio-temporeel model gebruikt om het nitraatgemiddelde van 2003 in het studiegebied te vergelijken met het algemene gemiddelde en de gemiddeldes van de meest recente jaren. In een eerste gevallenstudie wordt aangetoond dat het nitraatgemiddelde in het studiegebied in 2003 heel significant lager is dan het algemene gemiddelde van de volledige bemonsteringsperiode ($p < 0.01$). Tevens blijkt het

nitraatgehalte in de 4 meetpunten van de IJzer gemiddeld significant lager te zijn dan het gemiddelde van de metingen in 2002 en 2001 ($p = 0.03$). In een tweede gevallenstudie wordt de detectie van een niet-lineaire trend in het studiegebied beoogd. Uit de analyse blijkt zich een significant dalende trend voor te doen in het nitraatgehalte tussen september 1999 en januari 2002 ($\alpha = 0.05$). De daling start dus tussen het invoeren van het mestactieplan van 1996 (MAPI) en het mestactieplan van 2000 (MAPII). In de laatste studie wordt nagegaan of beide MAP's een trendbreuk teweeg brachten in de kans op de overschrijding van de nitraatnorm. Het model detecteert een trendbreuk na de implementatie van MAPII. Bovendien blijkt de trendbreuk voldoende groot te zijn om een daling te veroorzaken in de kans dat de nitraatnorm wordt overschreden. De drie gevallenstudies geven een sterke indicatie dat de introductie van de mestactieplannen een gunstig effect heeft op de nitraatstatus in het studiegebied.

Curriculum vitae

Personalia

Naam	Lieven Clement
Geslacht	man
Nationaliteit	Belg
Geboortedatum	22/12/1977
Geboorteplaats	Torhout, België
Burgelijke staat	gehuwd
e-mail adres	Lieven.Clement@UGent.be

Opleiding

10/95 – 07/00	Bio-ingenieur in de milieutechnologie (grote onderscheiding), Universiteit Gent **Scriptie**: Microbiële biosensoren voor organische micropolluenten in gassen. Promotor: Prof. Dr. ir. W. Verstraete

Loopbaanoverzicht-werkervaring

03/03 – heden Assistent, vakgroep LA10, FBW, Universiteit Gent
Ondersteuning voor de opleidingsonderdelen Informatica, Modelleren en Simuleren van Biosystemen, Applied Statistics for the Food Sciences, Informatics en dienstverlening (statistische consulting)

10/02 – 02/03 Doctoraatsbursaal bijzonder onderzoeksfonds, vakgroep LA10, FBW, Universiteit Gent
Modellering binnen de aquacultuur: optimaal ontwerp, optimalisatie en procescontrole van een continu recirculatiesysteem voor rotifeerproductie

08/00 – 09/01 Doctoraatsbursaal, vakgroep LA10, FBW, Universiteit Gent
Onderzoek met betrekking tot het ontwerp van een dynamische data validatie methode van inkomende gegevens in de oppervlaktewater meetdatabank van de VMM

Artikels in tijdschriften die opgenomen zijn in Web of Science

1. Volcke, E.I.P., Clement, L., Van de Steene, M. and Vanrolleghem, P. A. (2001). An overview of the posters presented at Watermatex 2000. I : Sensor / Monitoring, control and decision support systems. *Water Science & Technology* 43(7), 381-386.

2. Clement, L., Thas, O., Vanrolleghem, P.A. and Ottoy, J.P. (2006). Spatio-temporal statistical models for river monitoring networks. *Water Science & Technology* 53(1), 9-15.

3. Clement, L. and Thas, O. (2007). Estimating and modelling spatio-temporal correlation structures for river monitoring networks. *Journal of Agricultural, Biological, and Environmental Statistics*, 12(2), 161-176.

4. Clement, L., Thas, O., Ottoy, J.P. and Vanrolleghem, P.A. (2007). Data management of river water quality data - a semi-automatic procedure for data validation. *Water Resources Research*, accepted.

Congresproceedings

1. Clement, L., Thas, O., Vanrolleghem P.A., Ottoy J.P. (2003). Statistical validation of water quality data. In: Proceedings of Statistics in Public Resources and Utilities and in Care of the Environment 2003 (SpruceVI), Lund, Sweden, June 15-19, 2003.

2. Clement, L., Thas, O., Vanrolleghem, P.A. and Ottoy, J.P. (2004) Spatio-temporal statistical models for river monitoring networks. In: Proceedings of the 6th International Symposium on Systems Analysis and Integration Assessment (WATERMATEX 2004), Beijing, China, November 3-5, 2004.

3. Clement, L. and Thas, O. (2005). Spatio-temporal river monitoring network modelling. In: Proceedings of the Workshop on Recent Advances in Modelling Spatio-Temporal Data, Southampton, United Kingdom, May 25-26, 2005.

4. Clement, L. and Thas, O. (2005) Intervention analysis and trend detection in river monitoring network data. In: Proceedings of the 13th Annual Meeting of the Belgian Statistical Society, Corsendonk, Belgium, October 14-15, 2005.

5. Clement, L. and Thas, O. (2006) Nonparametric trend detection in spatio-temporal river monitoring networks. In: Proceedings of the 3th International workshop on spatio-temporal modelling (METMA3), Pamplona, Spain, September 27-29,2006.

Conferenties, studiedagen en symposia (Mondelinge bijdrage)

1. SPRUCE VI, 2003, Lund, Sweden, June 15-19, 2003.

2. WATERMATEX 2004, 6th International Symposium on Systems Analysis and Integration Assessment, Beijing, China, November 3-5, 2004.

3. 13th Annual Meeting of the Belgian Statistical Society, Corsendonk, Belgium, October 14-15, 2005.

4. International workshop on spatio-temporal modelling (METMA3), Pamplona, Spain, September 27-29, 2006.

Conferenties, studiedagen en symposia (Poster bijdrage)

1. Clement, L. and Thas, O. (2005) Spatio-temporal river monitoring network modelling. Workshop on Recent Advances in Modelling Spatio-Temporal Data, Southampton, United Kingdom, May 25-26, 2005: Award for Best Poster and of the Royal Statistical Society bursary.

Conferenties, studiedagen en symposia (Deelname)

1. WATERMATEX 2000, 5th International Symposium on Systems Analysis and Integration Assessment, Ghent, Belgium, September 18-20, 2000: Member of the local organising commitee.
2. Symposium on Supervision, Control and Optimization of Biotechnological Processes, Ghent, Belgium, February 4, 2002: secretaris.
3. Symposium on Environmental Statistics, Ghent, Belgium, January 20, 2003.
4. 12th Annual Meeting of the Belgian Statistical Society, Vielsalm, Belgium, October 8-9, 2004.
5. Symposium on Statistical Genetics, Ghent, Belgium, May 17, 2005.

Reviewed papers of International journals

Journal of Hydrology: 1 paper

Didactische activiteiten

2005	Oefeningen van de module “Niet-parametrische Methoden” in de opleiding praktijkgerichte statistiek, IVPV, FTW, Universiteit Gent.
2003–heden	Oefeningen in het opleidingsonderdeel “Applied Statistics for the Food Sciences”, Master of Science in Food Science and Technology, FBW, Universiteit Gent.

2003–heden Oefeningen in het opleidingsonderdeel “Modelleren en Simuleren van Biosystemen”, 3de Bachelor in de bio-ingenieurswetenschappen, FBW, Universiteit Gent.

2004–2006 Oefeningen in het opleidingsonderdeel “Informatica”, 2de Bachelor in de bio-ingenieurswetenschappen, FBW, Universiteit Gent.

2003–heden Oefeningen in het opleidingsonderdeel “Informatics and Statistics”, Master of Science in Physical Land Resources, Environmental Sanitation, Aquaculture, FBW, Universiteit Gent.

2006–2007 Begeleiding van Katrijn Van Bastelaere bij haar thesis “Spatio-temporele modellering van de oppervlaktewaterkwaliteit van het IJzerbekken op basis van het fysico-chemisch meetnet van de Vlaamse Milieumaatschappij” voor het behalen van het diploma Master in de statistische data-analyse, FWE, Universiteit Gent.

2006–2007 Begeleiding van Alain Visscher bij zijn thesis “Trend detectie in tijdreeksen van waterkwaliteitsdata” voor het behalen van het diploma Master in de statistische data-analyse, FWE, Universiteit Gent.