Reconstruction of data from the aquatic environment

Søren Nymand Lophaven

LYNGBY 2001
EKSAMENSPROJEKT
NR. 01/2001

IMM

Trykt af IMM, DTU

Preface

This project is a master thesis accounting for 30 of the 300 points required to obtain the engineering master degree at the Technical University of Denmark. The work has been carried out at the new department Informatics and Mathematical Modelling (IMM), Technical University of Denmark, in co-operation with the National Enviromental Research Instistute of Denmark (NERI).

I would like to thank my supervisor Associate Professor Helle Holst (IMM) and Senior Research Scientist Jacob Carstensen (NERI) for their engaged guidance and supervision during my work in the last six months.

Lyngby, the 31th of January 2001

Søren Nymand Lophaven

Abstract

This thesis describes, applies and compares statistical methods for reconstruction of concentrations of Dissolved Inorganic Nitrogen (DIN) and Dissolved Inorganic Phosphorus (DIP) in Kattegat. The measurements are taken in the period from 1993 to 1997 within the monitoring program, which was implemented by the adoption of the Action Plan on the Aquatic Environment in 1987.

The aim of the reconstruction methods is to estimate the concentration of the two species for each week in the five year period, and at any location in Kattegat. The methods are general and could be applied to other parameters.

The spatial distribution of DIN and DIP have been computed by three different variants of kriging, i.e. ordinary kriging, universal kriging and cokriging. In order to have a sufficient number of observations per week, methods for temporal reconstruction of data have to be applied prior to the computation of spatial predictions. Two methods have been used for this purpose, these are the General Linear Model and locally weighted regression. The methods are compared from a statistical and a physical point of view.

Furthermore, the thesis applies different 3 dimensional approaches. These are 3 dimensional kriging, locally weighted regression and an ARIMA model. The first two methods are applied to raw data, i.e. measurements, while the ARIMA model is applied to different stations, where the time series are

temporally reconstructed by the General Linear Model. Kriging is applied to the parameters of the model, in order identify these at any location in Kattegat. Such 3 dimensional methods are much less developed and described in literature compared to methods for analysis of strictly temporal or spatial data.

Keywords:

Dissolved Inorganic Nitrogen, Dissolved Inorganic Phosphorus, the General Linear Model, locally weighted regression, (cross) semivariogram, anisotropy, ordinary kriging, universal kriging, cokriging, sequential conditional simulation, ARIMA modelling

Abstract (in Danish)

I dette eksamensprojekt beskrives, anvendes og sammenlignes statistiske metoder til rekonstruktion af koncentrationen af Opløst Uorganisk Kvælstof (DIN) og Opløst Uorganisk Fosfor (DIP) in Kattegat. Målingerne er udført i perioden fra 1993 til 1997 indenfor rammen af det måleprogram, der blev implementeret i forbindelse med vedtagelsen af Vandmiljøplanen i 1987.

Formålet med rekonstruktionsmetoderne er at estimere koncentrationen af de to stoffer for hver uge i den 5-årige periode, og for enhver lokalitet i Kattegat. Metoderne er generelle, og kan således anvendes til rekonstruktion af andre parametre.

Den spatielle fordeling af DIN og DIP er blevet beregnet med tre forskellige varianter af kriging, disse er ordinær kriging, universal kriging og cokriging. For at have et tilstrækkeligt antal observationer pr. uge anvendes metoder til tidslig rekonstruktion inden de spatielle prædiktioner kan beregnes. To forskellige metoder er blevet anvendt til dette formål, hvilke er den Generelle Lineære Model og lokalt vægtet regression. Metoderne sammenlignes fra en statistisk og fysisk synsvinkel.

Endvidere er forskellige 3 dimensionelle metoder blevet anvendt i eksamenprojektet. Disse er 3 dimensionel kriging, lokalt vægtet regression og en ARIMA model. De to førstnævnte kan anvendes på rådata, dvs. målinger, mens ARIMA modellen anvendes på forskellige stationer, hvor tidsrækken er tidsligt rekonstrueret med den Generelle Lineære Model. Kriging anven-

des til beregning af model parametrene for en vilkårlig lokalitet i Kattegat. Disse 3 dimensionelle metoder er langt mindre udviklede, og beskrevet i litteraturen, sammenlignet med metoder til analyse af strengt tidslige eller spatielle data.

Nøgleord:

Opløst Uorganisk Kvælstof, Opløst Uorganisk Fosfor, den Generelle Lineære Model, lokalt vægtet regression, (kryds) semivariogram, anisotropi, ordinær kriging, universal kriging, cokriging, sekventiel betinget simulation, ARIMA modellering

Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Action Plan on the Aquatic Environment

In connection with water quality and protection of the aquatic environment, the overall goal of the Danish Government is to work towards ensuring that the Danish watercourses, lakes and marine waters are clean and of satisfactory quality as regards health and hygiene.

Since the mid 1980s, a number of plans of action and strategies have been adopted. The primary goal of these was to reduce the nitrogen pollution from agricultural sources. The Action Plan on the Aquatic Environment (APAE) was adopted in 1987, and contained reduction targets for nitrogen and phosphorus by 50 % and 80 % respectively, before 1993. This corresponds to a reduction of annual discharge of nitrogen from 283,000 tonnes to 141,600 tonnes, and of phosphorus from 9,120 tonnes to 1,820 tonnes. Separate goals for reduction of nitrogen and phosphorus were set up for three different sections, which are

- Agriculture
- Municipal wastewater treatment plants
- Individual industrial discharges

The reduction of nitrogen discharge from areas of agriculture accounted for approximately 95 % of the total reduction of nitrogen, which is a reduction

of 127,000 tonnes per year, and was the most critical part of the action plan. The reduction targets for the agricultural sector were to be attained through a number of different measures. The agricultural sector had to establish sufficient capacity to store 9 months of manure production, in order to be sure that it could be stored until the crop growth season started, and also establish crop rotation and fertilization plans to ensure that the nitrogen content of the fertilizer was optimally exploited and absorbed. The fields had to have a green cover during the winter period, which could take up nitrogen in this period. Furthermore, limits on how much livestock manure may be applied to the fields were establised, [Danish-EPA, 2000].

It soon became clear that it would not be possible to reach the reduction targets before 1993, and APAE was therefore further tightened in 1991 in the Action Plan for Sustainable Agriculture, in which the time frame was extended to year 2000. In this plan, it was assumed that APAE would account for a reduction from agriculture of 50,000 tonnes of nitrogen per year, and measures were set up for the remaining 77,000 tonnes.

In 1998 Parliament adopted the Action Plan on the Aquatic Environment II (APAE II) as a supplement to APAE. This plan assumed that the Action Plan for Sustainable Agriculture would account for a reduction from agriculture of 40,000 tonnes of nitrogen per year, which means that together with APAE the total reduction would be 90,000 tonnes of nitrogen from agriculture. In APAE II measures were set up for a reduction of 37,000 tonnes of nitrogen from agriculture before 2003, which will result in a total reduction of annual discharge of nitrogen from agriculture of 127,000 tonnes, which was the original goal of APAE in 1987. This means that if the aim of APAE II is attained, the original goal from 1987 for reduction of nitrogen discharge from agriculture will be reached with a delay of 10 years.

In connection with the adoption of the Action Plan on the Aquatic Environment in 1987, a monitoring program was established to demonstrate the effects of the measures contained in the plan. The monitoring program was revised in 1992 and again in 1997-1998, and resulted in the implementation of the Danish Aquatic Monitoring and Assessment Program, commonly referred to as NOVA-2003. NOVA-2003 contains a number of subprograms, which are:

- Inputs and discharges to soil and water from point sources
- Atmospheric deposition on the sea

- Agricultural monitoring catchments
- Groundwater
- Lakes
- Marine waters

Before the implementation of NOVA-2003 the marine monitoring program aimed specifically at demonstrating effects of the Action Plan on the Aquatic Environment, while NOVA-2003 with its subprograms encompasses the environmental quality of both surface- and groundwater in a broader sense, [Danish-EPA, 2000]. The data that will be used in this thesis is measured within the frame of the Nationwide Monitoring Program under the Action Plan on the Aquatic Environment, in the period from 1993 to 1997.

1.2 Description of data

Data have been measured at 71 different stations in Kattegat during a five year period from 1993 to 1997. The locations of the different stations in Kattegat are shown in figure 1.1. Measurements are carried out by the four Danish counties Frederiksborg, Western Zealand, Århus and Northern Jutland, and by the National Environmental Research Institute of Denmark and some Swedish institutions. A number of parameters, for describing the environmental state, have been measured and can be separated into three groups;

- Biomass
- Nutrients
- Physical parameters

To describe the amount and production of algae, phytoplankton biomass, chlorophyll and primary production have been measured. The nutrient concentrations that are measured, are the concentrations of Dissolved Inorganic Nitrogen (DIN), Dissolved Inorganic Phosphorus (DIP) and Dissolved Silica (DSi). The shortenings DIN, DIP and DSi will be used in this thesis. Physical parameters are temperature and salinity, and difference in salinity has, moreover, been calculated as the difference between salinity in the upper and lower part of the water column.

Data, except for salinity difference and primary production, are given as the average of measurements from the upper 10 meters of the water column.

Primary production is measured as the uptake of CO_2 by phytoplankton per m^2 per day. The units of phytoplankton biomass and chlorophyll are μgC/l and μg/l, respectively, while DIN, DIP and DSi are measured in μmol/l.

The uncertainty of a measurement of one of the variables is a sum of the uncertainty, caused by microvariability, and the uncertainty for the measurement itself, in the laboratory. Microvaribility is the dominating part of the uncertainty. It describes the fact that the true value of a variable at the same location is different within even very small intervals of time, [Carstensen et al., 1999]. Data are inhomogeneously measured in time and space, and this fact leads to the aim of this thesis.

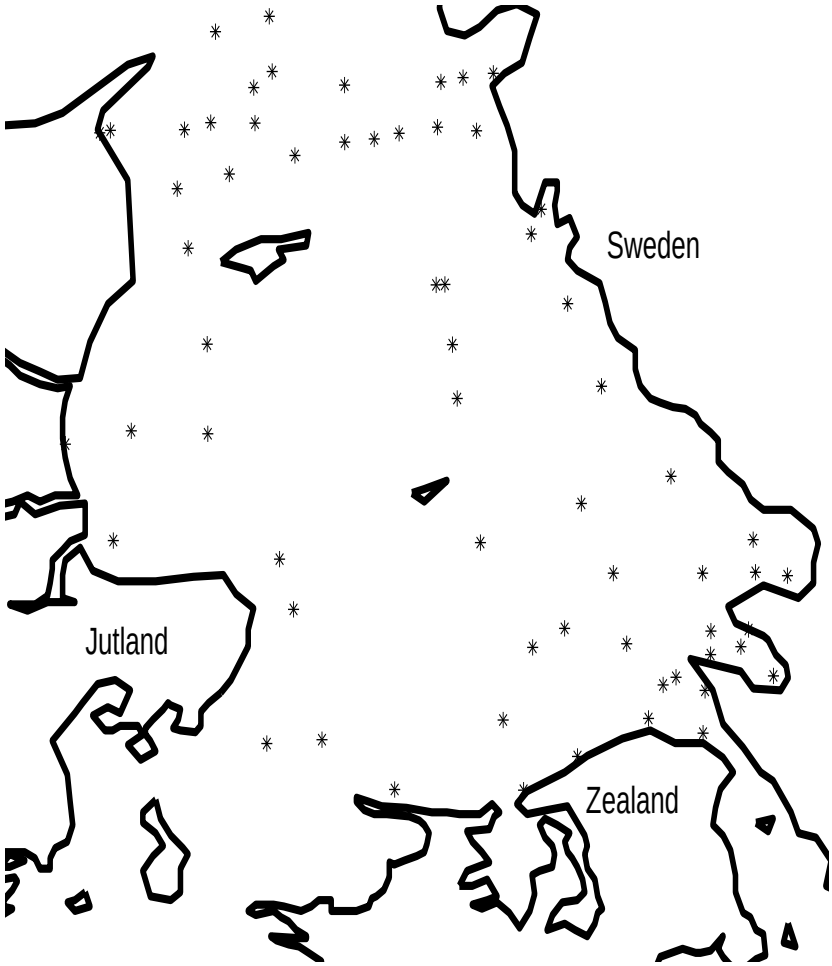

Figure 1.1: *Location of stations in Kattegat. The names of the stations are not shown.*

1.3 The aim of the thesis

The aim of this thesis is to **apply statistical methods for temporal and spatial reconstruction of aquatic environmental data from Kattegat**, i.e. to estimate the missing observations in both time and space, and thereby compute estimates for any time and location in Kattegat.

Such methods are important, since data are inhomogeneously measured in time and space, and in order to obtain improved knowledge of biological processes in the marine environment, the application of reconstruction methods is necessary. Examples of the use of the reconstructed values could be as an input for more complicated models, e.g. models for analysis of time series, for calculating budgets for nutrients and biomass, or for calibration of hydrodynamic, deterministic models. Furthermore, the methods can be used to design monitoring programs, because it is possible to optimize sampling locations in both time and space.

The reconstruction methods that will be described and applied are general, i.e. they can be applied to any of the measured variables, but the work in the present thesis will be limited to applying the methods to measurements of the nutrients DIN and DIP.

1.4 Nutrients in Kattegat

The statistical methods which are described in the thesis, are applied to the nutrients Dissolved Inorganic Nitrogen (DIN) and Dissolved Inorganic Phosphorus (DIP), and a short description of nutrients in Kattegat will therefore be given, as well as a description of the transportation of water.

Kattegat is dominated by transportation of low-saline water from the Baltic Sea and water with a high salinity from the North Sea. The difference in salinity causes a stratification in a depth of 15-20 meters, depending on the season and weather conditions. The general pattern of flow of water in Kattegat is that the low-saline water, in the upper part of the water column is tranported to the north, while the high-saline water at the buttom is transported to the south. This general pattern is strongly affected by the wind.

1.4.1 Discharges of nutrients to Kattegat

Dissolved Inorganic Nitrogen

Dissolved Inorganic Nitrogen (DIN) is the fraction of nitrogen which is implicitly available for biomass growth and includes nitrate NO_3^-, nitrite NO_2^- and ammonium NH_4^+. Approximately 30 % of the total yearly discharge of nitrogen to Kattegat comes from atmospheric deposition. The atmospheric deposition is, besides the natural content of nitrogen in the atmosphere, caused by combustion of fossil fuels at power plants and by other industries, as well as by motorized traffic. Such combustion produces different oxidized nitrogen species, referred to as NO_x. Another source of atmospheric deposition is ammonia, NH_3, which is caused by evaporation from fertilized fields. The atmospheric deposition in Kattegat is highest near the eastern coast of Jutland, [Hansen et al., 2000].

Approximately 70 % of the total yearly discharge of nitrogen originates from land based sources in Denmark and Sweden. The land based sources can be separated into point sources, such as wastewater treatment plants and industry, and non-point sources, e.g. cultivated fields. The main land based input of nitrogen is from fertilized fields. Nitrogen from fields is transported to Kattegat via rivers and fjords or by groundwater. The main

Danish direct input is from Gudenåen and Limfjorden, which drain large agricultured areas in Jutland, and transport water to Kattegat. Also Göta Elven in Sweden transports high amounts of nitrogen to Kattegat.

A part of the water which is transported to Kattegat from the North Sea, comes from the so-called Jutland Current. This current transports water with high concentrations of nutrients, coming mainly from the central European rivers like the Rhine and the Elbe, from the German Bight towards the north along the western coast of Jutland. Depending on weather conditions, this current transports water into Kattegat, and results in increased concentrations of nutrients in the northern part.

Dissolved Inorganic Phosphorus

Phosphate (PO_4^{3-}) is the dissolved fraction of the phosphorus, which is available for biomass growth. Phosphorus is not transported around in the air in the same degree as nitrogen. A few years ago phosphorus was discharged into Kattegat mainly from point sources like wastewater treatment plants and industry. The building of many new wastewater treatment plants in the 1980s has caused the fraction of phosphorus coming from point sources to decrease, as well as the total discharge of phosphorus. Today agriculture discharges almost the same amount of phophorus to Kattegat as point sources. The agricultural part of the discharge is transported to Kattegat via rivers and fjords or in the groundwater, as is the case for nitrogen. [Danish-EPA, 2000]

1.4.2 Dynamics of nutrients

The growth of biomass in Kattegat is throughout most of the year limited by DIN. In the spring, enhanced light conditions and increased temperature cause growth of biomass (phytoplankton), and nutrients are depleted in the photic surface layer. This phenomenon is often referred to as the spring bloom. During the summer the concentration of nutrients remains low, due to high rates of primary production. At the same time, the phytoplankton biomass sediments out of the photic zone and sinks to the seafloor. With the first autumn storms the water column becomes mixed, and nutrient-rich water below the photic zone is mixed into the surface layer. This causes the concentration of nutrients to increase in the upper part of the water

column. In shallow water areas of Kattegat there is a flux of nutrients to the photic zone released from the sediments. [Carstensen et al., 1999]

1.5 Former work

This section summarizes the results of the former work, that has been carried out by the present author, within the area of applying statistical methods for reconstruction of data. It will be shown how the General Linear Model (GLM) and locally weighted regression (LOESS) works, when used for temporal reconstruction of DIN and DIP. Moreover a short summary of the former results of spatial reconstruction will be given. It has been found that the nutrients DIN and DIP are log-normal distributed and a log-transformation of the measured concentrations has therefore been performed prior to the statistical analysis. The theory of GLM and LOESS is described in section 2.2 and 2.3, respectively, while the theory of spatial reconstruction is given in chapter 3.

1.5.1 The General Linear Model

Different variants of the General Linear Model (GLM) has been applied to DIN and DIP. These variants will be referred to as model 1, 2 and 3. Model 1 is a two-sided analysis of variance. Exemplified for DIN it can be written as

$$\log(DIN_{ij}) = station_i + week_j + \epsilon_{ij} \tag{1.1}$$

In order to try to obtain a better fit, especially for stations located in the open sea, model 1 was slightly modified by introducing a new qualitative factor, called *sum_open*. This factor has two levels, i.e. zero and one. If a station is located in the open sea and sampling has been carried out in the summertime, then *sum_open* is one. In all other cases *sum_open* is zero. Open sea is here defined as stations located where the water depth is ≥ 30 meters. Summertime is defined as the months June, July and August. The model is referred to as model 2, and can be written as

$$\log(DIN_{ijk}) = station_i + week_j + sum\_open_k + \epsilon_{ijk} \tag{1.2}$$

Model 2 is actually a three-sided analysis of variance. A third model has been used which is actually the same as model 1, but applied to two different

sets of data, where the first contains open sea stations, and the second coastal stations. This is referred to as model 3, and can be written

$$\begin{aligned} \log(DIN_{ij}) &= station_i + week_j + \epsilon_{ij} \\ &\quad \text{for depth} \geq 30 \text{ meter} \\ \log(DIN_{ij}) &= station_i + week_j + \epsilon_{ij} \\ &\quad \text{for depth} < 30 \text{ meters} \end{aligned} \tag{1.3}$$

The useful thing about the GLM is that it uses the information from surrounding stations in the estimation. However the GLM is only able to compute estimations for stations and weeks where measurements have been carried out.

Model 1

The result of the temporal reconstruction when applying model 1 will be presented for two coastal station, i.e. station 20004 and 4410, and a station located in the open sea, i.e. station 1001. The model has been applied to DIN and DIP.

For station 20004, shown in the upper part of figure 1.2, the estimated time series seems to fit the observations quite well, even though it overestimates the concentration in wintertime for the years 1993, 1995 and 1996.

On the other hand the model underestimates the concentrations for station 1001, which is shown in the lower part of figure 1.2. The reason is that most of the observations, that are used in the estimation, are from coastal stations, and the estimation of DIN therefore resembles the dynamics of the coastal stations.

For station 4410, shown in figure 1.3, the overestimation of the model is very large. This station is located in the county of Northern Jutland, which has a high detection limit for DIN, and this fact is probably the reason for the overestimation. The problem of the detection limits will be discussed further in section 2.2.1.

The temporal reconstruction by applying model 1 has been done for DIP for the stations 20004, 4410 and 1001, and is shown in figure 1.4 and 1.5.

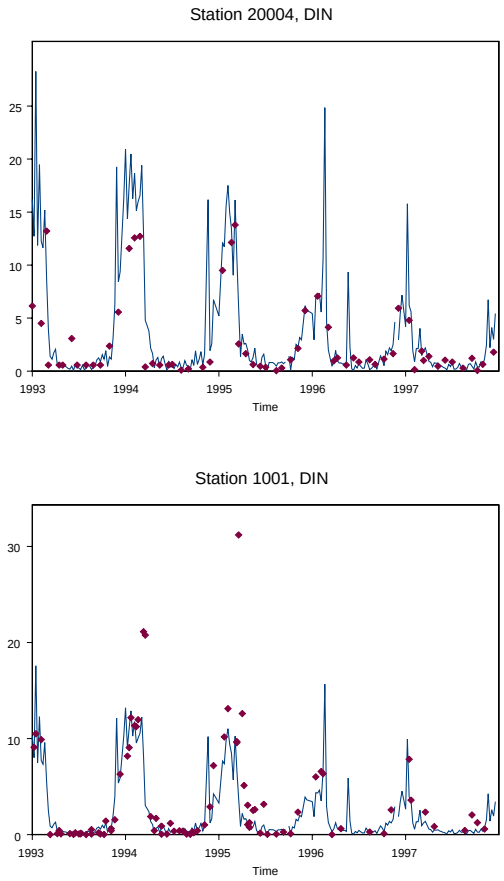

Figure 1.2: *Observed DIN concentrations and the reconstructed time series for stations 20004 and 1001 when applying model 1.*

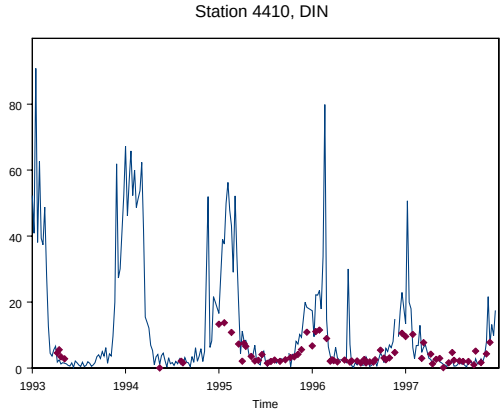

Figure 1.3: *Observed DIN concentrations and the reconstructed time series for station 4410 when applying model 1.*

By comparing the upper and lower part of figure 1.4, the temporal reconstruction of DIP for station 1001, which is an open sea station, seems to be the most accurate. This is the opposite of what was the case for DIN, and the opposite of what would be expected. One reason could be, that the sampling frequency of DIP at station 1001 is very high, which results in a better reconstruction. Furthermore, the recirculation of DIP is greater than for DIN, and therefore the difference in concentrations between open sea and coastal stations is not as big, as is the case for DIN. The model underestimates the DIP concentrations for station 20004. The modelling of DIP at station 4410 works much better than was the case for DIN, because the detection limit for DIP in the county of Northern Jutland is lower and more reasonable.

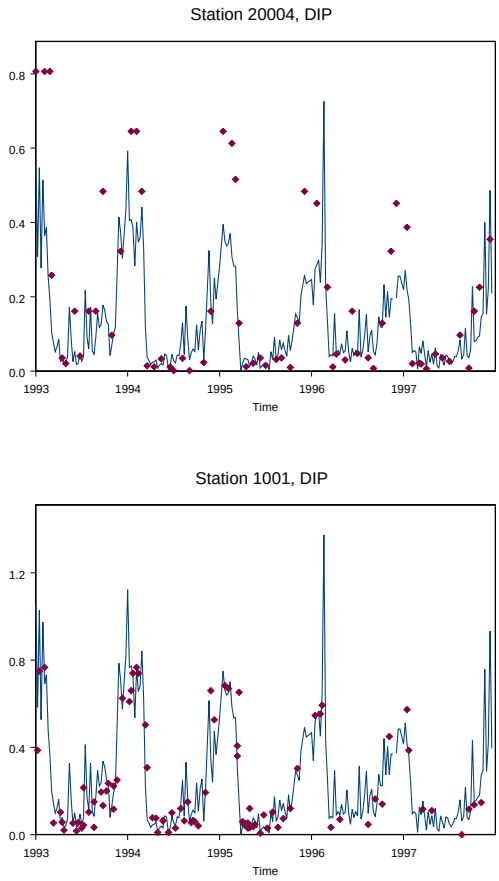

Figure 1.4: *Observed DIP concentrations and the reconstructed time series for stations 20004 and 1001 when applying model 1.*

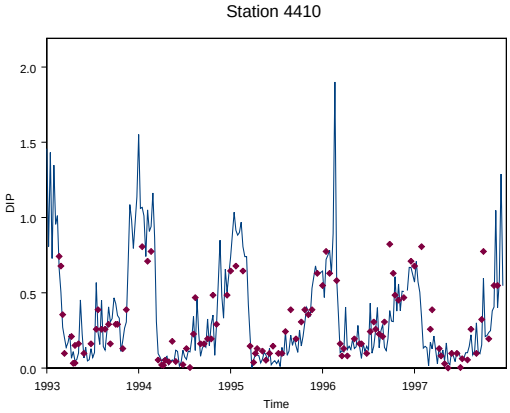

Figure 1.5: *Observed DIP concentrations and the reconstructed time series for station 4410 when applying model 1.*

Model 2 and 3

The results when applying models 2 and 3 are not shown here, but a few comments will be given. Models 2 and 3 were applied to take into account the difference in concentrations between coastal and open sea stations. The difference between the results obtained from model 1 and 2 is inconsiderable, and the introduction of an additional parameter in model 2 cannot be justified.

Model 3 fits better than model 1 and 2, but it does not reconstruct the same amount of data, i.e. it does not fill out as many gaps in the time series, as the two other models. The reason is that the estimation is based on fewer observations, and the model is not able to calculate an estimate for weeks where no observations have been done. This is a disadvantage because the aim of applying the GLM method is to fill out gaps in the time series.

Comparison of the models

The three models have been evaluated using cross validation. In this method a single year from one station is left out of the model estimation each time. Five stations are considered, these are stations 4410, 413, 190004, 1001 and 20004. Afterwards the estimated values are assigned to data, which were left out of the estimation. Exemplified by DIN the goodness of the model is calculated as

$$\text{Goodness Of Model} = \sum_{i=1}^{5}\sum_{j=1}^{5}\left(\frac{1}{m-1}\sum_{k=1}^{m}(\log(\widehat{DIN}_{ijk}) - \log(DIN_{ijk}))^2\right) \tag{1.4}$$

In (1.4) i is the index of station, j is the index of year and k is the index of the m residuals for a given combination of year and station. The result of the cross validation is shown in table 1.1, and shows that model 3 is the best for describing the dynamics of DIN, while model 1 is the best for DIP. The result of the cross validation together with the amount of temporal

Variable	Model	Goodness of model
DIN	1	39.2660
DIN	2	38.2706
DIN	3	37.3285
DIP	1	24.4280
DIP	2	24.4465
DIP	3	25.3985

Table 1.1: *Results of cross validation of model 1, 2 and 3 for DIN and DIP.*

reconstruction, which were computed by the different models, lead to the recommendation of model 1 for temporal reconstruction.

1.5.2 Locally weighted regression

Locally weighted regression (LOESS) only uses information from the station under consideration, and it can therefore only be used for the stations with the highest sampling frequencies. The number of data points to use in the local regression is given as a fraction of the total number of observations, and is determined by Akaike's information criterion, see section 2.3.

This fraction is referred to as the bandwidth. A second order polynomia is used for the local fitting, and the parameters of the polynomia are found by weighted least squares, where the weight is high for data points close to the point where we want to compute an estimation.

Akaike's information criterion for DIN as a function of the size of the local area is shown in figure 1.6 for station 20004 and 1001, and the following values of the bandwidth are used:

Station 20004: 0.18
Station 1001: 0.18

This means that 18% of the total number of data points are used in the local estimation for both stations. The result of the temporal reconstruction of

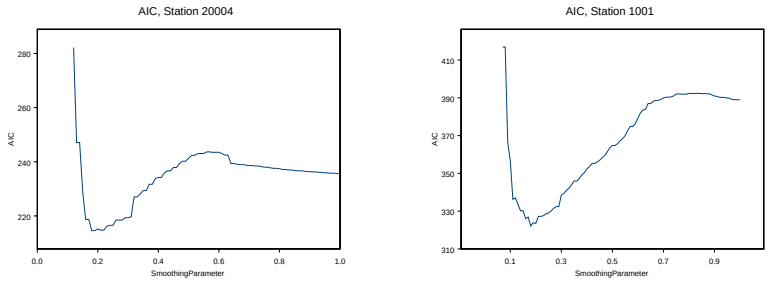

Figure 1.6: *Akaike's information criterion for DIN, used for determination of the size of the local area. Left: Station 20004. Right: Station 1001.*

DIN is shown in figure 1.7. This estimation is a smoother curve, than what was found when using the GLM, and it does not estimate the extreme peaks which seem to be caused by overfitting of the GLM.

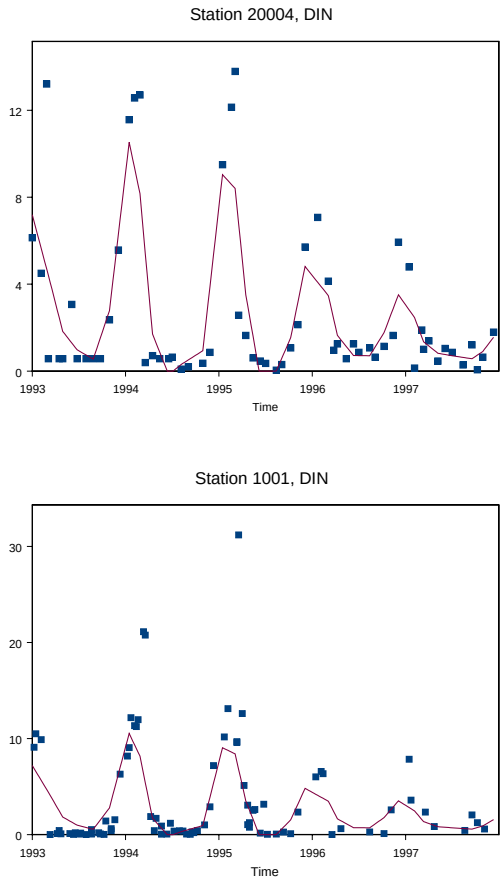

Figure 1.7: *Result of temporal reconstruction of DIN using LOESS. Upper: Station 20004. Lower: Station 1001.*

1.5.3 Spatial reconstruction

The spatial reconstruction was based on the estimations from the GLM, and was computed by ordinary kriging and cokriging assuming isotropy. The methods gave very similar results. The highest concentrations were calculated in the coastal areas, especially along the eastern coast of Jutland and in the north-eastern part of Kattegat.

1.6 Content of thesis and reading guide

It is the intention of the present author to write a thesis which can be read by people with interest in marine ecology and environment and by statisticians, although the main part of the work is within statistics. The thesis consists of seven chapters, which are:

Chapter 1: Introduction
Chapter 2: Temporal data analysis
Chapter 3: Spatial data analysis
Chapter 4: Spatiotemporal data analysis
Chapter 5, 6 and 7: Discussion, conclusion and future work

- The introduction has been given in the present chapter, and is about the Action Plan on the Aquatic Environment, the former work that has been carried out and a description of data that will be used for the analysis. This chapter is not very theoretical and does not assume any particular skills within statistics.

- Chapter 2 and 3 describe how gaps in time series at the different stations in Kattegat are filled out by estimated values, and how to predict the size of a physical magnitude at any location in space, from a number of point observations. Each of the two parts are split up into a number of sections, each one describing a method for either temporal or spatial reconstruction of data. The theory of the methods is described in each section, and the result of the application of the theory is shown in the same section. This is done to avoid having one big part describing the theory, and afterwards another part showing the results. The theory of many of the methods implies some statistical knowledge, although the aim has been to write these sections in an understandable way. The results of each chapter are summarized at the end.

- What is done in chapter 2 and 3, is to estimate the size of a physical magnitude at any time and location in two steps, i.e. first a temporal reconstruction is computed, and afterwards the results from this are used in the spatial reconstruction. Chapter 4 describes how these two steps can be directly combined. Some of the methods are basically the same as those that are used in chapter 2 and 3, and the theory is therefore not repeated.

- The last chapters discuss the performance of the methods, and from this discussion, make the final conclusions of the work, which has been done. At the end suggestions for future work are given.

Chapter 2
Temporal data analysis

2.1 Introduction to temporal data analysis

Temporal data analysis includes methods for estimating the size of a physical magnitude Z at any time t, i.e. reconstruction of time series by filling out gaps with estimated values. The temporal resolution that will be used is one week. It is not possible to use a higher resolution because the sampling frequency is not high enough. The dynamics of nutrients and biomass is fast, and consequently a lower resolution, of e.g. one month, is too low to describe the dynamics in a reasonable way.

Two different approaches will be used for reconstruction; these are the General Linear Model (GLM) and locally weighted regression (LOESS). Temporal reconstruction using GLM can only be done, at locations and for weeks, where measurements have been carried out, while LOESS is able to compute weekly estimations for stations with a high sampling frequency. GLM uses information from the surrounding stations, while LOESS only uses data from the station under consideration.

2.2 The General Linear Model

This section describes how temporal reconstruction is calculated using the General Linear Model (GLM). The results will not be presented and explained here, because the most important part of these has already been given in section 1.5.1. Instead it will be shown how different substitutions, of values below the detection limit, affect the result of the estimation, and the effect of the choice of time interval will also be examined. Temporally reconstructed values are used as the basis for the spatial interpolation.

The dependent variable in the linear model can be affected by qualitative or quantitative factors, where qualitative factors refer to non-numerical values or levels, and quantitative factors refer to numerical values. The linear model that will be used, is the following.

$$Z_{ij} = station_i \cdot week_j \cdot \epsilon_{ij} \tag{2.1}$$

The temporal reconstruction will be done for variables Z, which are assumed to be log-normal distributed and the multiplicative model (2.1) becomes additive as shown in (2.2).

$$\log(Z_{ij}) = station_i + week_j + \epsilon_{ij} \tag{2.2}$$

i is the index of *station*, and goes from 1 to the number of stations in the model, and j is the index of *week*, and goes from 1 to the number of weeks in the model. The fitted value of Z can only be calculated for weeks and stations where measurements of Z have been carried out. The model does not contain any cross effects since the highest sampling frequency is one week.

Both *week* and *station* are qualitative factors, and the model is actually a 2-sided analysis of variance. The model can be described with a matrix notation in a general form.

$$\underline{Y} = \underline{\boldsymbol{X}} \cdot \underline{\beta} + \underline{\epsilon} \tag{2.3}$$

The vector $\underline{Y}$ contains the log-transformed n observations of Z. $\underline{\epsilon}$ has mean 0 and a variance $\sigma^2\underline{\boldsymbol{\Sigma}}$, and it contains n elements. $\underline{\boldsymbol{X}}$ is the designmatrix with dimension $n \times k$ containing indicator variables. Indicator variable

means a one for a combination of *week* and *station* for which a measurement has been done, and a zero at all the other places in the matrix. This means that the designmatrix has two ones in each row. $\underline{\beta}$ is a vector containing the k parameters of the model. $\underline{\beta}$ is estimated as

$$\underline{\hat{\beta}} = (\underline{\boldsymbol{X}}^T \underline{\boldsymbol{\Sigma}}^{-1} \underline{\boldsymbol{X}})^{-1} \underline{\boldsymbol{X}}^T \underline{\boldsymbol{\Sigma}}^{-1} \underline{Y} \tag{2.4}$$

with a dispersion matrix

$$D(\underline{\hat{\beta}}) = \sigma^2 (\underline{\boldsymbol{X}}^T \underline{\boldsymbol{\Sigma}}^{-1} \underline{\boldsymbol{X}})^{-1} \tag{2.5}$$

The dispersion matrix in (2.5) has the dimension $k \times k$. The fitted log-transformed data are calculated as

$$\underline{\hat{Y_f}} = \underline{\boldsymbol{X_f}} \cdot \underline{\hat{\beta}} \tag{2.6}$$

$\underline{\boldsymbol{X_f}}$ has the dimension $m \times k$, where m is the number of possible combinations of the two factors *week* and *station*. $\underline{\hat{Y_f}}$ contains m elements and has the dispersion matrix

$$D(\underline{\hat{Y_f}}) = \underline{\boldsymbol{X_f}} D(\underline{\hat{\beta}}) \underline{\boldsymbol{X_f}}^T \tag{2.7}$$

with a dimension of $m \times m$. The fitted values of Z can now be calculated as

$$\underline{\hat{Z_f}} = \exp(\underline{\hat{Y_f}} + diag(D(\underline{\hat{Y_f}}))/2) \tag{2.8}$$

The calculation operations in (2.8) are done for each element in the vectors, [Edwards, 1985], [Edwards, 1984] and [Carstensen et al., 2000].

2.2.1 Results

When GLM is used for temporal reconstruction, the size of the detection limit is an important factor for the stations, where this magnitude is high. Statistically the detection limit defines the limit below which measurements are assumed to be zero, i.e. it defines the lowest concentration which can be measured. Detection limits depend on the measured variable and the laboratory that performs the analysis. In the four Danish counties Århus, Northern Jutland, Western Zealand and Frederiksborg, concentrations below the detection limit are given the value of the detection limit. The detection limits are not known, but have been found as cut-off values, by

County	DIN	DIP
Århus	0.71	0.065
Northern Jutland	1.43	0.065
Western Zealand	0.32	0.048
Frederiksborg	0.93	0.16

Table 2.1: *Detection limits for DIN and DIP in four danish counties, found as cut-off values.*

investigating plots of the variables against time. The values for the four Danish counties are shown in table 2.1 for DIN and DIP. Table 2.1 shows that the detection limit for DIN in Northern Jutland county is high. This means that the lowest possible concentration of DIN is 1.43 μmol/l in this county. The general level of concentrations at stations in Northern Jutland county is therefore high, and results in an overestimation when using GLM. Furthermore, the measured summer concentrations of DIN at the stations in Northern Jutland are much higher than corresponding concentrations at other stations. This might be due to problems with the measurements or the sampling, e.g. contamination of the samples, which results in a lot of measurements in the interval from 1.43 μmol/l to 3.0 μmol/l.

It is shown in section 1.5.1, that the temporal reconstruction of DIN using GLM works quite well for stations 190004 and 1001. Here measurements below the detection limit are substituted by random numbers between zero and the detection limit, from a uniform distribution. The result of the corresponding computation for station 4410 in Northern Jutland county is shown at the upper plot in figure 2.1. This shows that the GLM is overestimating the concentration of DIN, due to the problems listed above. The lower plot shows the reconstruction of DIP for the same station. In this case the GLM performs much better.

One could try to improve the result, by substituting the measurements below the detection limit in another way, and also cope with the many high measurements of summer concentrations, by substituting these as well. Two examples are shown in figure 2.2. Note that the y-axes in the figure are scaled differently, and also different from figure 2.1. The upper part shows the result after substituting all measurements below 2 μmol/l, by random numbers between 0 and 0.1 from a uniform distribution, while the

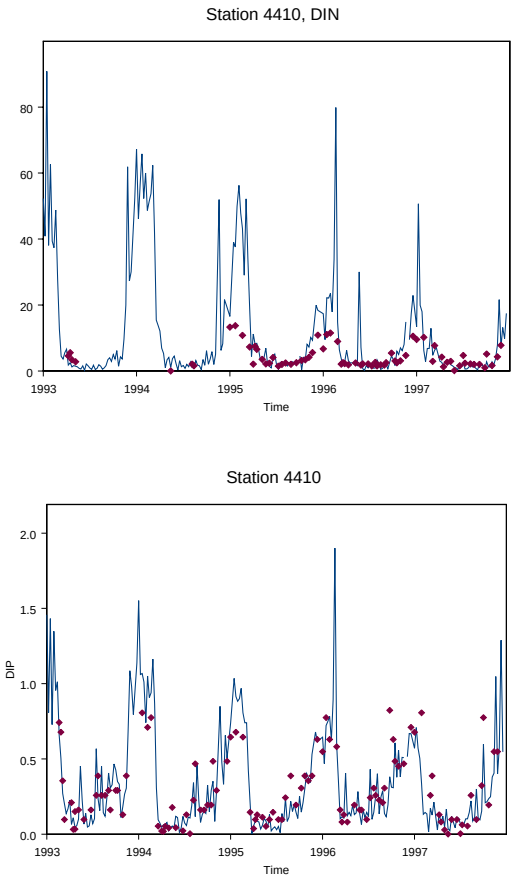

Figure 2.1: *Temporal reconstruction at station 4410 using GLM. Upper: DIN. Lower: DIP.*

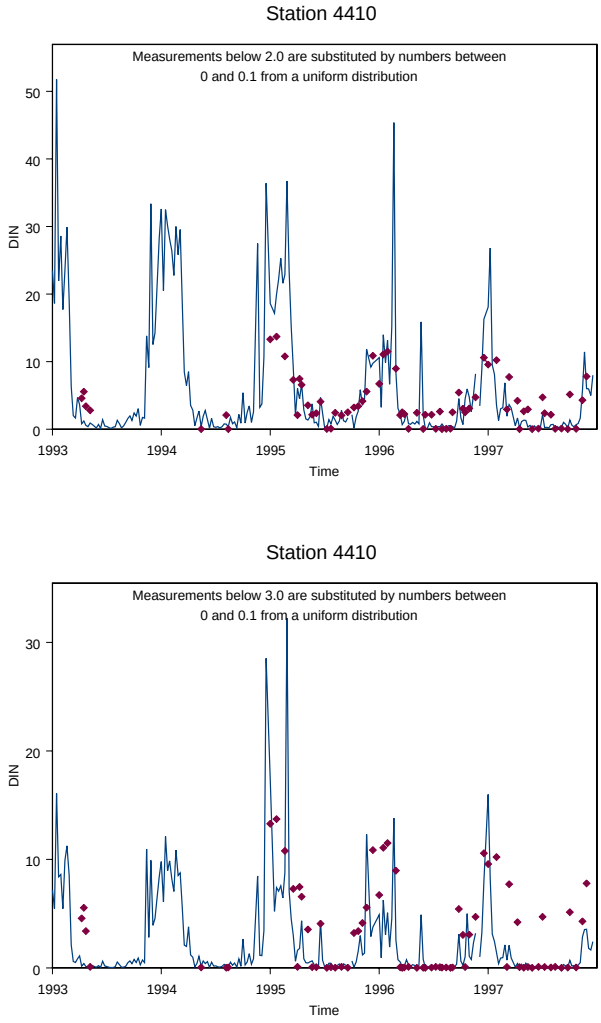

Figure 2.2: *Temporal reconstruction at station 4410 using GLM. Upper: The results in the case where measurements below 2 µmol/l are substituted by random numbers between 0 and 0.1 from a uniform distribution. Lower: The result in the case where measurements below 3 µmol/l are substituted by random numbers between 0 and 0.1 from a uniform distribution.*

lower part shows a similar substitution of all measurements below 3 μmol/l. Thus it is assumed that the concentration of DIN is almost 0 in summertime. It is seen that allowing the concentration of DIN to reach 0 in the summer significantly improves the estimation.

The low concentrations of DIN could also be computed by modelling, i.e. by an observed relationship between DIN and a number of independent variables. Significant relationships have been found between the log-transformed concentrations of DIN and temperature for the four stations of Northern Jutland, where measurements of DIN have been taken. These are 3302, 3310, 4402 and 4410. An example of the modelling is shown in figure 2.3, which is based on measurements from stations 4402 and 4410, because the regression line can be assumed to be the same for these two stations. The regression line on the figure is given by

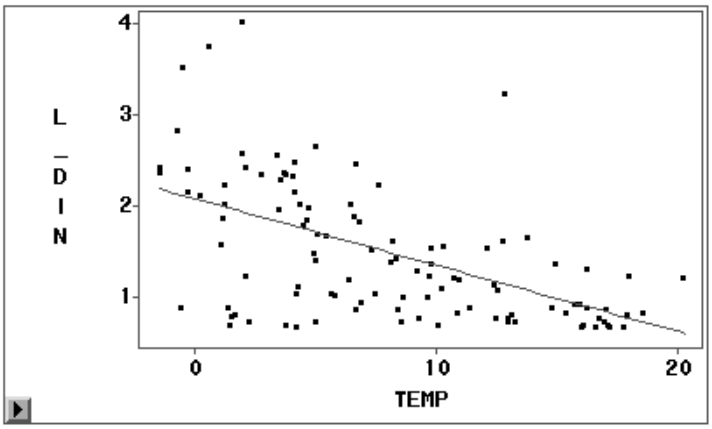

Figure 2.3: *The relationship between log-transformed DIN and temperature for the stations 4402 and 4410.*

$$\log(DIN) = 2.14 - 0.079 \cdot Temp \tag{2.9}$$

The main problem is that both log(DIN) and temperature are measured with uncertainty. If instead temperature is modelled as a function of

log(DIN), the regression line is given by

$$\begin{aligned} Temp &= 14.25 - 4.11 \cdot \log(DIN) \Longleftrightarrow \\ \log(DIN) &= 3.47 - 0.2435 \cdot Temp \end{aligned} \tag{2.10}$$

As shown in (2.10), the equation can be recalculated to describe log(DIN) as a function of temperature, and the two equations for computing log(DIN) are not the same. If either (2.9) or (2.10) is used to compute low values of log(DIN), it results in a bad temporal reconstruction for the four stations.

Instead it is assumed that the concentration of DIN is almost 0 in summertime, and all measurements below 2 μmol/l are therefore substituted by random numbers between 0 and 0.1 from a uniform distribution. In order not to substitute measurements which are actually correct, values in the interval between 2 μmol/l and 3 μmol/l are not substituted. Some of these are from spring and autumn, and they are therefore more realistic.

To obtain a high amount of information, a weekly temporal resolution will be used. This resolution is high compared to the sampling frequency. Consequently the individual estimations are based on a small number of observations, and extreme measurements therefore result in extreme estimations, i.e. GLM is overfitting. Instead a lower temporal resolution could be used, which results in a smoother curve of the reconstructed values, as shown in figure 2.4 and 2.5 for the time intervals one, two and four weeks.

The model corresponding to each of the three different time intervals has been evaluated using cross validation, in the same way as described in section 1.5.1, i.e. measurements from a single year and station are left out of the estimation each time. Five stations are considered, these are station 4410, 413, 190004, 1001 and 20004. Afterwards the estimated values are assigned to data, which were left out of the estimation, and the goodness of the model is calculated as

$$\text{Goodness Of Model} = \sum_{i=1}^{5}\sum_{j=1}^{5}\left(\frac{1}{m-1}\sum_{k=1}^{m}(\log(\widehat{DIN}_{ijk}) - \log(DIN_{ijk}))^2\right) \tag{2.11}$$

In (2.11) i is the index of station, j is the index of year and k is the index of the m residuals for a given combination of year and station. The result of the cross validation is shown in table 2.2. The GLM performs slightly better when time intervals of two and four weeks are used. The

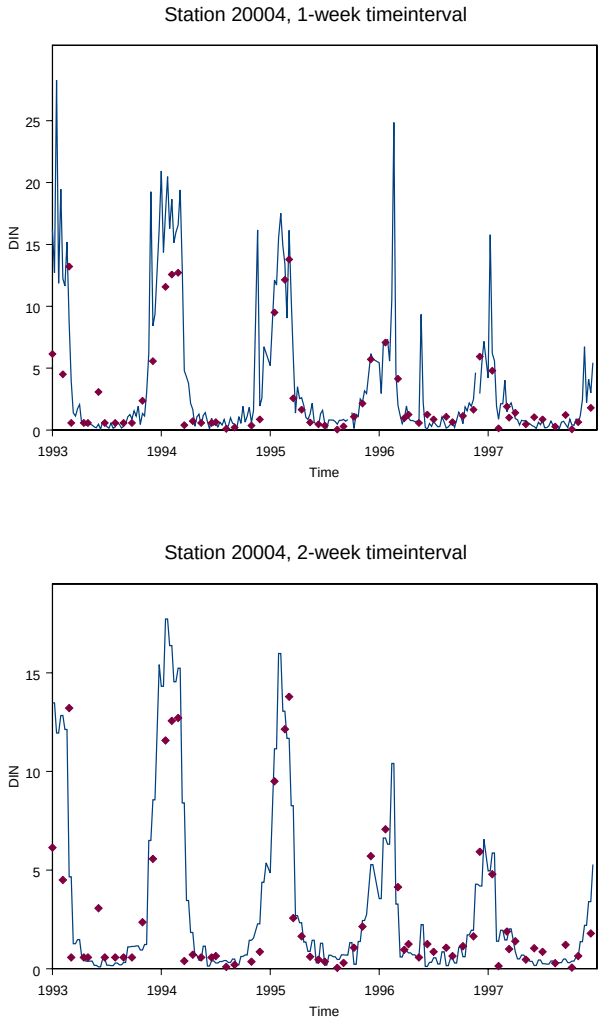

Figure 2.4: *Temporal reconstruction of DIN at station 20004 using GLM with different time intervals. Upper: The results for a time interval of one week. Lower: The result for a time interval of two weeks.*

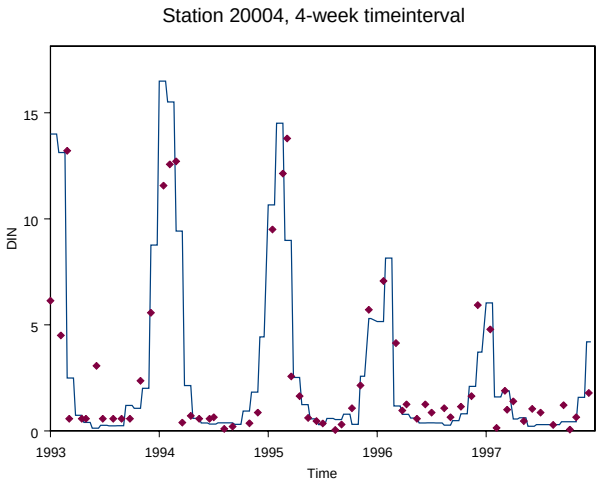

Figure 2.5: *Temporal reconstruction of DIN at station 20004 using GLM with a time interval of four weeks.*

Time interval	Goodness Of Model
1 week	39.266
2 weeks	36.672
4 weeks	37.400

Table 2.2: *Cross validation of the General Linear Model for three different intervals of time.*

disadvantage is that estimated values are not computed each week, and the weekly temporal resolution will therefore be kept.

2.3 Locally weighted regression

This section describes how temporal reconstruction is calculated using locally weighted regression (LOESS). The results will not be presented and explained here, because the most important part of these is shown in section 1.5.2. LOESS will in this thesis be applied to 1 and 3 dimensional data. The theory described in this section is general, and will not be repeated when applying LOESS in 3 dimensions.

LOESS is a nonparametric estimation method, which computes an unknown regression function. Suppose that the dependent variable is described by n observations of y_i. The dependent variable is a function of p independent variables. The n corresponding observations of the independent variables are denoted by $x_i = (x_{i1}, \ldots, x_{ip})$. In the case of temporal reconstruction, the dependent variable is only a function of one independent variable, i.e. time. The relationship between observations of the dependent variable y_i and observations of the independent variables x_i is

$$y_i = g(x_i) + \epsilon_i \tag{2.12}$$

where g is the regression function and ϵ_i are independent normal distributed variables with a mean 0 and a variance σ_ϵ^2.

The idea of LOESS is that data within a neighbourhood around a point x in the p-dimensional space can be approximated by fitting a regression surface, i.e. a regression function $\hat{g}(x)$, to data. The estimation of the dependent variable $\hat{y}$ in the point x is given by $\hat{g}(x)$. The approximation is done by weighted least squares, where points in the neighbourhood are weighted according to the distance from x. Points close to x are given a higher weight than those further away. The size of the neighbourhood is chosen by the value of a smoothing parameter $f = q/n$, which is the fraction of the total number of observations n, that are used in LOESS. The parameter f is usually called the nearest neighbour bandwidth. The above definition of f means that the q nearest neighbours are used. The locally

weighted regression requires a weight function, which is often defined as

$$W(u) = \begin{cases} (1-|u|^3)^3 & |u| < 1 \\ 0 & |u| > 1 \end{cases} \tag{2.13}$$

The weight corresponding to the ith observation in the neighbourhood of a point x is calculated as

$$w_i(x) = W\left(\frac{\parallel x - x_i \parallel}{d(x)}\right) \tag{2.14}$$

where $\parallel x - x_i \parallel$ is the Euclidean distance between x and x_i. $d(x)$ is the distance of the q-nearest x_i to x. By combining (2.13) and (2.14) it is seen that the weights $w_i(x)$ decrease when x_i increase in distance from x, [Cleveland, 1979], [Cleveland, 1988] and [Nielsen, 1997]. If the independent variables x are in the 1-dimensional space, and polynomias of second order are used in the estimation, then the estimated value $\hat{y_i}$ can be written as

$$\hat{y}_i = \hat{g}(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 \tag{2.15}$$

The values of β are found by minimizing (2.16).

$$\sum_i w_i(x)(y_i - \hat{y}_i)^2 \tag{2.16}$$

When applying LOESS the bandwidth f has to be chosen. This can be done using Akaike's information criterion (AIC). Different variants of AIC have been suggested. In the case of a parametric regression AIC is given as

$$AIC = n \log \hat{\sigma_\epsilon}^2 + 2(p+1) \tag{2.17}$$

where n is the number of observations, p is the number of parameters and $\hat{\sigma_\epsilon}^2$ given as

$$\hat{\sigma_\epsilon}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 \tag{2.18}$$

The common idea of the different AIC variants, is that the criterion is a function of the goodness of the fit and the complexity of the model, i.e. the criterion has the form

$$\log(\hat{\sigma_\epsilon}^2) + \psi(L) \tag{2.19}$$

where ψ is a so-called penalty function, which decreases by increasing smoothness of the fit. L is the smoothing matrix, that satisfies

$$\hat{y} = Ly \tag{2.20}$$

Akaike's information criterion is a general method for determination of the number of parameters in a model. It could also be used to calculate the goodness of model for the GLM approach. When using LOESS it is very important to incorporate the complexity of the model. If not, one could continue to improve the estimation by lowering the bandwidth, which causes an overfitting. The principle is shown in figure 2.6. The smoothing

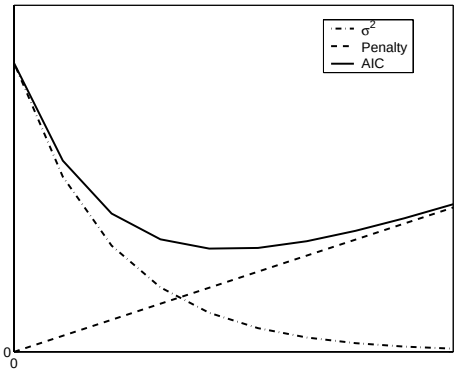

Figure 2.6: *The principle of Akaike's information criterion.*

parameter is selected as the one that minimizes the criterion, i.e. as the one corresponding to the minimum of the AIC curve in figure 2.6. For nonparametric regression methods the trace of the matrix L, i.e. the sum of the diagonal elements, $\sum l_{ii}$, can be interpreted as the effective number of parameters. In this case AIC is

$$\begin{aligned} AIC &= n \log \hat{\sigma_\epsilon}^2 + 2(\text{trace}(L) + 1) \qquad \text{where} \\ \hat{\sigma_\epsilon}^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 = \frac{y^T(I-L)^T(I-L)y}{n} \end{aligned} \tag{2.21}$$

The AIC criterion has a tendency to include too many parameters, leading to an overfitting. This bias can be corrected for by introducing a new variant of AIC, called AIC_{c1}, which is shown in equation (2.22), [Hurvich and Simonoff, 1998], [Awad, 1996] and [McQuarrie et al., 1997].

$$AIC_{c1} = n \log(\hat{\sigma_\epsilon}^2) + n\frac{\delta_1/\delta_2(n+\nu_1)}{\delta_1^2/\delta_2 - 2} \tag{2.22}$$

where

- n = the number of observations.
- $\delta_1 = \text{trace}(I - L)^T(I - L)$
- $\delta_2 = \text{trace}((I - L)^T(I - L))^2$
- $\nu_1 = \text{trace}(L)$

2.3.1 Results

Estimations of the concentrations of DIN and DIP have been computed using locally weighted regression in one dimension. A second order polynomia is used, and the bandwidth is found by AIC_{c1}. The computation can only be done at the stations with the highest sampling frequencies, i.e. at 24 stations in Kattegat. The stations and optimal bandwidths are shown in appendix B. Univariate statistics for AIC_{c1} are shown in table 2.3.

Variable	Minimum	Maximum	Mean	Variance
log(DIN)	0.12	0.38	0.25	0.0023
log(DIP)	0.12	0.33	0.23	0.0028

Table 2.3: *Univariate statistics of AIC_{c1} for the one dimensional LOESS.*

2.4 Summary of temporal data analysis

Two methods for temporal reconstruction of DIN and DIP have been applied. These are the General Linear Model (GLM) and locally weighted regression (LOESS). The GLM uses the information from surrounding stations, while LOESS only uses information from the station under consideration. A temporal resolution of one week is used, i.e. weekly estimations are computed by both methods.

The results of temporal reconstruction using GLM are shown in section 1.5.1, while the importance of substitution of DIN concentrations below the detection limit has been examined in this chapter. The detection limit of DIN in the county of Northern Jutland is high, and the GLM therefore overestimates the concentration of DIN for stations in this county. To allow the concentration to reach 0 in the summer, it was chosen to substitute all values below 2 μmol/l by random numbers between 0 and 0.1 from a

uniform distribution.

It has also been examined how the estimation of DIN depends on the choice of time interval, when GLM is used for temporal reconstruction. The estimation is improved when time intervals of two and four weeks are used, compared to a weekly resolution, but weekly estimations are not computed, and consequently we do not obtain the same amount of information. The weekly temporal resolution is therefore kept.

Estimations computed by LOESS do not depend a lot on the choice of time interval or the way that measurements below the detection limit are substituted. When applied to temporal data, it can only be used at stations with a high sampling frequency. The bandwidth, i.e. the fraction of observations to include in the local regression, is optimized by a bias corrected version of Akaike's information criterion.

By applying the two methods for temporal reconstruction of data, we have computed two sets of data which can be used for weekly spatial predictions. These datasets consist of reconstructed data for weeks where no measurement has been carried out, while the measured value is used for weeks where sampling has been done. The next chapter describes methods for computing spatial predictions, which is based on the two sets of data, as illustrated in figure 2.7. The sets of data generated by GLM contain a higher number of observations, i.e. reconstructed or measured values, than the corresponding dataset generated by LOESS. The number of observations for each week in the two cases are:

GLM: 65
LOESS: 24

The reason for computing spatial predictions based on both sets of data, is that we want to examine whether or not the spatial distribution of DIN and DIP can be computed from only 24 observations per week, i.e. from data temporally reconstructed by LOESS. This would indicate that measurements from stations with a low sampling frequency are unnecessary.

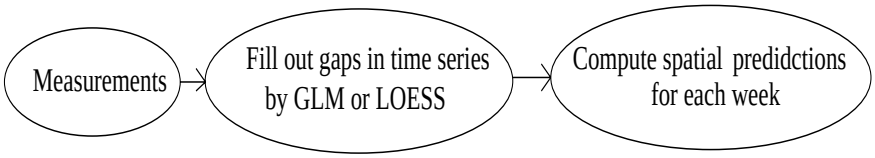

Figure 2.7: *Principle of the computation of predictions. The temporally reconstructed data are used to compute spatial predictions.*

Chapter 3
Spatial data analysis

3.1 Introduction to spatial data analysis

The section gives an introduction to spatial statistics, which includes methods for predicting the size of a physical magnitude Z at any location s. Kriging is such a method, which uses the spatial correlation in data for prediction of the spatial distribution of a physical magnitude. Predictions are done on the basis of a number of point observations $Z(s_i)$ at locations s_i. An unknown value of the magnitude at location s_0 can be calculated as

$$\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i) \tag{3.1}$$

where λ_i is the weight for the i'th point observation and n is the number of point observations. It is often required that $\sum \lambda_i = 1$, and in this case the kriging predictor is a weighted mean of the point observations where the observations which are close to $Z(s_0)$ are given a higher weight than those further away. This is done in a way that minimizes the prediction variance. Figure 3.1 shows the prediction problem. The computation of the kriging weights is based on a semivariogram, which describes the spatial variability of a physical magnitude. Before the weights λ_i and kriging predictions can be calculated, the semivariogram has to be computed and modelled. The spatial prediction therefore consists of the intermediate steps shown

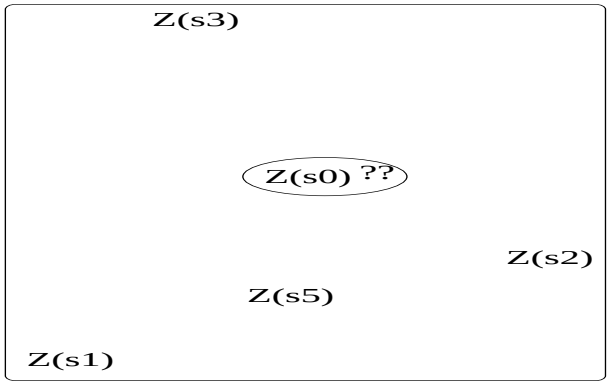

Figure 3.1: *The prediction problem of kriging.*

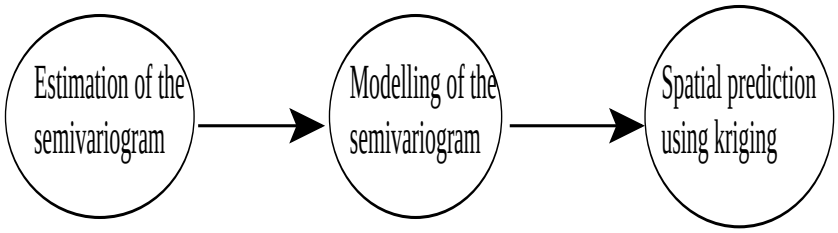

Figure 3.2: *The different steps for computing kriging predictions.*

in figure 3.2. In spatial statistics observations of a physical magnitude are considered as realizations of a stochastic process. At each location in a domain $\mathcal{D}$ a regionalized measureable variable $z(s)$ exists, which is a realization of a stochastic variable $Z(s)$. The quantity of stochastic variables $\{Z(s)|s \in \mathcal{D}\}$ makes up a stochastic function Z. The stochastic variable $Z(s)$ has a mean given by

$$E[Z(s)] = \mu(s) \tag{3.2}$$

and a function of covariance given by

$$Cov[Z(s+h), Z(s)] = C(s,h) \tag{3.3}$$

where h is a vector characterized by a length, which is called the separation distance, and a direction between observations. The separation distance is

$$h = \sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2} = \sqrt{h_x^2 + h_y^2 + h_z^2} \tag{3.4}$$

A stochastic function is first order stationary if the mean of the function is constant $\mu(s) = \mu$ in the domain $\mathcal{D}$. In this case the function of covariance is written as

$$C(s,h) = E[(Z(s+h) - \mu)(Z(s) - \mu)] \tag{3.5}$$

Second order stationarity is obtained if the mean is constant $\mu(s) = \mu$, and the function of covariance only depends on the distance between observations, i.e. $C(s,h) = C(h)$. $C(0) = \sigma_z^2$ is the variance of the stochastic variable $Z(s)$. In many practical applications of kriging, a hypothesis of second order stationarity of Z, with a finite variance $C(0) = \sigma_z^2$ is not satisfied by data. Instead, the intrinsic hypothesis is assumed to be fulfilled. The intrinsic hypothesis is less stringent. It assumes that the first order increments of Z have a finite variance, which means that the increments themselves are second order stationary. If the function of covariance of Z is second order stationary, the intrinsic hypothesis will always be fulfilled, but the opposite is not true. The intrinsic hypothesis leads to the definition of the semivariogram, [Nielsen, 1994] and [Ersbøll, 1994].

3.2 Spatial variability

Determination of the spatial variability is often based on a semivariogram, while the spatial variability between two correlated physical magnitudes is

described by the cross semivariogram. This section goes through estimation and modelling of the semivariogram and cross semivariogram. Furthermore it is shown how anisotropy is taken into account. Modelling of semivariograms and cross semivariograms is done in the same way, and to denote this the term (cross) semivariogram is used. The estimation and modelling of the semivariogram is computed for Dissolved Inorganic Nitrogen (DIN) and Dissolved Inorganic Phosphorus (DIP).

3.2.1 Estimation of the semivariogram

When the intrinsic hypothesis is assumed the variance of increments of Z defines a new function called the semivariogram, which is used to describe the spatial variability of data. The semivariogram is defined as

$$\gamma(h) = \frac{1}{2}var[Z(s+h) - Z(s)] = \frac{1}{2}E[(Z(s+h) - Z(s))^2] \tag{3.6}$$

The semivariogram only depends on the distance between observations, and is independent of the location of point observations. This means that the semivariogram is a constant function in the area under consideration. If second order stationarity is assumed the relationship between the function of covariance and the semivariogram is given as

$$\gamma(h) = C(0) - C(h) \tag{3.7}$$

where $C(0) = \sigma_z^2$ is the variance of the stochastic variable Z. To estimate the semivariogram, ergodicity of the increments of Z are assumed. The ergodic assumption is said to be fulfilled if a second order stationary process's second order variation can be estimated consistently, i.e. the estimated semivariogram $\hat{\gamma}(h)$ converges to the true semivariogram, as the sample size increases. If the process is assumed to be second order stationary, and Z is normally distributed, then the ergodic assumption can be examined by plotting the function of covariance. If $C(h) \to 0$ for $h \to \infty$ the process can be assumed to be ergodic. If this is not the case it might mean, that the estimation is unreliable, no matter how large the sample size is, [Cressie, 1993]. The estimation of the semivariogram is calculated in the following way

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{k=1}^{N(h)} [z(s_k + h) - z(s_k)]^2 \tag{3.8}$$

where $N(h)$ is the number of pairs of observations separated by the distance h. With n observations the number of pairs becomes $N(h) = \frac{n(n-1)}{2}$. Equation (3.8) is often referred to as the classical estimator. Another way of estimating the semivariogram is by using the robust estimate.

$$\hat{\gamma}(h) = \frac{\frac{1}{2}\left(\frac{1}{N(h)}\sum_{i=1}^{N(h)}|z(s+h)-z(s)|^{1/2}\right)^4}{\left(0.457+0.494/N(h)\right)} \tag{3.9}$$

This estimate is more resistant to outliers. By estimating and plotting both kinds of semivariograms, the influence of outliers can be examined, [Cressie, 1993].

3.2.2 Estimation of the cross semivariogram

The spatial variability between two correlated physical magnitudes is described by the cross semivariogram. When the intrinsic hypothesis is assumed, it is defined as

$$\gamma_{ij}(h) = \gamma_{ji}(h) = \frac{1}{2}E[(Z_i(s+h)-Z_i(s))(Z_j(s+h)-Z_j(s))] \tag{3.10}$$

where i and j denote two different variables. Moreover the cross semivariogram is symmetric in h, i.e.

$$\gamma_{ij}(h) = \gamma_{ij}(-h) \tag{3.11}$$

An estimator of the cross semivariogram is

$$\hat{\gamma}_{ij}(h) = \frac{1}{2N(h)}\sum_{k=1}^{N(h)}[z_i(s_k+h)-z_i(s_k)][z_j(s_k+h)-z_j(s_k)] \tag{3.12}$$

where $N(h)$ is the number of pairs of observations separated by the distance h. The estimation can also be based on semivariograms, by introducing a new variable $Z_i(s) + Z_j(s)$, and estimating the semivariogram $\hat{\gamma}_{i+j}(h)$ for this. In this case the estimator becomes

$$\hat{\gamma}_{ij}(h) = \frac{1}{2}[\hat{\gamma}_{i+j}(h) - \hat{\gamma}_i(h) - \hat{\gamma}_j(h)] \tag{3.13}$$

The estimated (cross) semivariogram is often referred to as experimental or empirical, [Cressie, 1993].

3.2.3 Modelling the (cross) semivariogram

Modelling of semi- and cross semivariograms is done in the same way. To denote this the term (cross) semivariogram is used. The estimated (cross) semivariogram is fitted with a model, and the best model is used in the kriging estimation. The most widely used models are the spherical, the exponential and the Gaussian model. Other models for fitting a (cross) semivariogram exist, but will not be presented here. The spherical model is given by

$$\gamma^*(h) = \begin{cases} 0 & h = 0 \\ C_0 + C_1\left(\frac{3}{2}\frac{h}{R} - \frac{1}{2}\frac{h^3}{R^3}\right) & 0 < h < R \\ C_0 + C_1 & h \geq R \end{cases} \tag{3.14}$$

where C_0 is called the nugget effect, R is the range and $C_0 + C_1$ is the sill. The nugget effect is caused by measurement errors and microvaribility. The sill $C_0 + C_1$ is defined as $\sigma^2 = lim_{h\to\infty}\gamma(h)$. The form of the exponential model is given by

$$\gamma^*(h) = \begin{cases} 0 & h = 0 \\ C_0 + C_1\left(1 - \exp(-\frac{h}{R})\right) & 0 < h \end{cases} \tag{3.15}$$

and the Gaussian model by

$$\gamma^*(h) = \begin{cases} 0 & h = 0 \\ C_0 + C_1\left(1 - \exp(-\frac{h^2}{R^2})\right) & 0 < h \end{cases} \tag{3.16}$$

The three (cross) semivariogram models are shown in figure 3.3. The estimated (cross) semivariogram can also be modelled as a sum of the basic models, $\gamma_i^*(h)$, presented above

$$\gamma^*(h) = \sum_{i=1}^{n} \gamma_i^*(h) \tag{3.17}$$

In this case each basic model $\gamma_i^*(h)$ is called a nested structure. When modelling the (cross) semivariogram it is important not to use too many nested structures, because this would lead to an overfitting of the estimated semivariogram.

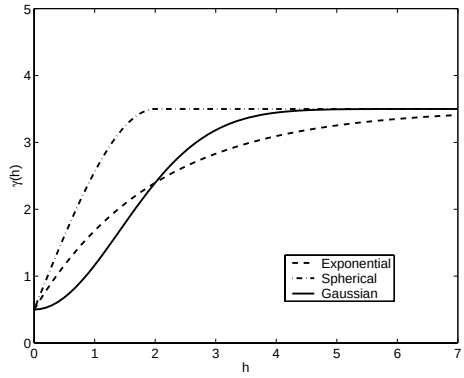

Figure 3.3: *Spherical, Gaussian and exponential (cross) semivariogram models. The model parameters are: R=2 , C_0=0.5 , C_1=3.*

If only a single structure is used for modelling the estimated (cross) semivariogram, some rules for choosing the best kind of model exist. The estimated (cross) semivariogram often shows a parabolic behaviour near the origin, and the Gaussian model will in this case provide the best fit, see figure 3.3. The Gaussian model without a nugget effect has a tendency to lead to a singular solution when the kriging system of linear equations is solved. A nugget effect $\neq 0$ should therefore always be specified when the Gaussian model is used. If, on the other hand, the estimated (cross) semivariogram has linear behaviour near the origin, the exponential or spherical model should be used, [Isaaks and Srivastava, 1989].

Traditionally fitting of the (cross) semivariogram is done by eye, because it has been shown that predictions computed by kriging are reasonably insensitive to the specification of the (cross) semivariogram model. However to get most information out of data, the best model should be used, and the fitting of the model parameters by eye should, instead, be used as starting values in the optimization algorithm, [Cressie, 1985]. The best (cross) semivariogram model can be found using the least squares criterion, i.e. the model parameters are taken as being those, which minimize the sum of

the squared residuals.

$$\sum_{j=1}^{k}[\hat{\gamma}(h_j) - \gamma^*(h_j, \theta)]^2 \tag{3.18}$$

In (3.18) $\hat{\gamma}(h_j)$ is the estimated value of the (cross) semivariogram in lag j, and $\gamma^*(h_j, \theta)$ the value of the (cross) semivariogram model in lag j. The unknown parameters are denoted by θ. Instead of this criterion [Cressie, 1985] suggests minimizing the weighted sum of squares

$$\sum_{j=1}^{k} N(h_j)\Big[\frac{\hat{\gamma}(h_j)}{\gamma^*(h_j, \theta)} - 1\Big]^2 \tag{3.19}$$

$N(h_j)$ is the number of pairs of data in lag j. The criterion in (3.19) gives higher weights to (cross) semivariogram values, where the estimation is based on a higher number of data pairs.

Now that the different (cross) semivariogram models and the parameters have been introduced, it will be explained how the kriging predictions are affected by the choice of model and model parameters. We recall from section 3.1 that the kriging prediction at location s_0 in calculated as

$$\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i) \tag{3.20}$$

where λ_i is the weight for the i'th point observation and n is the number of point observations. Figure 3.3 shows that the Gaussian model computes lower values of the semivariance for small separation distances compared to the exponential. This means that points close together are modelled as being more correlated when the Gaussian model is used, and points close to s_0 are therefore given higher weights λ_i, than the weights for the same points, when the exponential model is used. The prediction therefore becomes more local when using the Gaussian model compared to the exponential. The spherical model is even more local, because the slope of this (cross) semivariogram function is steeper for small separation distances, i.e. it reaches the constant sill for a lower value of h.

Increasing the nugget effect, while the other parameters are kept constant, causes the kriging weights to be more similar, because smaller weights are

given to points close to s_0. This means that higher weights must be given to points further away from s_0. If the nugget effect is decreased the opposite argumentation is used. Changing the sill, while the other parameters are kept constant, does not affect the kriging weights, but the kriging variance is changed, i.e. increasing the sill increases the kriging variance. Changing the range, while the other parameters are kept constant, is known to have only a minor effect on the kriging weights, [Isaaks and Srivastava, 1989]. If the range is decreased more points will be uncorrelated, and the prediction becomes more local. When the range is changed the effect on the kriging variance is more important than the effect on the kriging weights. Increasing the range means that data points are more correlated, and the kriging variance is therefore decreased. The largest kriging variance occurs when all data points are uncorrelated.

3.2.4 Handling anisotropy

The observed spatial correlation or variability, described by the (cross) semivariogram, is not always the same in different directions. This is due to the underlying physical process, which evolves differently in space. The phenomenon is referred to as anisotropy and results in different (cross) semivariograms in different directions. The directional (cross) semivariograms are estimated by only including pairs of data which are located in a certain direction relative to each other. In case of isotropy, where the spatial correlation is the same in all directions, the omnidirectional (cross) semivariogram, based on all data pairs is used for describing the spatial variability.

In literature there is some disagreement about the terminology used for describing the different kinds of anisotropy. Traditionally the term geometric anisotropy is used when the range changes with direction, while the sill is constant. This is the kind of anisotropy mostly observed in nature. [Journel and Huijbregts, 1978] use the term zonal anisotropy for all kinds of anisotropy that are not geometric, e.g. the sill or both sill and range change with direction. [Zimmerman, 1993] uses the more descriptive terms range anisotropy, sill anisotropy and nugget anisotropy, while [Isaaks and Srivastava, 1989] use geometric anisotropy to describe changes of range with direction and constant sill, and zonal anisotropy when the sill changes while the range remains constant. In this thesis the last definitions of geometric- and zonal anisotropy will be used. However only

geometric anisotropy will be described in the following part of the section, because this is the kind of anisotropy which is mostly observed in nature. An example of geometric anisotropy is shown in figure 3.4. Anisotropy is

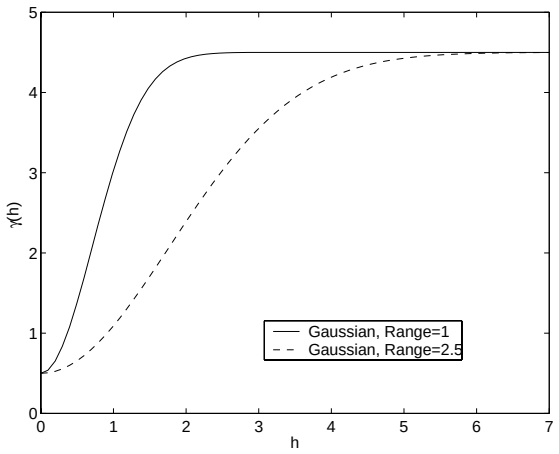

Figure 3.4: *(Cross) semivariograms for two different directions, showing geometric anisotropy.*

described by a direction and an anisotropy ratio. The first thing to do is to identify the directions for the maximum and minimum range. This can be done by investigating a contour map of the (cross) semivariogram surface. These directions determine the axes of anisotropy, i.e. the axes of a new coordinate system. The anisotropy ratio is the maximum range divided by the minimum. Anisotropy ratios $>$ 2-3 change the kriging predictions markedly, compared to the situation where the area is considered as isotropic, [Kaluzny et al., 1998]. We now want to achieve isotropy by making a rotation and a recalculation of the data coordinate system. The original data coordinate system is rotated until the axes of the coordinate system coincide with the axes of anisotropy. In the case of a three dimensional system, two rotations are done, as shown in figure 3.5. The rotation angles are given by the directions of maximum and minimum range. The new components of the separation distance, using a matrix formulation,

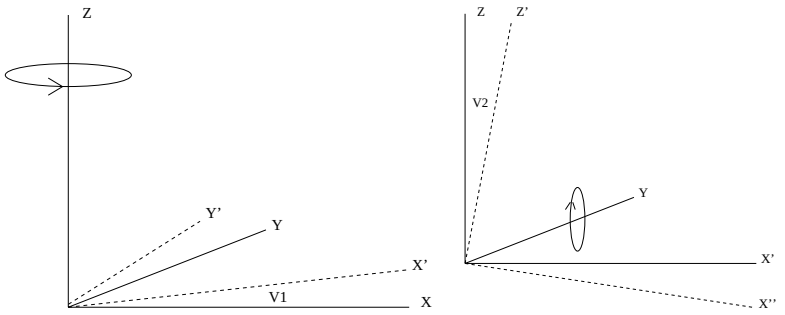

Figure 3.5: *The rotation of the data coordinate system. The definition of v1 and v2 is shown.*

are given by

$$\begin{bmatrix} h'_x \\ h'_y \\ h'_z \end{bmatrix} = \begin{bmatrix} \cos(v1)\cos(v2) & \sin(v1)\cos(v2) & \sin(v2) \\ -\sin(v1) & \cos(v1) & 0 \\ -\cos(v1)\cos(v2) & -\sin(v1)\sin(v2) & \cos(v2) \end{bmatrix} \begin{bmatrix} h_x \\ h_y \\ h_z \end{bmatrix} \tag{3.21}$$

Equation (3.21) rotates the data coordinate system, given by the original axes x, y and z, into new axes x'', y' and z', which define the anisotropic coordinate system.

The recalculation is done by dividing the components, h_x, h_y and h_z, of the original separation distance h by the directional ranges, a_x, a_y and a_z, i.e. changing the separation distance from h to $h_1 = h/a$. In two and three dimensions the new separation distance becomes

$$h_1 = \sqrt{(\frac{h_x}{a_x})^2 + (\frac{h_y}{a_y})^2} \quad \text{and} \quad h_1 = \sqrt{(\frac{h_x}{a_x})^2 + (\frac{h_y}{a_y})^2 + +(\frac{h_z}{a_z})^2} \tag{3.22}$$

Zonal anisotropy can be catered for by dividing the separation distance with a large value of the directional range. To obtain a new coordinate system in which the (cross) semivariogram is isotropic, both a rotation and a recalcuation has to be done, i.e

$$h'_1 = TRh \tag{3.23}$$

where

$$h_1' = \begin{bmatrix} h_{1x}' \\ h_{1y}' \\ h_{1z}' \end{bmatrix}, \quad h = \begin{bmatrix} h_x \\ h_y \\ h_z \end{bmatrix}, \quad T = \begin{bmatrix} \frac{1}{a_x} & 0 & 0 \\ 0 & \frac{1}{a_y} & 0 \\ 0 & 0 & \frac{1}{a_z} \end{bmatrix},$$

$$R = \begin{bmatrix} \cos(v1)\cos(v2) & \sin(v1)\cos(v2) & \sin(v2) \\ -\sin(v1) & \cos(v1) & 0 \\ -\cos(v1)\cos(v2) & -\sin(v1)\sin(v2) & \cos(v2) \end{bmatrix}$$

The new separation distance becomes

$$h_{transformed} = \sqrt{h_{1x}'^2 + h_{1y}'^2 + h_{1z}'^2} \tag{3.24}$$

3.2.5 Results

The spatial variability has been determined for the variables DIN and DIP for data, reconstructed or measured, from four different weeks. The four weeks represent the four seasons, and are

Week 1776: A week in the middle of January 1994.
Week 1837: A week in the middle of March 1995.
Week 1854: A week in the middle of July 1995.
Week 1919: A week in the middle of October 1996.

Both omnidirectional and directional semivariograms are estimated and modelled. The optimal parameters and semivariogram model are found as those which minimize the sum of the squared residuals. The parameters of the omnidirectional semivariograms are shown in table 3.1, for data which are measured or temporally reconstructed by GLM, while table 3.2 shows the parameters for data which are measured or temporally reconstructed by LOESS. A larger range is found when data are temporally reconstructed by LOESS.

Figure 3.6 and 3.7 shows estimated and modelled omnidirectional semivariograms for log-transformed DIN and DIP for a week in the middle of January 1994. The semivariogram in figure 3.6 is based on values which are measured or reconstructed by GLM, while the semivariogram in figure 3.7 is based on values which are measured or reconstructed by LOESS.

The spherical semivariogram model is preferred, since the experimental semivariogram seems to be linear near the origin.

Week	Variable	Model	Range	Sill	Nugget effect
1776	log(DIN)	Spherical	19	0.50	0.10
1776	log(DIP)	Spherical	25	0.20	0.10
1837	log(DIN)	Spherical	22	0.52	0.25
1837	log(DIP)	Spherical	34	0.32	0.00
1854	log(DIN)	Spherical	15	0.90	0.20
1854	log(DIP)	Spherical	19	0.40	0.00
1919	log(DIN)	Spherical	42	0.79	0.30
1919	log(DIP)	Spherical	34	0.41	0.00

Table 3.1: *Type of semivariogram model and the parameters range, sill and nugget effect of omnidirectional semivariograms. Data are temporally reconstructed with GLM.*

Week	Variable	Model	Range	Sill	Nugget effect
1776	log(DIN)	Spherical	97	1.10	0.00
1776	log(DIP)	Spherical	70	0.20	0.00
1837	log(DIN)	Spherical	48	0.20	0.00
1837	log(DIP)	Spherical	152	0.45	0.05
1854	log(DIN)	Spherical	30	3.90	0.70
1854	log(DIP)	Spherical	72	1.30	0.20
1919	log(DIN)	Spherical	92	1.80	0.60
1919	log(DIP)	Spherical	30	0.45	0.05

Table 3.2: *Type of semivariogram model and the parameters range, sill and nugget effect of omnidirectional semivariograms. Data are temporally reconstructed with LOESS.*

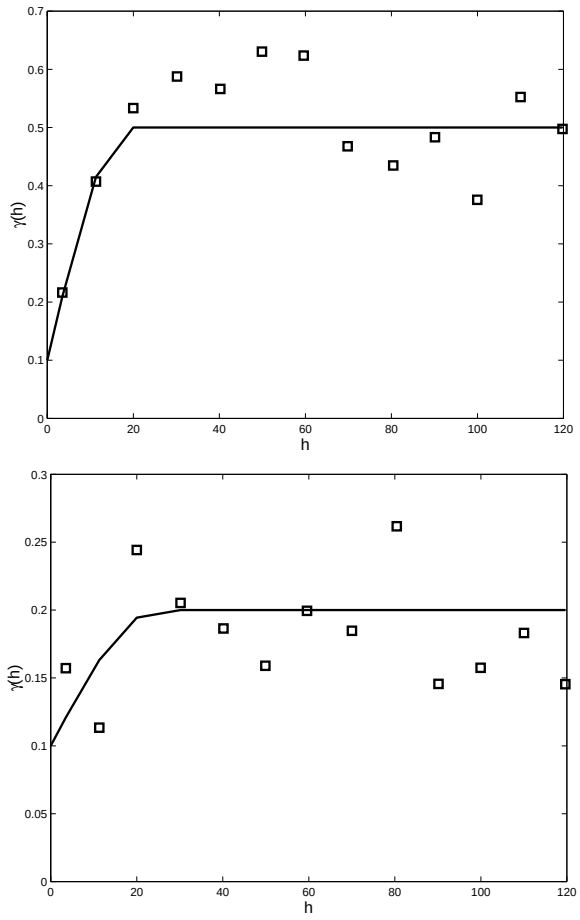

Figure 3.6: *Omnidirectional semivariograms, based on values which are measured or reconstructed by GLM, for a week in the middle of January 1994. Upper: Log-transformed DIN. Lower: Log-transformed DIP.*

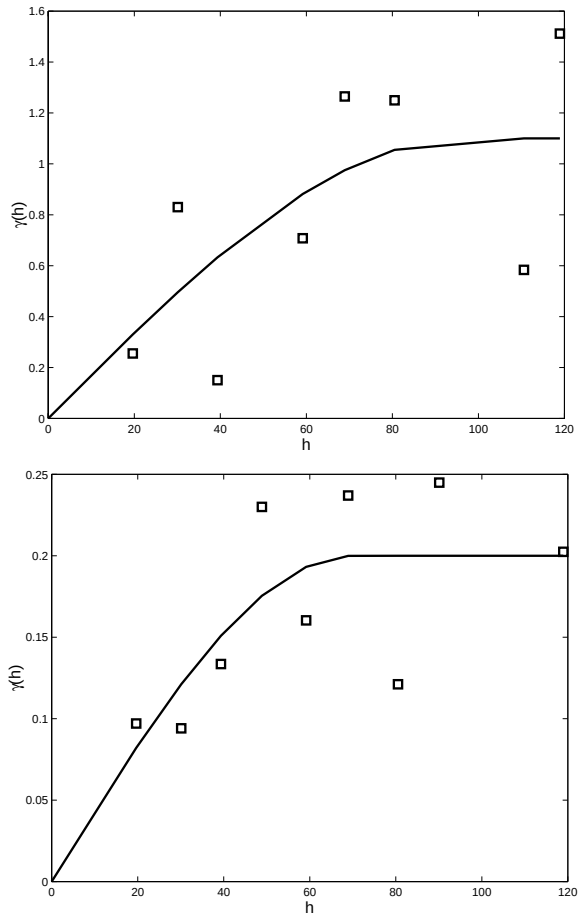

Figure 3.7: *Omnidirectional semivariograms, based on values which are measured or reconstructed by LOESS, for a week in the middle of January 1994. Upper: Log-transformed DIN. Lower: Log-transformed DIP.*

When GLM is used for temporal reconstruction a nugget effect is found for log-transformed DIN for all the four weeks, while it is only found for log-transformed DIP in the winter-week. The range of both log-transformed DIN and DIP are for most weeks found to be between 20 to 40 kilometers.

The spatial variability does not differ very much between the seasons, when GLM is used for temporal reconstruction. The reason is that data are dominated by reconstructed values. Purely reconstructed data result in the same semivariogram for different weeks. According to equation (2.2), reconstructed data from one week are computed by adding a constant contribution from the factor *week* to the different values of the factor *station*. This means that the only difference between reconstructed data from two different weeks is the addition of another constant, corresponding to the factor *week*.

When LOESS is used for temporal reconstruction a greater variability of the semivariogram model between the four weeks are found. The range is found to be higher, with values in the interval from 30 to 152 kilometers.

To identify the anisotropy directions in Kattegat, maps of the semivariogram surface have been computed for log-transformed DIN and DIP in the four weeks. This has only been done for data temporally reconstructed by GLM, because this results in a dataset with more points within one week. The set of data reconstructed by LOESS only includes values for stations with high sampling frequencies, and the number of data points within one week is not high enough to compute any directional semivariograms based on these values. The maps of the semivariogram surface are shown in figure 3.8 for a week in the middle of January 1994, while maps for the three other weeks are shown in appendix C. The maps in figure 3.8 make it easy to determine the directions of anisotropy. The highest range is found in the direction 135 degrees clockwise from the north, while the perpendicular direction, i.e. 45 degrees clockwise from the north, gives the minimum range for both variables. Consequently semivariograms have been estimated and modelled in these directions. In this thesis we will only allow for geometric anisotropy. When modelling the range in different directions it is therefore assumed that the sill and nugget effect are the same. The spherical semivariogram model is used, since this was found to be the best for fitting the experimental omnidirectional semivariogram.

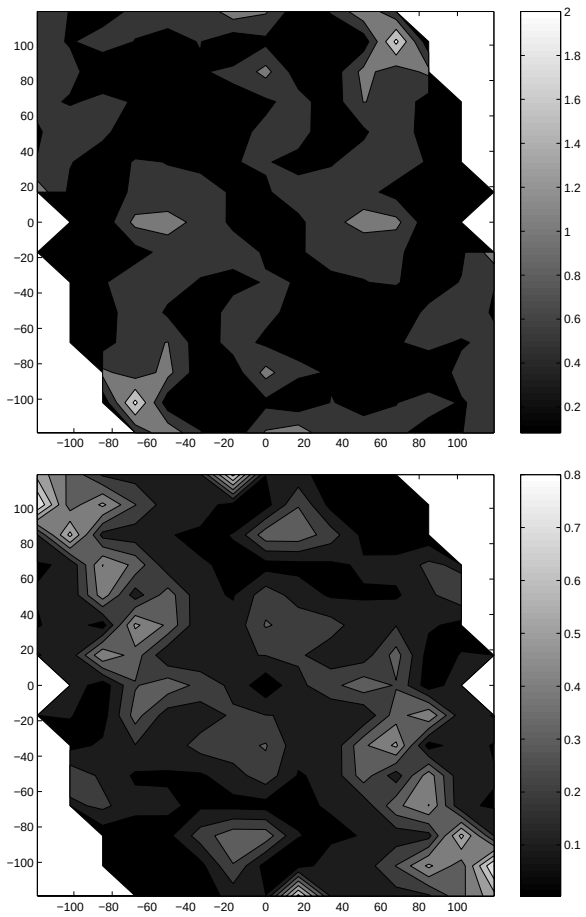

Figure 3.8: *Maps of the semivariogram surface in a week in the middle of January 1994. Data are temporally reconstructed by GLM. Upper: Log-transformed DIN. Lower: Log-transformed DIP.*

The result of modelling of the range in different directions is shown in table 3.3 for the two variables and four weeks. Figure 3.9 shows estimated and

Week	Variable	Direction	Range
1776	log(DIN)	135	40.0
1776	log(DIN)	45	17.0
1776	log(DIP)	135	34.0
1776	log(DIP)	45	22.0
1837	log(DIN)	135	55.0
1837	log(DIN)	45	13.0
1837	log(DIP)	135	73.0
1837	log(DIP)	45	24.0
1854	log(DIN)	135	21.0
1854	log(DIN)	45	5.00
1854	log(DIP)	135	23.0
1854	log(DIP)	45	17.0
1919	log(DIN)	135	59.0
1919	log(DIN)	45	27.0
1919	log(DIP)	135	34.0
1919	log(DIP)	45	34.0

Table 3.3: *Range for directional semivariograms. Data are temporally reconstructed by GLM.*

modelled directional semivariograms for log-transformed DIN and DIP for a week in the middle of January 1994. The anisotropy ratio is given as the range in the 135 degree direction, which is the highest range, divided by the range in the 45 degree direction, and is shown in table 3.4. In general anisotropy ratios higher than 2-3 change the kriging predictions considerably. The highest anisotropy ratios are found for DIN in the spring and autumn-week. Since temporally reconstruction with LOESS does not result in enough data points for computing weekly directional semivariogram, the same anisotropy ratios and directions of anisotropy will be assumed when using LOESS for computing spatial predictions.

The cross semivariogram has been estimated and modelled for the same four weeks as the semivariogram. The depth of water, measured in meters is used as the secondary variable. Other secondary variables have been tried, e.g. salinity. However the estimated cross semivariogram between log-transformed DIN and salinity, as well as between log-transformed DIP and salinity, did not show any spatial correlation. Table 3.5 shows the

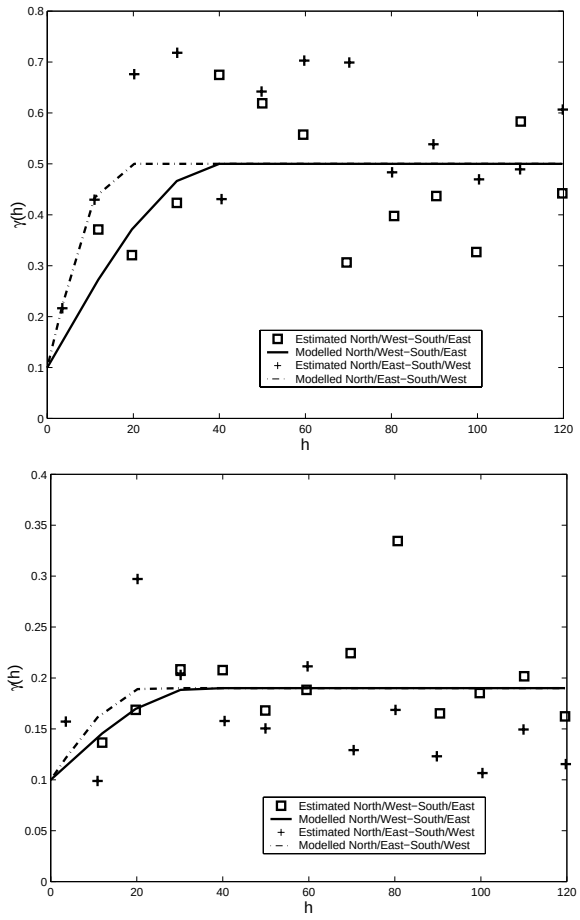

Figure 3.9: *Directional semivariograms in the directions of anisotropy, for a week in the middle of January 1994. Data are temporally reconstructed by GLM. Upper: Log-transformed DIN. Lower: Log-transformed DIP.*

Week	Variable	Range
1776	log(DIN)	2.35
1776	log(DIP)	1.55
1837	log(DIN)	4.23
1837	log(DIP)	3.04
1854	log(DIN)	4.20
1854	log(DIP)	1.35
1919	log(DIN)	2.19
1919	log(DIP)	1.00

Table 3.4: *Anisotropy ratios given as the range in the 135 degree direction divided by the range in the 45 degree direction.*

type of model, and the estimated values of the parameters, when data are temporally reconstructed by GLM. The models are very similar for the different weeks, but a very high range of the cross semivariogram between log-transformed DIP and depth of water is found for the spring week. The cross semivariograms have been computed for different directions, and it has been found that it can be assumed to be isotropic. The cross semi-

Week	Variables	Model	Range	Sill	Nugget effect
1776	Depth/log(DIN)	Spherical	40	-8.0	0.0
1776	Depth/log(DIP)	Spherical	42	-2.1	-0.1
1837	Depth/log(DIN)	Spherical	52	-8.0	0.0
1837	Depth/log(DIP)	Spherical	148	-1.3	0.0
1854	Depth/log(DIN)	Spherical	36	-7.5	0.0
1854	Depth/log(DIP)	Spherical	48	-2.8	0.0
1919	Depth/log(DIN)	Spherical	54	-12.0	-2.0
1919	Depth/log(DIP)	Spherical	38	-2.8	0.0

Table 3.5: *Type of cross semivariogram model and the parameters range, sill and nugget effect of omnidirectional cross semivariograms. Data are temporally reconstructed with GLM.*

variograms for the winter week are shown in figure 3.10. It has also been tried to estimate the corresponding cross semivariograms for data temporally reconstructed by LOESS. Due to lack of observations, i.e. only 24 observations are available for each week, these cross semivariograms did not make sense. For some lags negative values of the cross semivariance were found, while positive values were found for other lags. Finally the

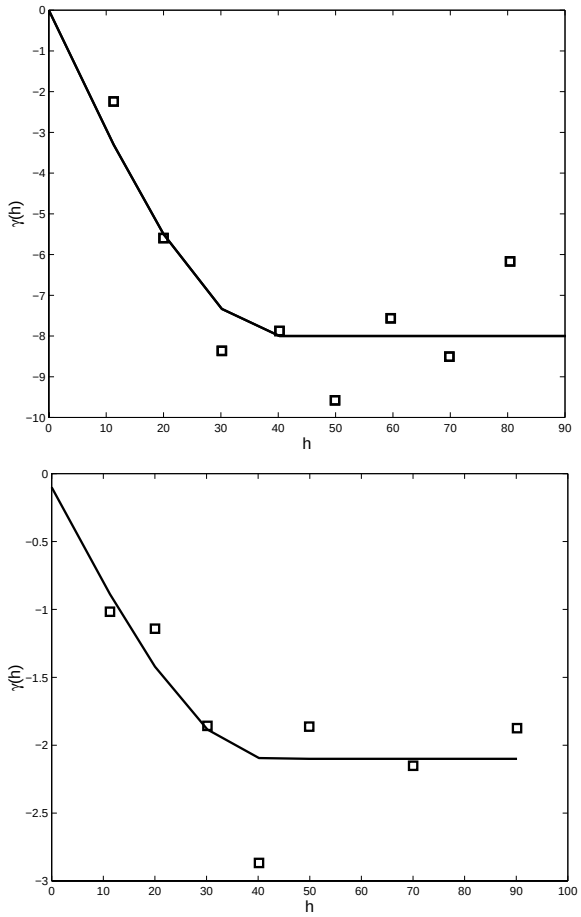

Figure 3.10: *Omnidirectional cross semivariograms, based on values which are measured or reconstructed by GLM, for a week in the middle of January 1994. Upper: Depth/log(DIN). Lower: Depth/log(DIP).*

semivariogram of the secondary varible, i.e. depth of water, has also been estimated and modelled. This has been done using a spherical semivariogram model with the parameters; range=80, C_0+C_1=568 and nugget effect=8. The semivariogram has been estimated in different directions, and based on this analysis, the area is assumed to be anisotropic. The omnidirectional semivariogram is shown in figure 3.11.

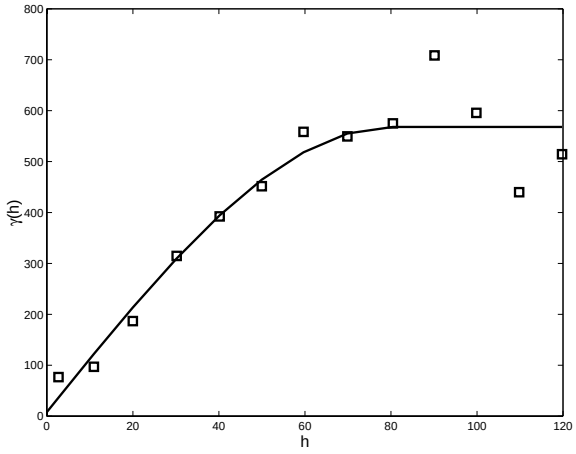

Figure 3.11: *Omnidirectional semivariogram for depth of water, which is used as a secondary variable, when computing spatial predictions using cokriging.*

3.3 Ordinary kriging

In the following it will be shown how the weights in equation (3.1) are calculated for ordinary kriging. In ordinary kriging the mean is unknown, and assumed to be constant for all the observations which are used in the estimation of $\hat{Z}(s_0)$, i.e. the model is

$$Z(s) = \mu + \epsilon(s) \tag{3.25}$$

We recall from equation (3.1) that the estimation in a point s_0 is calculated as

$$\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i) \tag{3.26}$$

The true value is denoted by $Z(s_0)$, and we have

$$\begin{aligned} E(\hat{Z}(s_0) - Z(s_0)) &= 0 \Longrightarrow \\ E(\hat{Z}(s_0)) &= E(Z(s_0)) \end{aligned} \tag{3.27}$$

The goal is to minimize the error of the estimation, i.e.

$$\text{Minimize} \quad E[(\hat{Z}(s_0) - Z(s_0))^2] \tag{3.28}$$

To obtain an unbiased estimate, the sum of the weights has to be 1, i.e.

$$\sum_i \lambda_i = 1 \tag{3.29}$$

The minimization given by (3.28) with the constraint (3.29), is calculated by introducing a Lagrange parameter m. This converts the constrained minimization problem into an unconstrained, and gives the following system of linear equations

$$\begin{aligned} \sum_j \lambda_j \gamma(s_i - s_j) + m &= \gamma(s_i - s_0) \\ \sum_i \lambda_i &= 1 \end{aligned} \tag{3.30}$$

Equation (3.30) can be written in matrix form

$$\begin{bmatrix} 0 & \gamma_{12} & \dots & \gamma_{1n} & 1 \\ \gamma_{21} & 0 & \dots & \gamma_{2n} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_{n1} & \gamma_{n2} & \dots & 0 & 1 \\ 1 & 1 & \dots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ m \end{bmatrix} = \begin{bmatrix} \gamma_{10} \\ \gamma_{20} \\ \vdots \\ \gamma_{n0} \\ 1 \end{bmatrix}$$

Here $\gamma_(s_i - s_j)$ is denoted γ_{ij}. The diagonals in the matrix is 0 since $\gamma_{ii} = 0$.

The ordinary kriging variance is given by

$$\sigma^2_{ok} = \sum_i \lambda_i \gamma(s_i - s_0) + m \tag{3.31}$$

In appendix D it is shown how the calculation from equation (3.28) to (3.30) is done, and the reader is therefore referred to this for more details. See also [Cressie, 1993], [Deutsch and Journel, 1992] and [Marsily, 1986].

3.3.1 Results

In the following the results when applying ordinary kriging to DIN and DIP will be shown for a week in the winter- and summerperiod (week 1776 and 1854). The results for a week in spring and autumn are shown in appendix E. Data which are used as the basis for predicting the spatial distribution of the variables, are constructed by GLM or LOESS, i.e. both reconstruction methods are used for temporal reconstruction. This means that we get two maps of the spatial distribution for each week, one based on temporal reconstruction using LOESS, and another based on data constructed by GLM. It will also be shown how the models can be evaluated and compared. For weeks where measurements of the variables have been carried out, these are used instead of the reconstructed values.[1]

Dissolved Inorganic Nitrogen

Spatial predictions of the concentration of DIN are computed for the log-transformed data, and the predictions are back transformed using equation (3.32).

$$\widehat{DIN} = \exp(\widehat{\log(DIN)} + V(\widehat{\log(DIN)})/2) \tag{3.32}$$

Figure 3.12 shows the spatial distribution of DIN in the winter using GLM for temporal reconstruction. High concentrations are calculated along the eastern coast of Jutland, especially in Hevring Bay and in the north/western part of Kattegat, as well as in the north/eastern part of Kattegat near the Swedish coast.

[1] The two sets of data will be referred to as temporally reconstructed by GLM or LOESS, even though the reconstructed values are only used for weeks where no measurements have been carried out

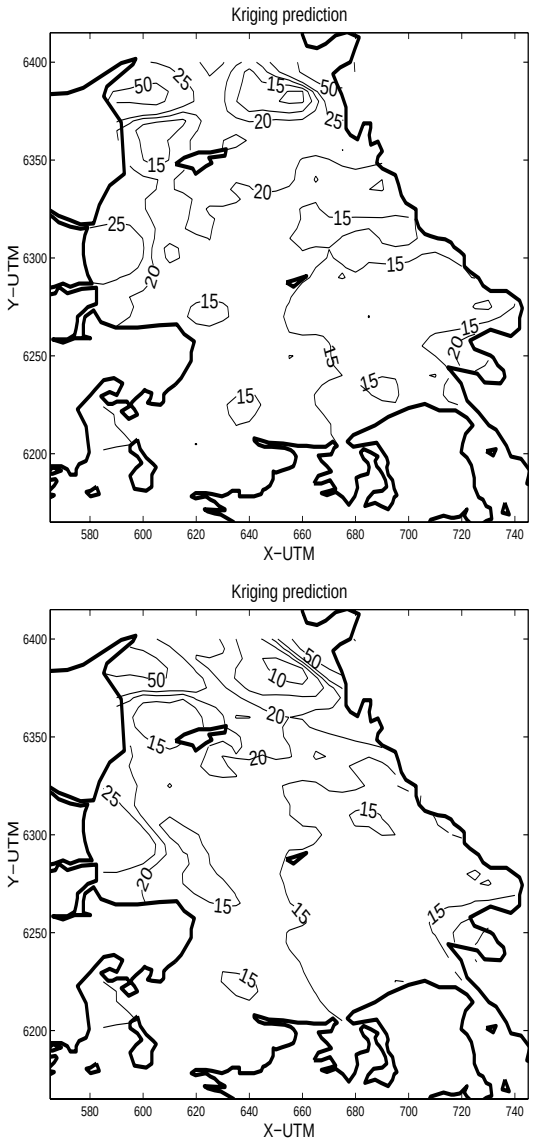

Figure 3.12: *Spatial distribution of DIN for a week in the middle of January 1994 computed by ordinary kriging. Temporal reconstruction is computed by GLM. Upper: Kattegat is assumed to be isotropic. Lower: Kattegat is assumed to be anisotropic.*

The high concentrations of DIN in Hevring Bay can be caused by upwelling. This phenomenon is known to take place when low-saline water from the Baltic Sea enters Kattegat through Øresund. The low-saline water flows on top of the lower layer of water with a higher salinity, and therefore increases the pressure on this high-saline water. The pressure generates a vertical movement of water, in the eastern direction, at the bottom of the sea, and causes an upwelling of high-saline water in Hevring Bay.

Another reason for the concentrations in Hevring Bay can be that water from Gudenåen is discharged into Kattegat through Randers Fjord. The water has a high content of nutrients, because Gudenåen drains large areas of agriculture in Jutland.

The high concentrations in the north/western part of Kattegat is caused by the Jutland Current, described in section 1.4.1, which occasionally flows into Kattegat. This current transports water with high concentrations of nutrients from the central European rivers, towards the north along the western coast of Jutland. In general the atmospheric deposition of nitrogen species on Kattegat is larger in the western part of Kattegat, than by the Swedish, which also causes the concentration of DIN to be higher by the eastern coast of Jutland. The high concentrations of DIN in the north/eastern part of Kattegat is due to discharge of freshwater from Göta Elven.

The upper map in figure 3.12 is computed for the case where Kattegat is assumed to be isotropic with respect to DIN. It has been found in section 3.2.5 that this is not the case. The north/western-south/eastern direction has the largest range, while the perpendicular direction has the smallest. When we correct for this anisotropy, the lower map in figure 3.12 is obtained. It is seen that the contours are tilted towards the direction of the higher range, meaning that observations located in this direction relative to each other are more likely to be the same. Figure 3.13 shows one of the very useful properties of applying kriging for determination of the spatial distribution, i.e. we are able to calculate the uncertainty, in this case given by the standard deviation, of the predictions. The back transformation of the variances are calculated using equation (3.33).

$$S(\widehat{DIN}) = \sqrt{\exp[2 \cdot \log(\widehat{DIN}) + V(\log(\widehat{DIN}))](\exp[V(\log(\widehat{DIN}))] - 1)} \tag{3.33}$$

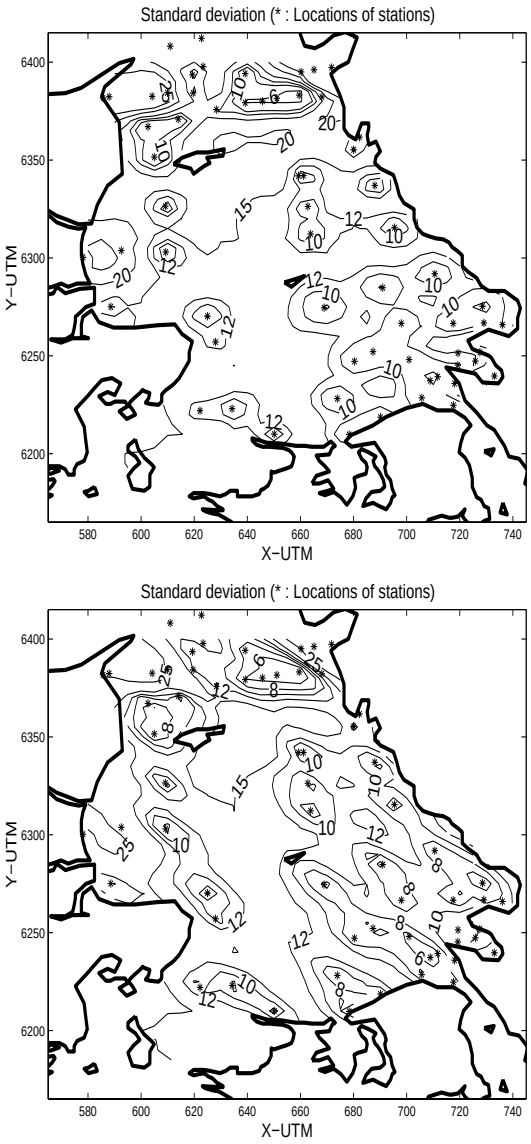

Figure 3.13: *Standard deviation of predictions of DIN for a week in the middle of January 1994 computed by ordinary kriging. Temporal reconstruction is computed by GLM. Upper: Kattegat is assumed to be isotropic. Lower: Kattegat is assumed to be anisotropic.*

The two maps in figure 3.13 correspond to the maps of concentrations of DIN shown in figure 3.12. The standard deviations are low near the location of observations. Again the contours are tilted, when the anisotropy of DIN is taken into account, and, as expected, the standard deviations are lower in the anisotropy case. Consequently, the model which incorporates anisotropy seems to be better than the one, where the area is considered as isotropic.

The upper map in figure 3.14 shows the spatial distribution of DIN, in the case where the temporal reconstruction is computed using LOESS. The predicted concentrations are lower than what is found for data constructed by GLM. Moreover high concentrations are found in the open sea, which do not seem to be reasonable. This is also found for the concentration of DIN for a week in the middle of March 1995. This map is shown in figure E.1 in appendix E. High concentrations are also found in the innermost part of Hevring Bay, indicating that the area of Kattegat, which is influenced by the discharge from Gudenåen, is not as large as it seems to be when GLM is used for temporal reconstruction.

The lower map in figure 3.14 shows the corresponding standard deviations. The values seem to be lower than the standard deviations computed when GLM is used for temporal reconstruction of data. This might be due to the larger range estimated when LOESS is used.

Figure 3.15 shows the spatial distribution of DIN for a week in the middle of July 1995. The prediction, which is based on data temporally reconstructed by GLM is shown on the upper map, while the prediction based on data generated by LOESS is on the lower map. The pattern of the spatial distribution of DIN is found to be the same in both cases, with the highest concentrations along the eastern coast of Jutland, and in the north/eastern part of Kattegat. However the concentrations on the lower map are higher,especially in the coastal areas. The computation of the spatial distribution of concentrations of DIN based on LOESS does only contain 24 observations. This means that extreme observations are able to dominate the prediction, and cause results which do not seem to be reasonable. This might also be the reason for the computation of high concentrations of DIN in the open sea, shown in figure 3.14.

The spatial distribution of the concentration of DIN for a week in the middle of October 1996 is shown in figure E.2 in appendix E. Here the

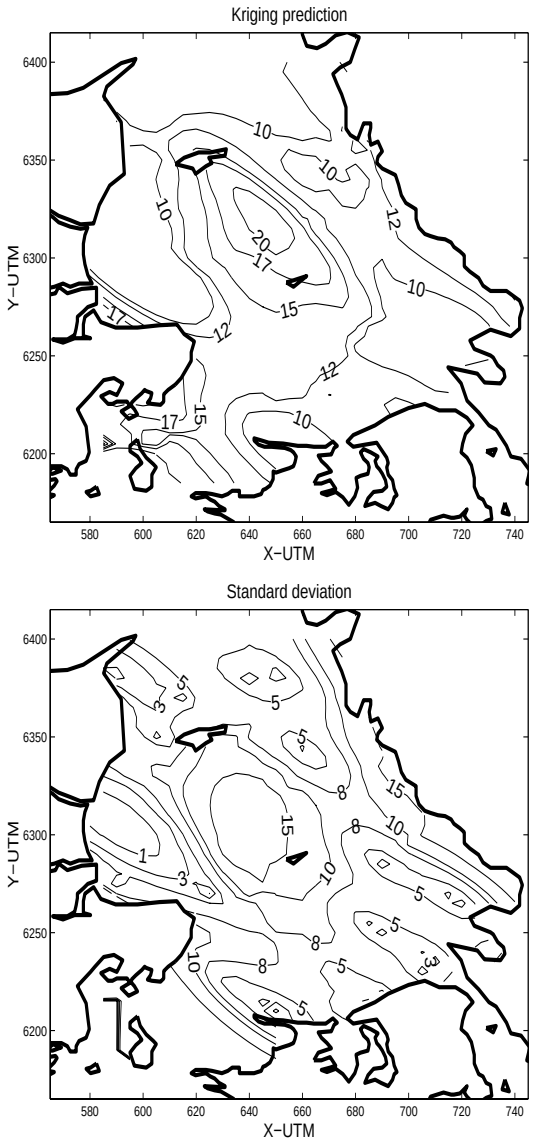

Figure 3.14: *Spatial distribution of the concentration of DIN, and the standard deviation of the predictions, computed by ordinary kriging, for a week in the middle of January 1994. Temporal reconstruction is computed by LOESS. Upper: Spatial distribution of the concentration of DIN. Lower: Standard deviation of the predictions.*

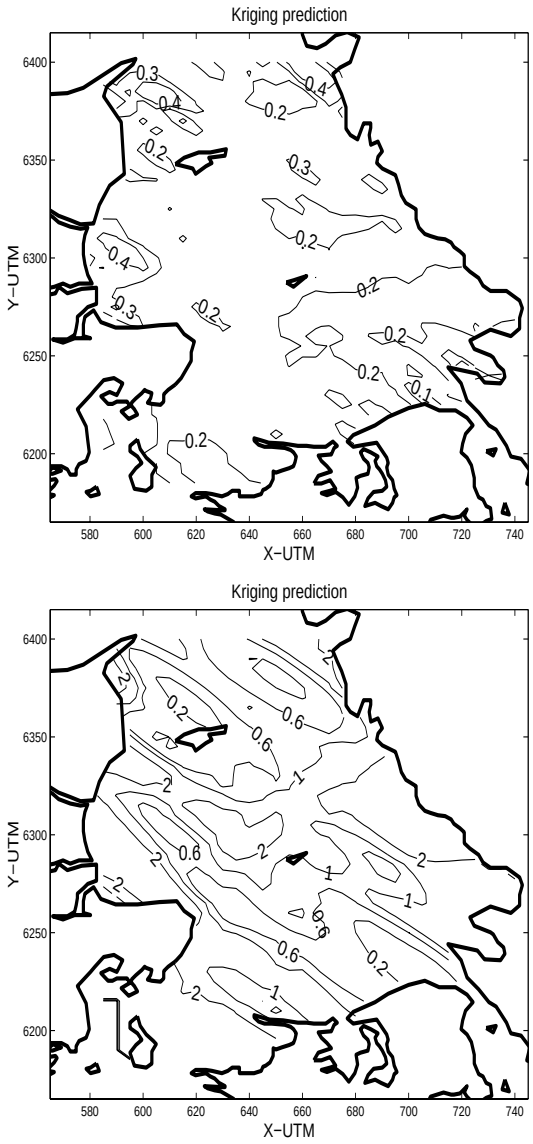

Figure 3.15: *Spatial distribution of DIN for a week in the middle of July 1995 computed by ordinary kriging. Upper: Temporal reconstruction by GLM. Lower: Temporal reconstruction by LOESS.*

same pattern of the spatial distribution is found as in figure 3.15, but the estimated concentrations are of the same order of magnitude when GLM and LOESS are used for temporal reconstruction.

Reconstruction	Week	$\sum(DIN - \widehat{DIN})^2$	m	Goodness Of Model
GLM	1776	2189	16	136.8
GLM	1837	1424	19	74.95
GLM	1854	2.50	19	0.136
GLM	1919	14.8	17	0.871
Total		3630	71	51.13
LOESS	1776	296.1	16	18.51
LOESS	1837	191.6	17	11.27
LOESS	1854	136.2	18	7.567
LOESS	1919	35.28	17	2.075
Total		659.2	68	9.694

Table 3.6: *Cross validation of ordinary kriging model for DIN.*

The kriging models, for the two different cases of temporal reconstruction of data, can be evaluated and compared for each of the four different weeks using cross validation. This is done by leaving a single value of log(DIN) out of the model each time. Afterwards the predicted value at the same location is assigned to the left out observation, and back transformed using equation (3.32). The Goodness Of Model is calculated as

$$\text{Goodness Of Model} = \frac{1}{m}\sum_{k=1}^{m}(DIN_k - \widehat{DIN_k})^2 \tag{3.34}$$

where m is the number of observations and DIN are the original, measured concentrations of DIN. This means that locations represented by a reconstructed value are not included in the goodness of model. The cross validation shows that the computation of the spatial distribution of DIN is better when GLM is used for temporal reconstruction for the summer -and autumn weeks, while the computation for winter and spring seems to be better when LOESS is used. The values of the Goodness of Model cannot be directly compared to those found in section 1.5.1 and 2.2.1.

Dissolved Inorganic Phosphorus

The spatial predictions of the concentration of DIP are, in the same way as for DIN, computed for the log-transformed data, and the predictions are back transformed.

Figure 3.16 shows the spatial distribution of DIP in the winter using GLM and LOESS for temporal reconstruction. The upper map shows that high concentrations are calculated along the eastern coast of Jutland and in the north/eastern part of Kattegat near the Swedish coast. This spatial pattern can be explained by the same phenomena as described for DIN, although phosphorus is not supplied from the atmosphere.

The lower map shows relatively high concentrations of DIP in the open sea, when data for the spatial prediction are temporally reconstructed by LOESS. This is also found for DIN in figure 3.14 and does not seem to be reasonable. The pattern of the spatial distribution of DIP shown in figure 3.16, is also what is found for a week in the middle of March 1995, and shown in figure E.3 in appendix E.

The standard deviations of the spatial predictions in figure 3.16 are shown in figure 3.17. These values are lower when the temporal reconstruction is computed by LOESS, because of the higher range used in this case.

The general pattern of the spatial distribution of DIP in summertime, shown in figure 3.18, is seen to be very similar for both kinds of temporal reconstruction, although the high concentrations, in the south/eastern part of Kattegat near the Swedish coast, are not found when GLM is used for temporal reconstruction. Also the level of concentrations is within the same order of magnitude.

The kriging models, for the two different cases of temporal reconstruction of data, have been evaluated and compared for each of the four different weeks, in the same way as it was done for DIN. The cross validation shows that the computation of the spatial distribution of DIP is better when LOESS is used for temporal reconstruction, compared to GLM for all the four weeks.

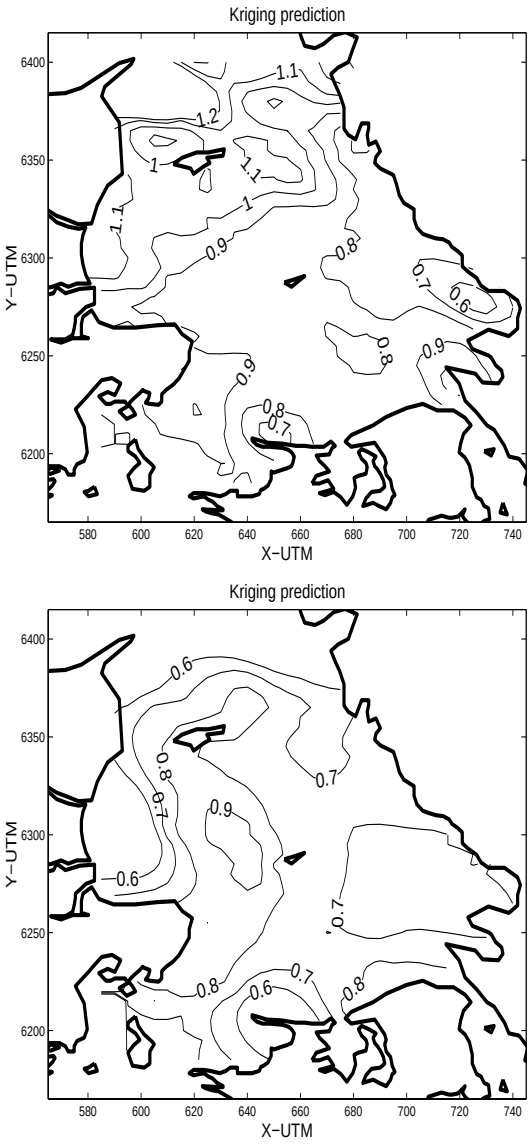

Figure 3.16: *Spatial distribution of DIP for a week in the middle of January 1994 computed by ordinary kriging. Upper: Temporal reconstruction by GLM. Lower: Temporal reconstruction by LOESS.*

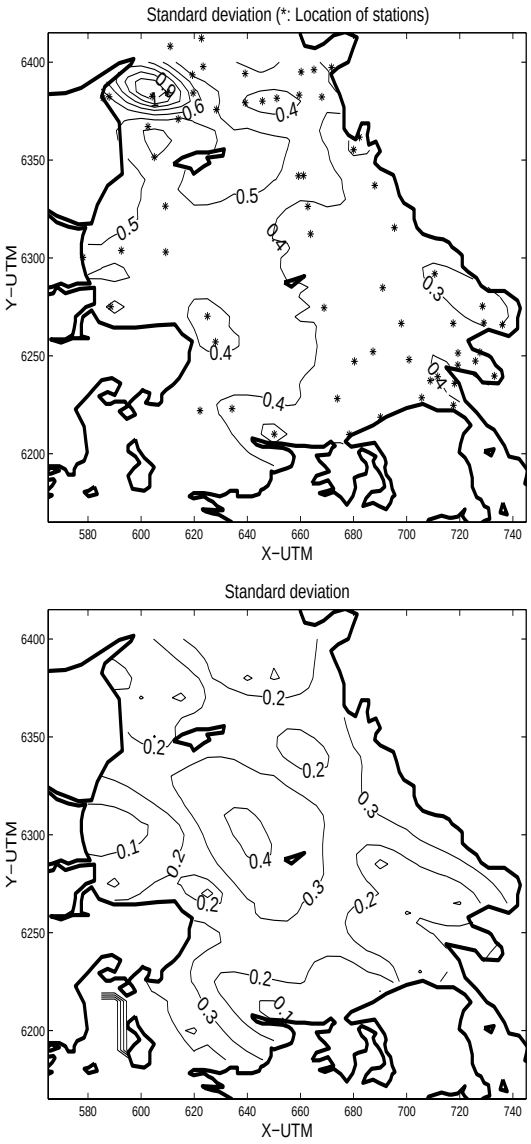

Figure 3.17: *Standard deviation of predictions of DIP for a week in the middle of January 1994 computed by ordinary kriging. Upper: Temporal reconstruction by GLM. Lower: Temporal reconstruction by LOESS.*

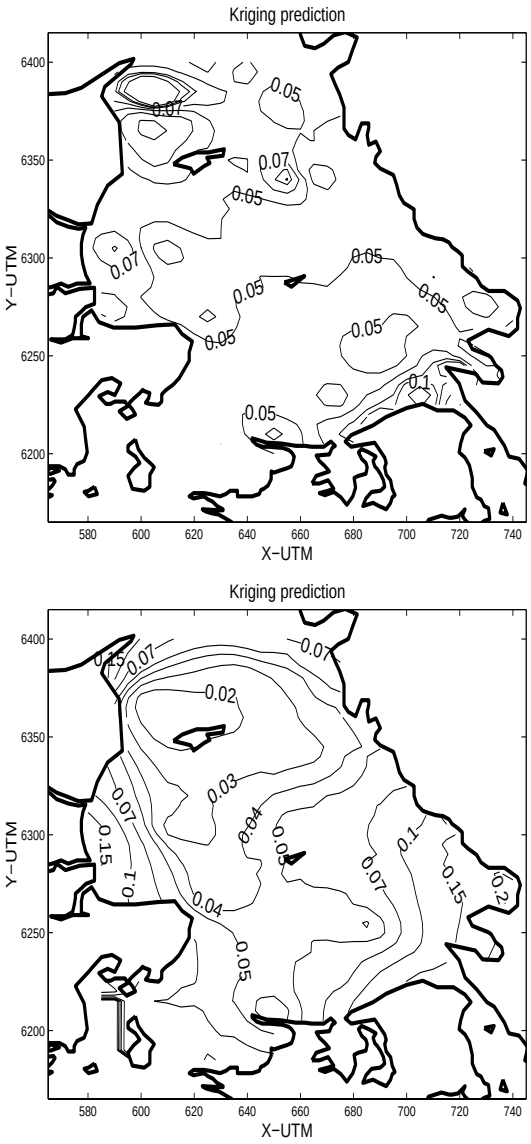

Figure 3.18: *Spatial distribution of DIP for a week in the middle of July 1995 computed by ordinary kriging. Upper: Temporal reconstruction by GLM. Lower: Temporal reconstruction by LOESS.*

Reconstruction	Week	$(DIP - \widehat{DIP})^2$	m	Goodness Of Model
GLM	1776	2.440	16	0.153
GLM	1837	1.355	18	0.075
GLM	1854	0.042	19	0.0022
GLM	1919	0.755	17	0.044
Total		4.592	70	0.066
LOESS	1776	0.322	16	0.0201
LOESS	1837	0.514	17	0.0302
LOESS	1854	0.035	19	0.0018
LOESS	1919	0.541	17	0.0318
Total		1.412	69	0.0205

Table 3.7: *Cross validation of ordinary kriging model for DIP.*

3.4 Universal kriging

The term universal kriging is used when the mean field can be described by a trend model. Spatial predictions are modelled as

$$Z(s) = \mu(s) + \epsilon(s) \tag{3.35}$$

where s is the location, given by x and y coordinates. This means that ordinary kriging is also universal kriging, with a "trend" model described by the constant mean. The trend component $\mu(s)$ can be modelled as a function of coordinates, x and y, in the 2-dimensional case, and x, y and z in the 3-dimensional case. This is shown in (3.36).

$$\mu(s) = \sum_{l=0}^{L} a_l f_l(s), \quad f_0(s) = 1 \tag{3.36}$$

The functions $f_l(s)$ are known, while the parameters a_l are unknown, but can be estimated from data. Usually low-order polynomias are used for describing $\mu(s)$, such as the first order polynomia

$$\mu(s) = \mu(x, y) = a_0 + a_1 x + a_2 y + a_3 xy \tag{3.37}$$

The computation of Z at an unmeasured location s_0 is similar to the estimation in the case of ordinary kriging.

$$\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i) \tag{3.38}$$

In this case we have

$$E(\hat{Z}(s_0)) = \sum_{i=1}^{n}\sum_{l=0}^{L} \lambda_i Z a_l f_l(s_i) \tag{3.39}$$

The goal is to minimize the error of the estimation

$$\text{Minimize} \quad E[(\hat{Z}(s_0) - Z(s_0))^2] \tag{3.40}$$

In order to obtain an unbiased estimate the following constraint is used

$$\sum_{i=1}^{n} \lambda_i f_l(s_i) = f_l(s_0), \quad l = 0, \cdots, L \tag{3.41}$$

By introducing L Lagrange parameters, m_l, we get the following system of linear equations

$$\begin{aligned} \sum_{i=1}^{n} \lambda_i \gamma(s_i - s_j) + \sum_{l=0}^{L} m_l f_l(s_j) &= \gamma(s_j - s_i), \quad j = 1, \cdots, n \\ \sum_{i=1}^{n} \lambda_i f_l(s_i) &= f_l(s_0), \quad l = 0, \cdots, L \end{aligned} \tag{3.42}$$

expressed in terms of the semivariogram, [Journel and Rossi, 1989] and [Clausen, 1980]. Written in matrix form we get

$$\begin{bmatrix} 0 & \gamma_{21} & \cdots & \gamma_{n1} & f_0(s_1) & \cdots & f_L(s_1) \\ \gamma_{12} & 0 & \cdots & \gamma_{n2} & f_0(s_2) & \cdots & f_L(s_2) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \gamma_{1n} & \gamma_{2n} & \cdots & 0 & f_0(s_n) & \cdots & f_L(s_n) \\ f_0(s_1) & f_0(s_2) & \cdots & f_0(s_n) & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ f_L(s_1) & f_L(s_2) & \cdots & f_L(s_n) & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ m_0 \\ \vdots \\ m_L \end{bmatrix} = \begin{bmatrix} \gamma_{10} \\ \gamma_{20} \\ \vdots \\ \gamma_{n0} \\ f_0(s_0) \\ \vdots \\ f_L(s_0) \end{bmatrix}$$

The universal kriging variance is given by

$$\sigma_{uk}^2 = \sum_{i=1}^{n} \lambda_i \gamma(s_i - s_0) + \sum_{l=0}^{L} \lambda_l f_l(s_j), \quad j = 1, \cdots, n \tag{3.43}$$

The use of universal kriging seems reasonable compared to ordinary kriging where the mean is assumed to be constant. But universal kriging is associated with some theoretical problems, which has lead to only a minor use of

the method. The problem is that the trend model and the semivariogram have to be estimated at the same time. To estimate the semivariogram of the residuals defined by

$$\hat{\gamma}(h) = \frac{1}{2}E[(z(s+h) - \mu(s+h)) - (z(s) - \mu(s))]^2 \tag{3.44}$$

the trend model $\mu(s)$ is needed. To compute the trend model we need to know the system of universal kriging equations, given by (3.42). Moreover the experimental semivariogram of residuals underestimates the true semivariogram. If the semivariogram of residuals is estimated anyway, it is still not straightforward to compute the trend or the true underlying semivariogram. However universal kriging can still be used, if it is assumed that the underlying semivariogram is known, [Armstrong, 1984].

3.4.1 Results

In the following the results when applying universal kriging to DIN and DIP will be shown for a week in the winter- and summerperiod (week 1776 and 1854), in the same way as it was done in the case of ordinary kriging. The results for a week in spring and autumn are shown in appendix F. Data which are used as the basis for predicting the spatial distribution of the variable, are constructed by GLM or LOESS, i.e. both reconstruction methods are used for temporal reconstruction.

Dissolved Inorganic Nitrogen

Figure 3.19 shows the estimation of the trend, $\mu(s)$, of DIN in the winter week, based on data temporally reconstructed by GLM. The trend represents the deterministic part of the universal kriging model, and is computed as

$$\mu(s) = \mu(x, y) = a_0 + a_1 x + a_2 y + a_3 x^2 + a_4 y^2 + a_5 xy \tag{3.45}$$

where the values of $a_0, \cdots, a_5$ are found from data. The map of the trend shows high concentrations of DIN in the coastal areas. The concentration decreases with a nearly constant gradient from the coastal area to the middle of the area of Kattegat. When adding the stochastic part, $\epsilon(s)$, of the kriging model to the trend, the upper map in figure 3.20 is obtained, while the lower map shows the spatial distribution of DIN, in the case where the

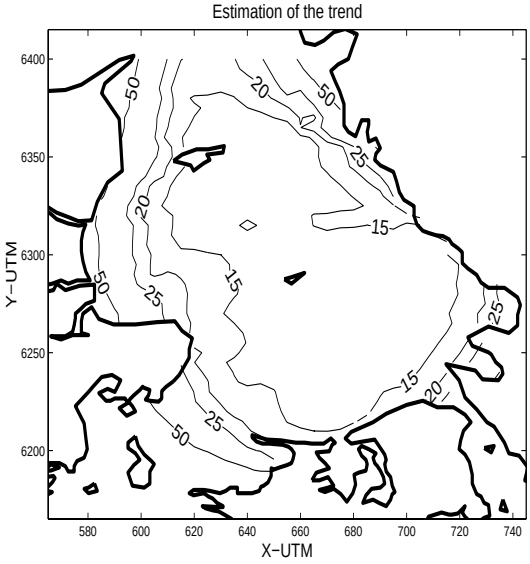

Figure 3.19: *Computation of the trend of DIN in the universal kriging model for a week in the middle of January 1994. Data are temporally reconstructed by GLM.*

predictions are based on data temporally reconstructed by LOESS.

Like at the corresponding map computed by ordinary kriging, this also shows high concentrations of DIN in the open sea. However when using universal kriging, the highest concentrations found in the open sea are not higher than in the coastal areas, and the map computed by universal kriging is therefore more understandable than the corresponding map computed by ordinary kriging.

The standard deviations of the predictions in figure 3.20 are shown in figure 3.21. The values of the standard deviation are not as high when LOESS is used for temporal reconstruction compared to GLM. This is due to the longer range used in the computation, when data are temporally reconstructed by LOESS.

The spatial distribution of DIN in summertime is shown in figure 3.22. The same spatial pattern is seen on both maps, but the concentrations in the coastal areas are extremely overestimated, when data are temporally reconstructed by LOESS. This is due to the extrapolation, which is caused by the inclusion of a trend in the kriging model. This extrapolation becomes extreme in the case where data are based on LOESS, because of the small number of observations. The same is seen in figure F.2 in appendix F, which shows the spatial distribution of DIN for a week in the middle of October 1996.

The kriging models, for the two different cases of temporal reconstruction of data, have been evaluated and compared for each of the four different weeks, in the same way as described for DIN in section 3.3.1. The cross validation shows that the computation of the spatial distribution of DIN is better when GLM is used for temporal reconstruction for the summer -and autumn weeks, while the computation for winter and spring is better when LOESS is used. The values of the Goodness Of Model when using universal kriging, can also be compared to those calculated when ordinary kriging is used for determination of the spatial distribution of DIN. These numbers are shown in table 3.6. It is seen that the model is not improved by inclusion of a trend in the kriging model. Actually only the computation of the spatial distribution of DIN in wintertime, based on data temporally reconstructed by GLM, is improved.

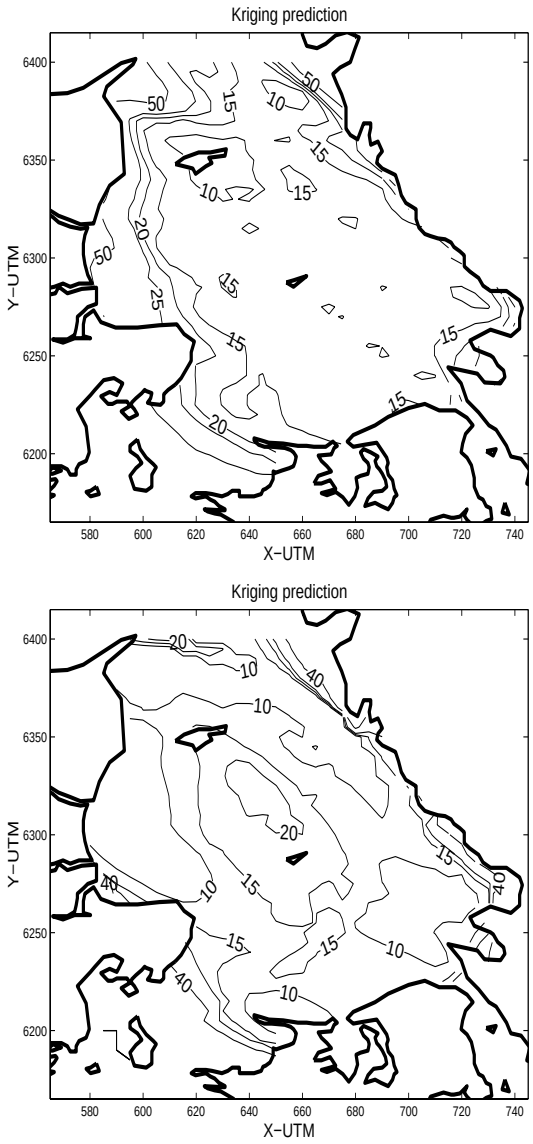

Figure 3.20: *Spatial distribution of DIN for a week in the middle of January 1994 computed by universal kriging. Upper: Temporal reconstruction by GLM. Lower: Temporal reconstruction by LOESS.*

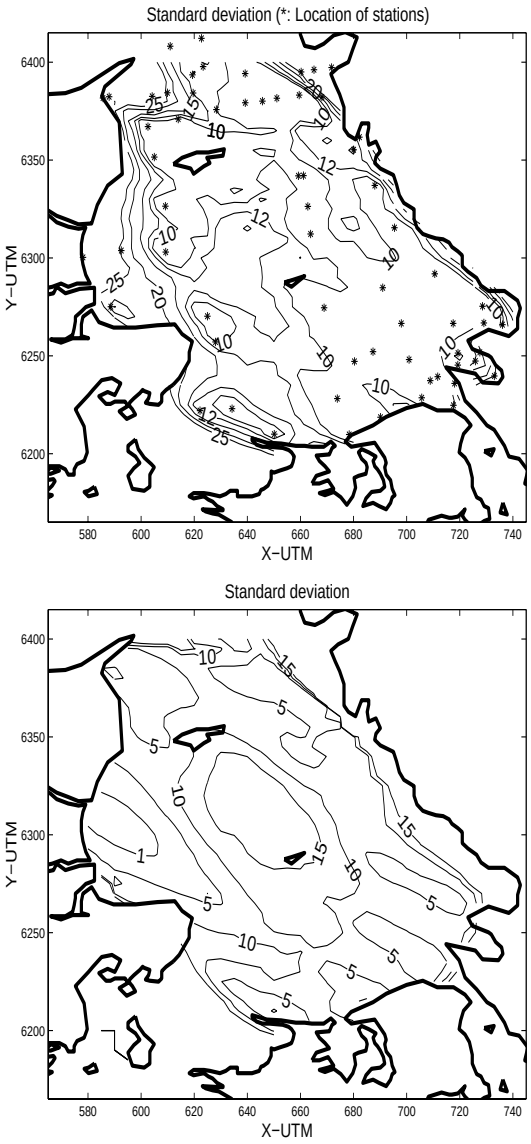

Figure 3.21: *Standard deviation of predictions of DIN for a week in the middle of January 1994 computed by universal kriging. Upper: Temporal reconstruction by GLM. Lower: Temporal reconstruction by LOESS.*

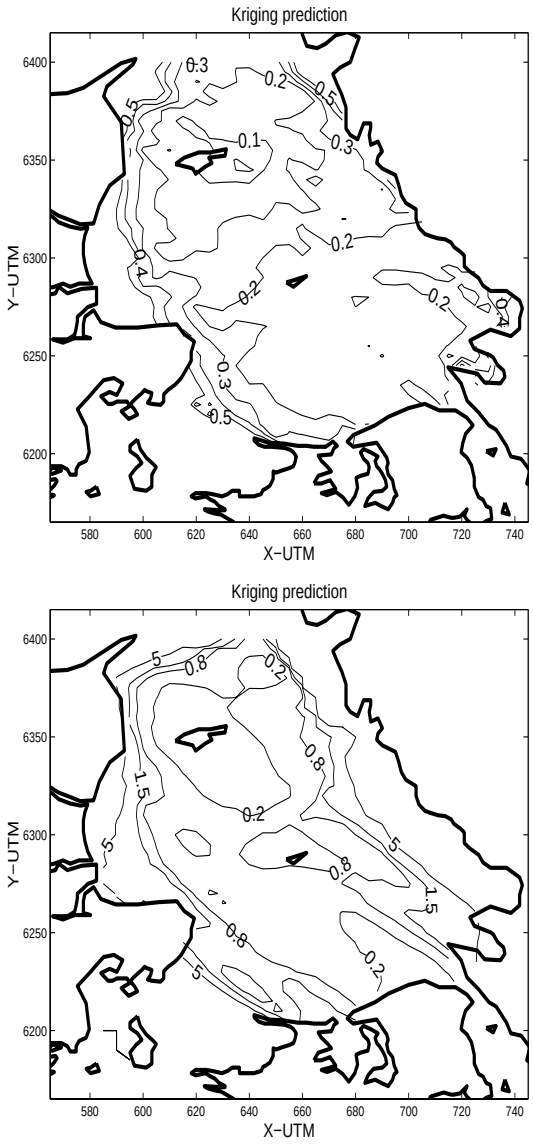

Figure 3.22: *Spatial distribution of DIN for a week in the middle of July 1995 computed by universal kriging. Upper: Temporal reconstruction by GLM. Lower: Temporal reconstruction by LOESS.*

Reconstruction	Week	$\sum(DIN - \widehat{DIN})^2$	m	Goodness Of Model
GLM	1776	1426	16	89.13
GLM	1837	3688	19	194.1
GLM	1854	2.915	19	0.153
GLM	1919	60.54	17	3.561
Total		5177	71	72.92
LOESS	1776	346.9	16	21.68
LOESS	1837	278.8	17	16.40
LOESS	1854	240.7	17	14.16
LOESS	1919	179.9	17	10.58
Total		1046	67	15.61

Table 3.8: *Cross validation of universal kriging model for DIN.*

Dissolved Inorganic Phosphorus

Figure 3.23 shows the estimation of the trend, $\mu(s)$, of DIP in the winter week, based on data temporally reconstructed by GLM. The trend is computed in the same way as in the case of DIN. As for DIN, the highest values of the trend are found in the coastal areas, but for DIP the gradient of the trend is in the north/western-south/eastern direction, with the highest concentrations in the north/western part of Kattegat.

When adding the stochastic part, $\epsilon(s)$, of the kriging model to the trend, the upper map in figure 3.24 is obtained, while the lower map shows the spatial distribution of DIP, in the case where the predictions are based on data temporally reconstructed by LOESS. The lower map shows the highest concentrations of DIP in the open sea, which is not rational from a physical point of view, but is probably due to the relatively small number of observations, in the case where predictions are based on data temporally reconstructed by LOESS. The upper map is very similar to the corresponding one, computed by ordinary kriging.

The standard deviations of the predictions in figure 3.24 are shown in figure 3.25. As in the case of prediction of DIN, the values of the standard deviation are not as high when LOESS is used for temporal reconstruction compared to GLM. This is due to the longer range used in the computation, when data are temporally reconstructed by LOESS. The spatial distribution of DIP in summertime is shown in figure 3.26. Almost the same spatial pattern is seen on both maps. Compared to the corresponding map com-

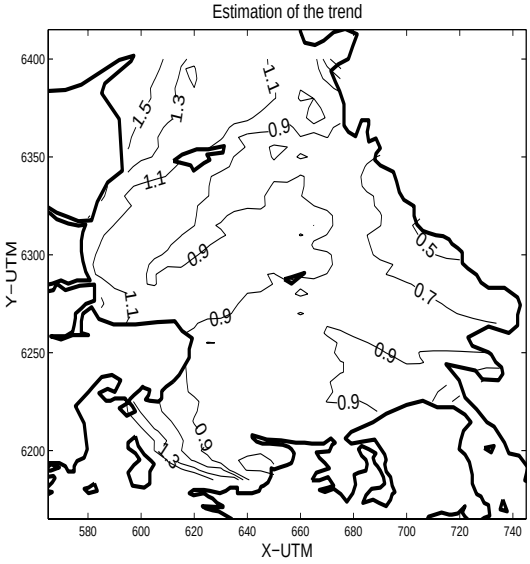

Figure 3.23: *Computation of the trend of DIP in the universal kriging model for a week in the middle of January 1994. Data are temporally reconstructed by GLM.*

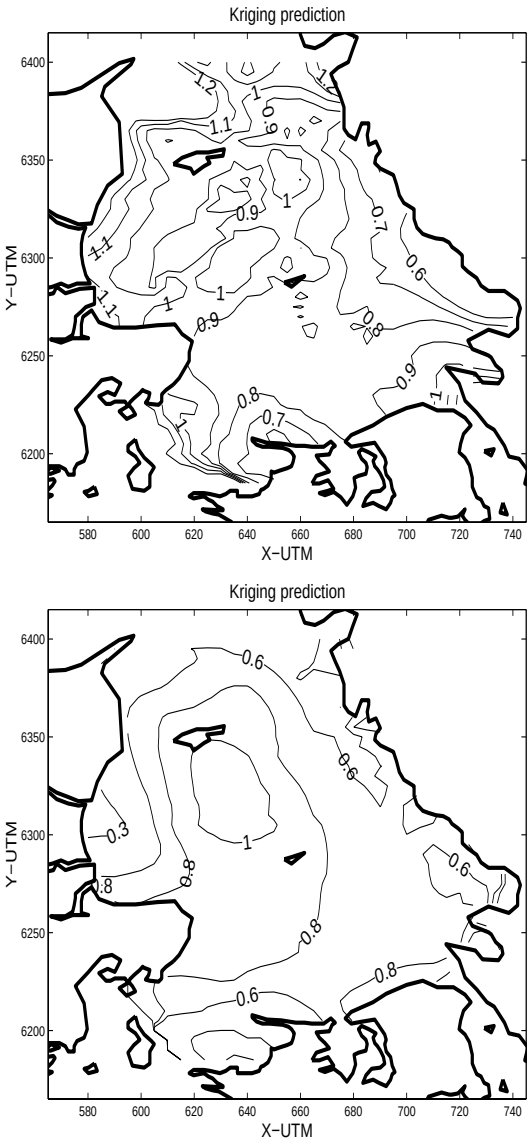

Figure 3.24: *Spatial distribution of DIP for a week in the middle of January 1994 computed by universal kriging. Upper: Temporal reconstruction by GLM. Lower: Temporal reconstruction by LOESS.*

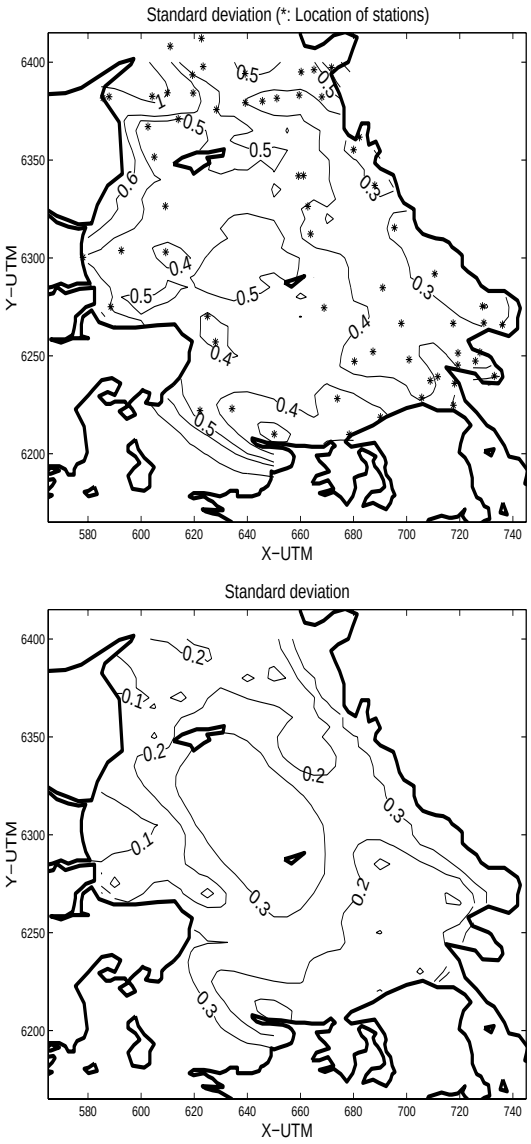

Figure 3.25: *Standard deviation of predictions of DIP for a week in the middle of January 1994 computed by universal kriging. Upper: Temporal reconstruction by GLM. Lower: Temporal reconstruction by LOESS.*

puted by ordinary kriging, higher concentrations of DIP in coastal areas are shown on the lower map in figure 3.26. This is due to the extrapolation caused by the inclusion of a trend in the kriging model.

The kriging models, for the two different cases of temporal reconstruction of data, have been evaluated and compared for each of the four different weeks, in the same way as described for DIN in section 3.3.1. The cross validation shows that the computation of the spatial distribution of DIN is better when GLM is used for temporal reconstruction for the summer -and autumn weeks, while the computation for winter and spring is better when LOESS is used. The values of the Goodness Of Model when using universal kriging, can also be compared to those calculated when ordinary kriging is used for determination of the spatial distribution of DIP. These numbers are shown in table 3.7. It is seen that the model is not improved by inclusion of a trend in the kriging model.

Reconstruction	Week	$(DIP - \widehat{DIP})^2$	m	Goodness Of Model
GLM	1776	3.045	16	0.190
GLM	1837	1.533	18	0.085
GLM	1854	0.034	19	0.0018
GLM	1919	1.456	17	0.086
Total		6.068	70	0.087
LOESS	1776	0.478	16	0.030
LOESS	1837	0.610	17	0.036
LOESS	1854	0.140	18	0.0077
LOESS	1919	2.409	17	0.142
Total		3.637	68	0.053

Table 3.9: *Cross validation of universal kriging model for DIP.*

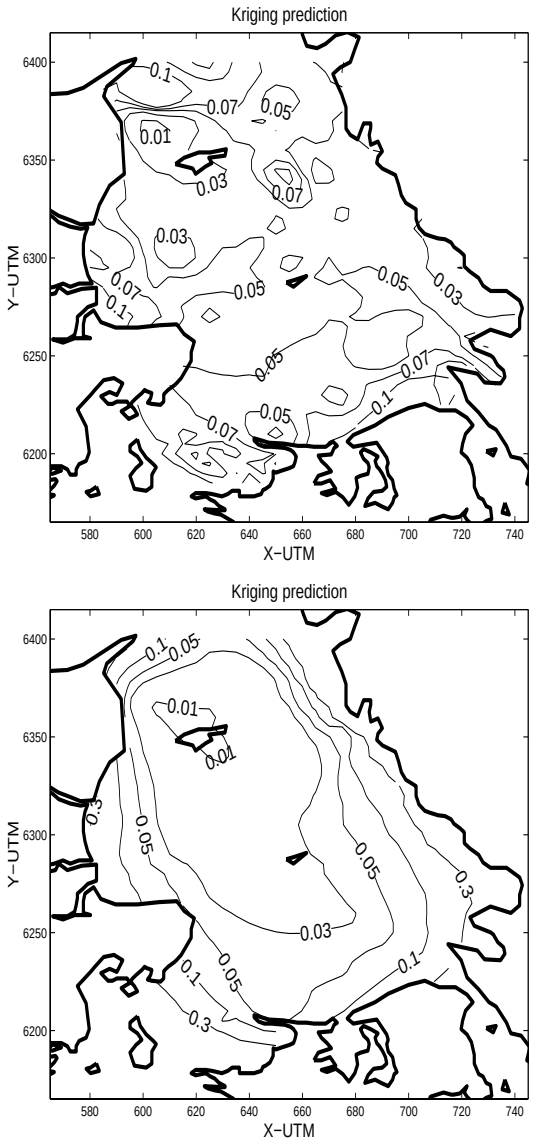

Figure 3.26: *Spatial distribution of DIP for a week in the middle of July 1995 computed by universal kriging. Upper: Temporal reconstruction by GLM. Lower: Temporal reconstruction by LOESS.*

3.5 Cokriging

Cokriging can be used when measurements of more than one variable, e.g. Z_1 and Z_2 have been done. The prediction of $\hat{Z}_1$ is done, not only on the basis of Z_1, but also on measurements of Z_2. In this section, the case where measurements of Z_1 and Z_2 are used to predict the primary variable $\hat{Z}_1$, will be explained. Z_2 is in this case the secondary variable. This can be generalized to prediction of $\hat{Z}_1$ from N variables. The reader is referred to [Cressie, 1993], [Myers, 1982] and [Marcotte, 1991] regarding this discussion. In the case of cokriging an additional index is needed to describe the variable. The prediction of the variable Z_1 at the location s_0 is given by

$$\hat{Z}_1(s_0) = \sum_{j=1}^{n} \lambda_1^j Z_1(s_j) + \sum_{l=1}^{m} \lambda_2^l Z_2(s_l) \tag{3.46}$$

j and l are indices of locations for observations of the two variables Z_1 and Z_2. Z_1 and Z_2 do not have to be measured at the same locations, and m and n are not neccesarily equal. Like in the case of ordinary and universal kriging, the function that we want to minimize is

$$\text{Minimize} \quad E[(\hat{Z}_1(s_0) - Z_1(s_0))^2] \tag{3.47}$$

To obtain an unbiased estimate the following constraints are needed

$$\sum_{j=1}^{n} \lambda_1^j = 1 \qquad \text{and} \qquad \sum_{l=1}^{m} \lambda_2^l = 0 \tag{3.48}$$

If both Z_1 and Z_2 are intrinsic, the cokriging equations become

$$\begin{aligned}
-\sum_{j=1}^{n} \lambda_1^j \gamma_{11}(s_i - s_j) - \sum_{l=1}^{m} \lambda_2^l \gamma_{12}(s_i - s_l) - m_1 &= -\gamma_{11}(s_0 - s_i) \\
\text{for} \quad i &= 1, \cdots, n \\
-\sum_{j=1}^{n} \lambda_1^j \gamma_{21}(s_k - s_j) - \sum_{l=1}^{m} \lambda_2^l \gamma_{22}(s_k - s_l) - m_2 &= -\gamma_{12}(s_0 - s_k) \\
\text{for} \quad k &= 1, \cdots, m
\end{aligned} \tag{3.49}$$

where m_1 and m_2 are Lagrange multipliers. The variance of the prediction is calculated as

$$\sigma^2_{cok} = \sum_{j=1}^{n} \lambda_1^j \gamma_{11}(s_0 - s_j) + \sum_{l=1}^{m} \lambda_2^l \gamma_{12}(s_0 - s_l) + m_1$$

3.5.1 Results

In the following the results when applying cokriging to DIN and DIP will be shown for a week in the winter- and summerperiod (weeks 1776 and 1854). The results for a week in spring and autumn are shown in appendix G. Data which are used as the basis for predicting the spatial distribution of the variable, are temporally reconstructed by GLM. It has been attempted to perform cokriging based on data temporally reconstructed by LOESS, but it did not result in a rational spatial distribution of neither DIN nor DIP. The depth of water at each station is used as the secondary variable. This means that we have 71 measures of the secondary variable and 65 of the primary.

Dissolved Inorganic Nitrogen

Figure 3.27 shows the spatial distribution, and corresponding map of standard deviations, of DIN in the winter. The same spatial pattern, as computed by ordinary or universal kriging, is seen. The high concentrations in the north-eastern and north-western part of Kattegat, are not as high as predicted by the other two variants of kriging, while the high concentrations by the eastern coast of Jutland, near Randers Fjord and Limfjorden, are higher when cokriging is used compared to ordinary kriging. This is due to the additional information included in the model, i.e. when the depth of water is lowered, the concentration of DIN should increase. The lower map in figure 3.27 shows that the computed standard deviations of the predictions are lower compared to those computed by the other variants of kriging. Again this is due to the additional information included in the model.

The spatial distribution of DIN in the summer is shown in figure 3.28. The same spatial pattern and order of magnitude of the concentration of DIN, as computed by ordinary or universal kriging when data are temporally reconstructed by GLM, is seen.

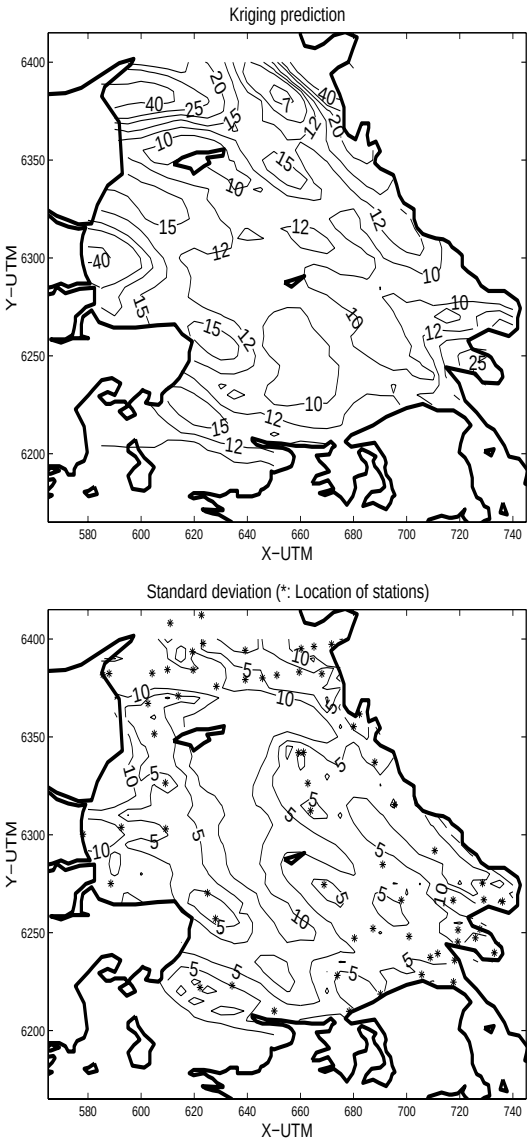

Figure 3.27: *Upper: Spatial distribution of DIN for a week in the middle of January 1994 computed by cokriging. Temporal reconstruction is computed by GLM. Lower: Corresponding map of the standard deviation.*

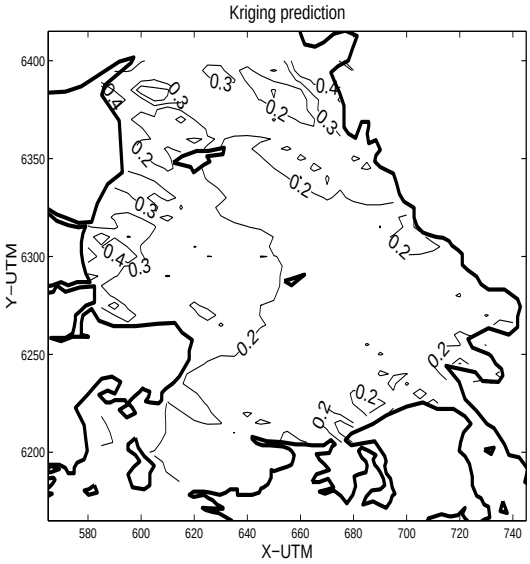

Figure 3.28: *Spatial distribution of DIN for a week in the middle of July 1995 computed by cokriging. Temporal reconstruction is computed by GLM.*

The cokriging model for DIN has been evaluated by cross validation in the same way as described in section 3.3.1. Only the primary variable is left out of the model when cross validation is performed. The values of the Goodness Of Model in table 3.10 can be compared with those calculated for ordinary and universal kriging. If we only consider the results, for the cases where data are temporally reconstructed by GLM, cokriging seems to perform better than the two other kriging variants. However the Goodness Of Model is lower when data are temporally reconstructed by LOESS. The performance of the different models is discussed further in chapter 5.

Week	$\sum(DIN - \widehat{DIN})^2$	m	Goodness Of Model
1776	1239	16	77.40
1837	1162	19	61.20
1854	2.32	19	0.120
1919	15.5	17	0.91
Total	2419	71	34.10

Table 3.10: *Cross validation of cokriging for DIN.*

Dissolved Inorganic Phosphorus

Figure 3.29 shows the spatial distribution, and corresponding map of standard deviations, of DIP in the winter. The same spatial pattern, as computed by ordinary or universal kriging, is seen. However the highest concentrations of DIP in the north-western part of Kattegat, are higher than those predicted by the other two variants of kriging. The lower map in figure 3.29 shows the computed standard deviations of the predictions. For DIN it was found that these were lower when spatial predictions are computed by cokriging, compared to ordinary and universal kriging. For DIP the standard deviations are of the same order of magnitude when the different variants of kriging are used.

The spatial distribution of DIP in the summer is shown in figure 3.30. The same spatial pattern and order of magnitude of the concentration of DIN, as computed by ordinary or universal kriging when data are temporally reconstructed by GLM, is seen.

The cokriging model for DIP has been evaluated by cross validation in the same way as described in section 3.3.1. Only the primary variable is left out of the model when cross validation is performed. The values of the Goodness Of Model in table 3.11 can be compared with those calculated for ordinary and universal kriging. If only the results, for the cases where data are temporally reconstructed by GLM, are considered, cokriging seems to perform better than the two other kriging variants. It also seems to perform better than universal kriging, in the case where data are temporally reconstructed by LOESS, while ordinary kriging based on data reconstructed by LOESS seems to be better, when only cross validation is used as a Goodness Of Model criterion. The performance of the different models is discussed further in chapter 5.

Week	$(DIP - \widehat{DIP})^2$	m	Goodness Of Model
1776	1.55	16	0.097
1837	1.10	18	0.061
1854	0.041	19	0.0022
1919	0.80	17	0.047
Total	3.49	70	0.050

Table 3.11: *Cross validation of cokriging for DIP.*

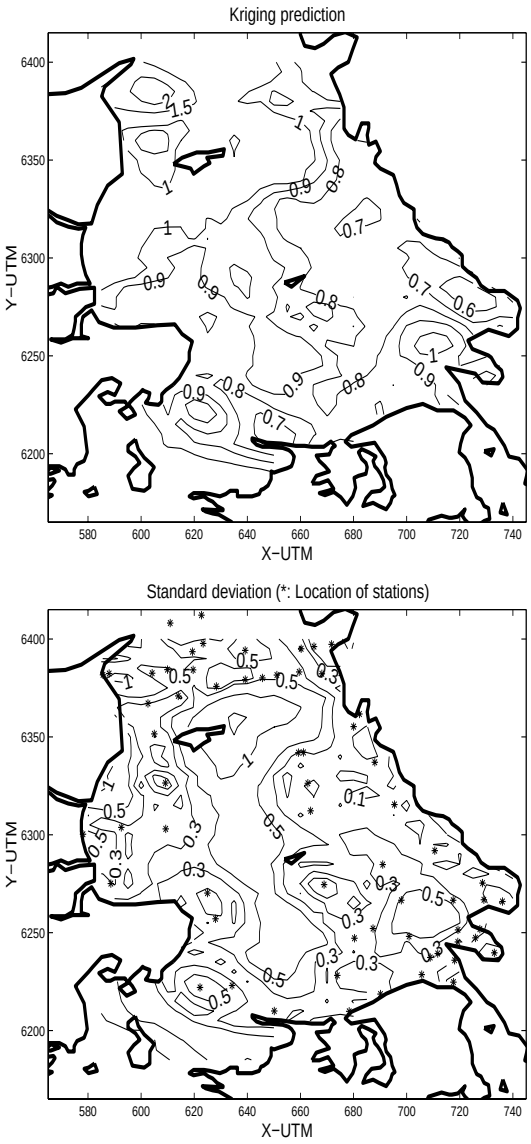

Figure 3.29: *Upper: Spatial distribution of DIP for a week in the middle of January 1994 computed by cokriging. Temporal reconstruction is computed by GLM. Lower: Corresponding map of the standard deviation.*

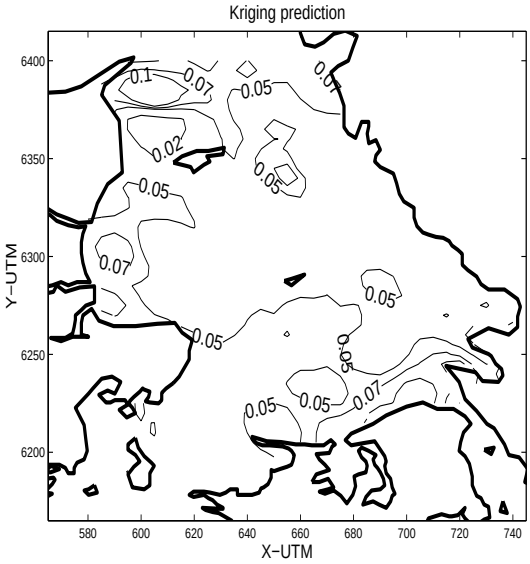

Figure 3.30: *Spatial distribution of DIP for a week in the middle of July 1995 computed by cokriging. Temporal reconstruction is computed by GLM.*

3.6 Sequential conditional simulation

What has been done until now is to compute the best possible predictions, by minimizing the prediction variance. This section deals with computation of the spatial distribution of a physical magnitude using sequential conditional simulation. This method produces a number of equally probable computations, referred to as realizations. Each realization reproduces the sample histogram -and semivariogram, while only the sample mean is reproduced when predictions are computed as described in the previous sections. The calculation of a number of realizations provides a good assessment of uncertainty, e.g. from 100 realizations one can make a histogram of 100 simulated values for each location. It can be shown that the average of a large number of realizations at a given location, is the best possible solution, i.e. the solution that can be computed by kriging. A histogram, of e.g. 100 realizations from a given location, describes the uncertainty with which we know the size of the physical magnitude at a given location. If a stochastic function is given by $Z(s)$, then the multivariate distribution of $Z(s)$ at the locations s_1, s_2, $\cdots$, s_n can be expressed as the product of its n univariate conditional distributions.

$$\begin{aligned} &f(s_1,\ s_2,\ \cdots,\ s_n;\ z_1,\ z_2,\ \cdots,\ z_n) = \\ &f(s_1;\ z_1) \times f(s_2;\ z_2 \mid Z(s_1) = z_1) \times \cdots\cdots \\ &\times f(s_n;\ z_n \mid Z(s_\alpha) = z_\alpha,\ \alpha = 1, \cdots, n-1) \end{aligned} \tag{3.51}$$

This means that $f(s_2;\ z_2 \mid Z(s_1) = z_1)$ represents the probability distribution of $Z(s_2)$ given that the outcome of $Z(s_1)$ is equal to z_1. Each realization, $z(s)$, is made up by drawing random values from each of the previous univariate and conditional distributions, e.g. the realization z_1 is drawn at random from the distribution $f(s_1;\ z_1)$, while the next realization, z_2, which is drawn from the conditional distribution $f(s_2;\ z_2 \mid Z(s_1) = z_1)$, depends on the simulated value z_1, and the method is therefore called sequential conditional simulation.

Multiple realizations can only be drawn if the distribution function is fully characterized. This is overcome by Gaussian sequential conditional simulation, which applies the sequential conditional simulation in (3.51) using the multivariate normal distribution, which is fully described by its covariance, i.e.

$$C(s,h) = E[(Z(s+h) - \mu)(Z(s) - \mu)] \tag{3.52}$$

Data are transformed prior to simulation into data which are normally distributed. Such a transformation is called a normal score transformation, and transformed data are referred to as normal score data. After the analysis a back transformation into original values is made. The transformation/back transformation can be written as

$$\begin{aligned} y(s) &= \phi\{z(s)\} \\ z(s) &= \phi^{-1}\{y(s)\} \end{aligned} \tag{3.53}$$

where z is original data, y is transformed data and Φ is the normal score transformation.

Simple kriging with a semivariogram model for the normal score data is used for computation of each realization. This method is very similar to ordinary kriging, described in section 3.3. In simple kriging a constant known mean is assumed, and kriging is performed for the residuals, i.e. predictions are calculated in the following way, [Rossi et al., 1993], [Caers, 2000] and [Deutsch and Journel, 1992].

$$\hat{z(s_0)} - \mu = \sum_{i=1}^{n} \lambda_i [z(s_i) - \mu] \tag{3.54}$$

3.6.1 Results

In the following the results of Gaussian sequential conditional simulation will be presented for DIN in wintertime, i.e. week 1776. Data for the computation are measured or temporally reconstructed by GLM. The method has also been applied to the corresponding set of data for DIP, but in order not to include too many figures in the text, these are shown in appendix H. The use of the method aims at exemplifying how knowledge of uncertainty of prediction at a given location in Kattegat can be obtained. The estimated semivariogram of the normal score data has been modelled by a spherical semivariogram model, with a nugget effect of 0.4, a sill of 1 and a range of 17 kilometers. The estimated and modelled semivariogram of the normal score data is shown in figure 3.31.

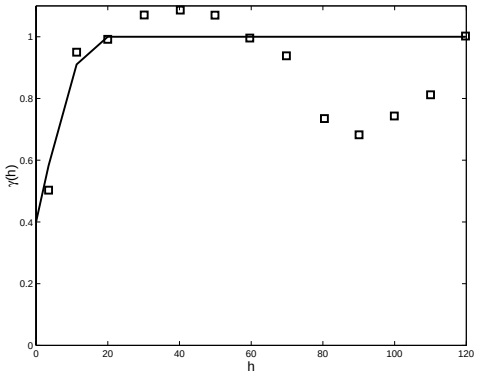

Figure 3.31: *Semivariogram for log-transformed normal score DIN data.*

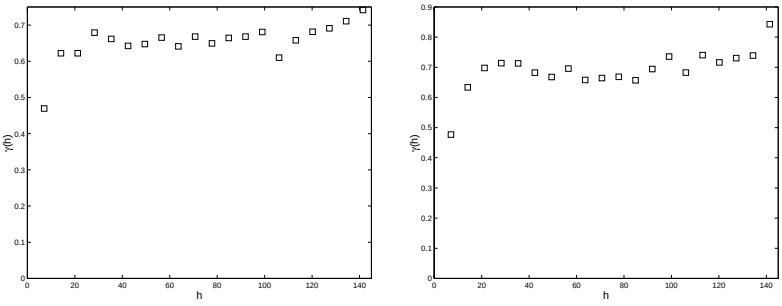

Figure 3.32: *Semivariograms of simulated log-transformed DIN data for two different realizations.*

The semivariogram of the normal score data is used to compute 100 realizations of DIN. To check whether or not the semivariogram of the simulated log-transformed concentrations of DIN is reproduced, it has been computed for two of the 100 realizations and is shown in figure 3.32. When comparing the semivariograms in figure 3.32 with the upper semivariogram in figure 3.6, a good agreement is found. Also the histograms of the two realizations,

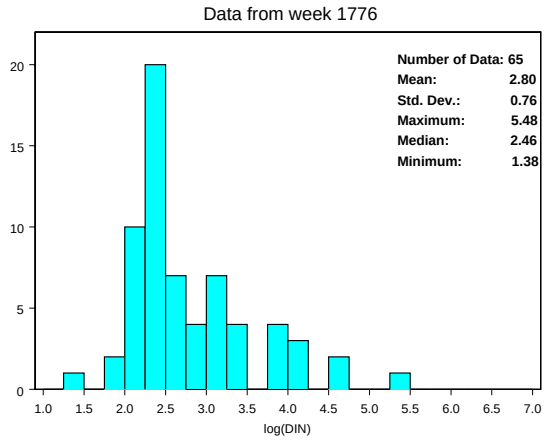

Figure 3.33: *Histogram of log-transformed observations of DIN. Data are measured or reconstructed by GLM.*

shown in figure 3.34 seem to be very similar to the one for the observed data, which is given in figure 3.33.

The two realizations of DIN are shown in figure 3.35, and it is seen that we do not obtain the same smooth map of the concentration of DIN, as is the case when the best possible solution is found using kriging.

From all the 100 realizations it is possible to obtain knowledge of uncertainty in any point in Kattegat. As an example the locations marked on the map in figure 3.36 have been chosen, and a histogram of the 100 simulated values at these two locations has been drawn.

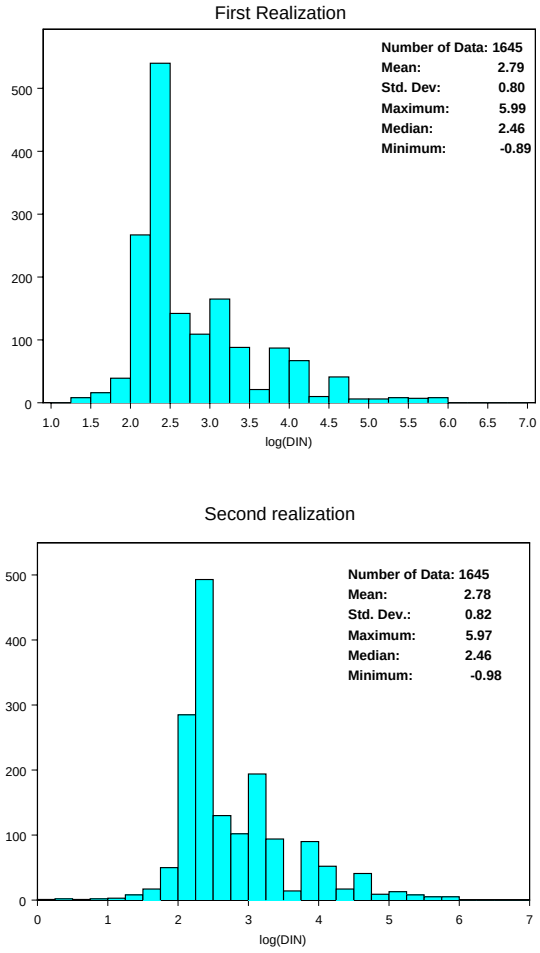

Figure 3.34: *Histogram of simulated log-transformed DIN data for two different realizations.*

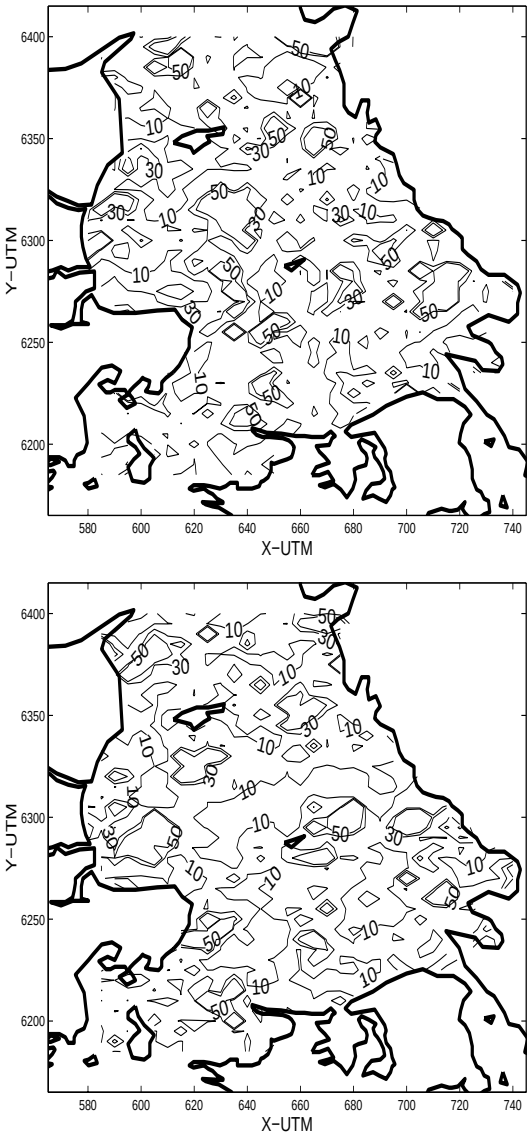

Figure 3.35: *Mapping of two different realizations of DIN.*

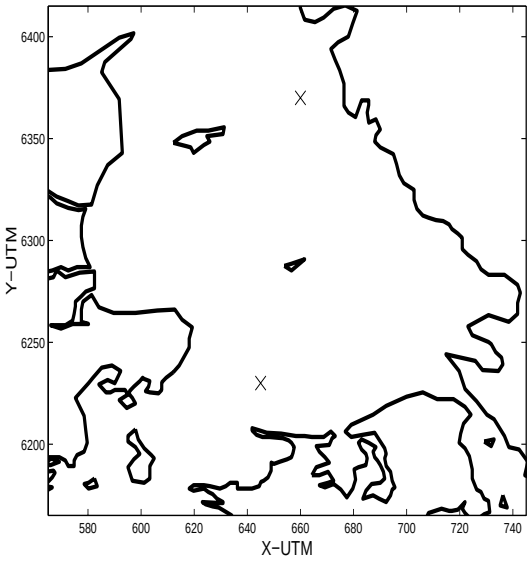

Figure 3.36: *The two locations where the histogram of 100 different realizations has been plotted.*

The histograms for the two locations are shown in figure 3.37. It is seen that both the standard deviation and the mean of the simulated values at the northern location is larger than for the southern. The coefficient of variance (CV) can be calculated as the standard deviation divided by the mean in order to account for the difference in mean value. This results in:

Southern location: CV=32.89/23.69=1.39
Northern location: CV=46.80/26.09=1.79

For DIP the corresponding results are shown in appendix H. For this variable a greater coefficient of variance is found for the southern location.

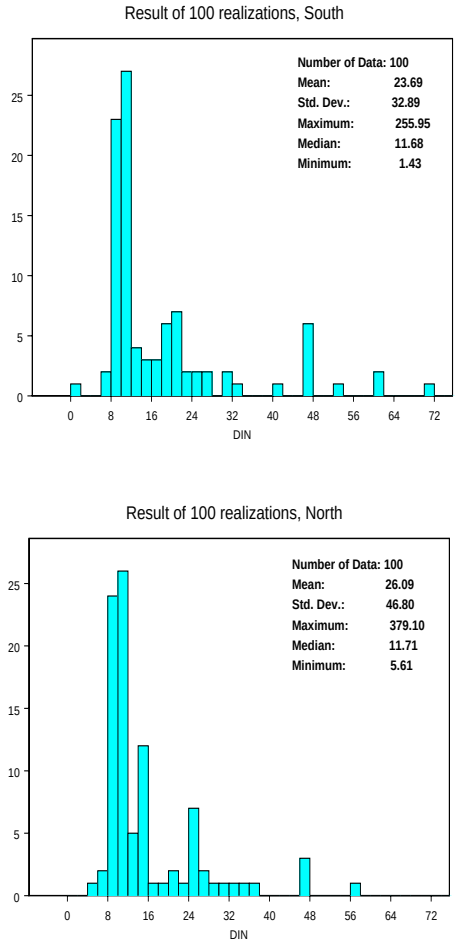

Figure 3.37: *Histograms of DIN for 100 realizations. Two different locations in Kattegat.*

3.7 Summary of spatial data analysis

In this chapter different methods for computing the spatial distribution of DIN and DIP have been described, applied and compared for four different weeks, representing the four seasons of the year. Data for the computation are temporally reconstructed by GLM or LOESS, when no measurement has been carried out. For weeks where measurements are available, these are used instead of the reconstructed values. Both GLM and LOESS are used for temporal reconstruction in order to examine whether or not the 24 stations with the highest sampling frequencies are sufficient to compute weekly maps of the spatial distribution of DIN and DIP.

Prior to the computation of predictions, the spatial variability has been described by estimation and modelling the semivariogram. This showed that Kattegat is anisotropic with respect to both variables. Data are more spatially correlated in the south/east - north/west direction, while the lowest correlation is found in the perpendicular direction. The range of the estimated semivariogram depends largely on the method used for temporal reconstruction, i.e. much higher ranges are found when data are temporally reconstructed by LOESS.

When ordinary kriging is used to compute the spatial distribution of DIN and DIP based on temporal reconstruction by GLM, high concentrations are found along the eastern coast of Jutland and in the north/eastern part of Kattegat. This seems rational from a physical point of view. When data for the spatial predictions are temporally reconstructed by LOESS, the highest concentrations are predicted in the open sea, which is not understandable. This phenomenon is explained by the low number of observations.

When universal kriging is used to compute the spatial distribution of DIN and DIP based on temporal reconstruction by GLM, high concentrations are predicted along the eastern coast of Jutland and in the north/eastern part of Kattegat, and the maps are very similar to those computed by ordinary kriging. When data for the spatial predictions are temporally reconstructed by LOESS, too high concentrations are predicted in the open sea. In the summer the trend causes a dramatically overestimation of the concentration of both DIN and DIP in the coastal areas.

The use of cokriging has only been based on data temporally reconstructed

by GLM using the depth of water as a secondary variable. The pattern of the spatial distribution of DIN and DIP is found to be the same as computed by ordinary -and universal kriging when data are temporally reconstructed by GLM.

Table 3.12 summarizes the results of the cross validation. It is seen that cokriging is the best method for computing the spatial distribution of DIN, if GLM is used for temporal reconstruction. However the use of LOESS seem to be preferably, which does not agree with the fact, that these results are difficult to explain from a physical point of view. The same is seen for DIP, but in this case cokriging with GLM-data is preferably compared to universal kriging with LOESS-data. The performance of the models will be further discussed in chapter 5.

Variable	Kriging	Reconstruction	Goodness Of Model
DIN	Ordinary	GLM	51.13
DIN	Universal	GLM	72.92
DIN	Cokriging	GLM	31.10
DIN	Ordinary	LOESS	9.694
DIN	Universal	LOESS	15.61
DIP	Ordinary	GLM	0.066
DIP	Universal	GLM	0.087
DIP	Cokriging	GLM	0.050
DIP	Ordinary	LOESS	0.021
DIP	Universal	LOESS	0.053

Table 3.12: *Comparison of Goodness Of Model values for DIN and DIP.*

The kriging methods compute the optimal spatial predictions, by minimizing the prediction variance. The last section of this chapter shows how Gaussian sequential conditional simulation can be used to obtain knowledge of uncertainty of the predictions. This is done by computing 100 equally probable realizations. The method is not used to make any conclusions about the performance of the different models.

Chapter 4
Spatiotemporal data analysis

4.1 Introduction to spatiotemporal data analysis

Spatiotemporal data refers to spatial data measured over time. Traditionally such datasets have been analyzed as strictly temporal or spatial data, either by excluding a part of the dataset in the computation or by eliminating the temporal or spatial dimensions in some way, e.g. the spatial dependence could be removed by computing spatial averages, leading to the use of time series models. The strictly spatial predictions are often computed by kriging, while time series models or some kind of regression model compute temporal predictions. These statistical methods have been applied to numerous datasets, and are described in many books, see e.g. [Cressie, 1993], [Isaaks and Srivastava, 1989] and [Shumway, 1988].

Alternatively the prediction could be computed using a spatiotemporal method, which includes both the temporal and the spatial dimensions, i.e. it includes more of the information from the dataset. This principle is illustrated in figure 4.1, which can be compared to the two step predictions illustrated in figure 2.7. Such statistical methods are much less developed than methods for analyzing strictly spatial or temporal data. This chapter

describes and tests different methods for analyzing spatiotemporal data.

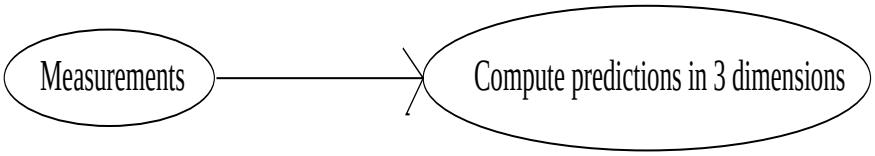

Figure 4.1: *Principle of the computation of predictions in 3 dimensions.*

4.2 Kriging in three dimensions

Kriging can be extended into three dimensions by using a semivariogram model, which describes the spatial variability in three dimensions. A general description of the theory of kriging and the semivariogram is given in chapter 3, and will therefore not be described further in this chapter. This section goes through the estimation and modelling of the semivariogram in three dimensions for log-transformed DIN and DIP. Furthermore ordinary kriging is applied for computing spatiotemporal predictions of the variables. Other variants of kriging could also have been used. Only measurements are used in the analysis, i.e. no method for temporal reconstruction of data is used prior to the spatiotemporal kriging model.

4.2.1 The three dimensional semivariogram

Dissolved Inorganic Nitrogen

The semivariogram of log-transformed DIN has been estimated and modelled in three dimensions, by including time in the separation distance. A time step of one week is used, and the separation distance is

$$h = \sqrt{h_x^2 + h_y^2 + h_t^2} \tag{4.1}$$

where t is time. The main problem of including time in the semivariogram is that one week is not the same as one kilometer. However this can be

accounted for by including anisotropy in the model, i.e. allowing for different spatial correlation in different directions. The way to do this is first to identify the directions of anisotropy in the horizontal plane, i.e. the plane defined by x and y. Afterwards the search angle is dipped in steps of 45 degress in order to find the vertical anisotropy direction, see figure 3.5. The horizontal directions of anisotropy are found by turning the search angle in steps of 45 degrees in the horizontal plane. The isotropic nugget effect and sill are found to be 0.3 and 3.3 respectively, when the spherical semivariogram model is used. The direction with the highest range=73 kilometers is found to be 45 degrees clockwise from the north, i.e. in the north/east-south/west direction, while the direction with the lowest range=55 kilometers is the perpendicular direction, i.e. 135 degrees clockwise from the north. Estimated and modelled horizontal semivariograms with highest and lowest range are shown in figure 4.2.

After the horizontal directions of anisotropy have been identified, the coordinate system is rotated according to these. When the search angle is dipped vertically in steps of 45 degress, the ranges shown in table 4.1 are found. The lowest range is found in a direction perpendicular to the horizontal plane, which means that no further rotation is needed.

Vertical rotation	Range
45 degrees	26
90 degrees	18
135 degrees	27

Table 4.1: *Estimated anisotropic ranges for log-transformed DIN, when the search angle is dipped into the vertical plane.*

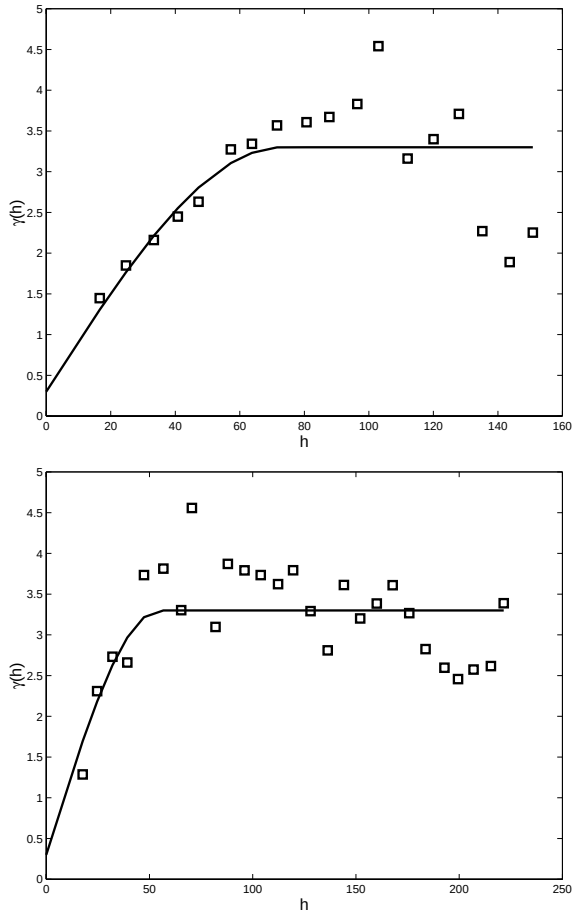

Figure 4.2: *Estimated and modelled horizontal semivariograms for log-transformed DIN. Upper: Search angle is 45 degrees clockwise from the north. Lower: Search angle is 135 degrees clockwise from the north.*

Dissolved Inorganic Phosphorus

The semivariogram of log-transformed DIP has been estimated and modelled in three dimensions, in the same way as for log-transformed DIN. The horizontal directions of anisotropy are found by turning the search angle

in steps of 45 degrees in the horizontal plane. The isotropic nugget effect and sill are found to be 0.4 and 1.7 respectively, when the spherical semivariogram model is used. The north/south direction, with a range of 100 kilometers, is found to be the direction with the highest range, while the direction with the lowest range is the perpendicular direction, i.e. the east/west direction. The range in this direction is found to be 66 kilometers. Estimated and modelled horizontal semivariograms with highest and lowest range are shown in figure 4.3.

A rotation of the horizontal coordinate system is not performed since the directions of anisotropy coincide with the axes of the coordinate system. The search angle is dipped vertically in steps of 45 degrees, and the ranges shown in table 4.2 are found. The lowest range is found in a direction per-

Vertical rotation	Range
45 degrees	32
90 degrees	17
135 degrees	30

Table 4.2: *Estimated anisotropic ranges for log-transformed DIP, when the search angle is dipped into the vertical plane.*

pendicular to the horizontal plane, which means that no rotation is needed.

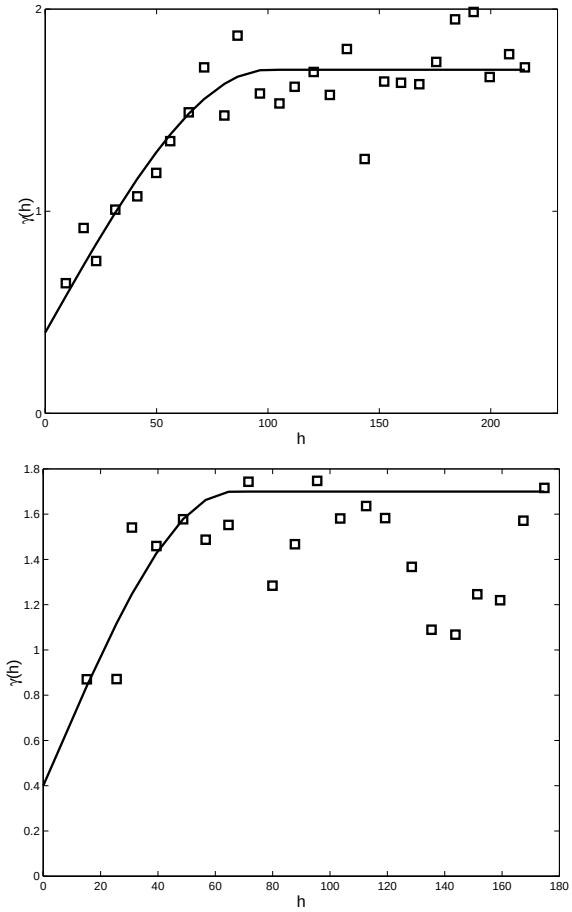

Figure 4.3: *Estimated and modelled horizontal semivariograms for log-transformed DIP. Upper: Search angle is in north/south direction. Lower: Search angle is in east/west direction.*

4.2.2 Ordinary kriging in three dimensions

Dissolved Inorganic Nitrogen

The spatial distribution of DIN computed by 3 dimensional kriging, for a winter- and a summer week, is shown in figure 4.4, while the corresponding maps for a spring- and a autumn week are shown in figure I.1 in appendix I. When the results for the winter week are compared to those computed in the 2 dimensional case, it is seen that the computed concentrations of DIN are in the same order of magnitude. The high concentrations in the northern part of Kattegat, computed by ordinary- and universal kriging in 2 dimensions, are not shown in the upper part of figure 4.4. It is also seen that we are not able to compute predictions in an area by the Swedish coast, when using the 3 dimensional model, because the number of points in the search area is not high enough to compute a stable result. For the week in spring the same area without predictions is found. The computed concentrations of DIN in this week, are of the same order of magnitude when using ordinary kriging in 2 and 3 dimensions. The computed concentrations in summertime seem to be too high, compared to what is usually found in Kattegat in the summer, and they are higher than what is computed by the 2 dimensional ordinary kriging model.

The 3 dimensional kriging models can be evaluated using cross validation, in the same way as it is done for the single weeks in the 2 dimensional case, i.e. a single value of log(DIN) is left out of the model each time. Afterwards the predicted value at the same location, given by x, y and t, is assigned to the left out observation, and back transformed using equation (3.32). The Goodness Of Model can be calculated as

$$\text{Goodness Of Model} = \frac{1}{m}\sum_{k=1}^{m}(DIN_k - \widehat{DIN_k})^2 \tag{4.2}$$

where m is the number of observations of DIN in the five year period. The cross validation results in a Goodness Of Model of 23.78, which can be compared to corresponding numbers computed by ordinary kriging, universal kriging and cokriging in two dimensions. This comparison is shown in table 4.3. The 3 dimensional kriging model seeems to perform better than the 2 dimensional, when GLM is used for temporal reconstruction. However 2 dimensional kriging based on data temporally reconstructed by LOESS seems to be the best model.

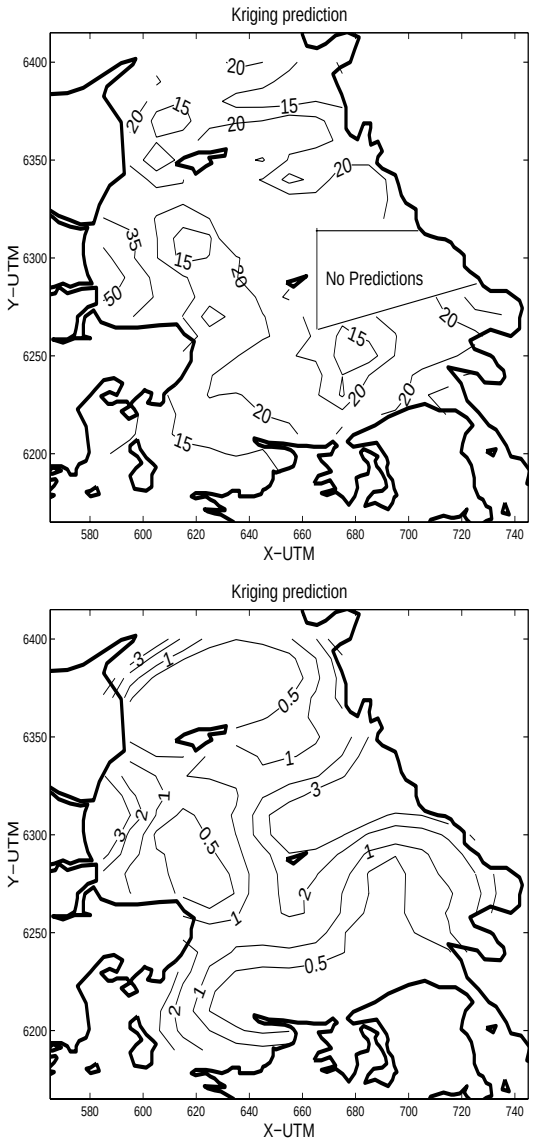

Figure 4.4: *Spatial distribution of DIN, computed by three dimensional ordinary kriging. Upper: A week in the middle of January 1994. Lower: A week in the middle of July 1995.*

Dimension	Kriging	Reconstruction	Goodness Of Model
2	Ordinary	GLM	51.13
2	Universal	GLM	72.92
2	Cokriging	GLM	34.10
2	Ordinary	LOESS	9.694
2	Universal	LOESS	15.61
3	Ordinary		23.78

Table 4.3: *Comparison of Goodness Of Model values for DIN.*

Dissolved Inorganic Phosphorus

The spatial distribution of DIP computed by 3 dimensional kriging, for a winter- and a summer week, is shown in figure 4.5, while the corresponding maps for a spring- and a autumn week are shown in figure I.2 in appendix I. A very good agreement, between the results computed with the 2 and 3 dimensional model, is found for all the four weeks. Both the order of magnitude and the spatial distribution of concentrations of DIP are found to be very similar. The 3 dimensional kriging model has been cross validated, and a value of the Goodness Of Model of 0.0324 is found. As for DIN, this value can be compared to those computed by the 2 dimensional kriging models. Such a comparison is shown in table 4.4. The 3 dimensional kriging model

Dimension	Kriging	Reconstruction	Goodness Of Model
2	Ordinary	GLM	0.066
2	Universal	GLM	0.087
2	Cokriging	GLM	0.050
2	Ordinary	LOESS	0.021
2	Universal	LOESS	0.053
3	Ordinary		0.032

Table 4.4: *Comparison of Goodness Of Model values for DIP.*

seems to perform better than the 2 dimensional, when GLM is used for temporal reconstruction, and also better than universal kriging based on data temporally reconstructed by LOESS. However 2 dimensional ordinary kriging based on data temporally reconstructed by LOESS seems to be the best model. The performance of the models will be further discussed in chapter 5.

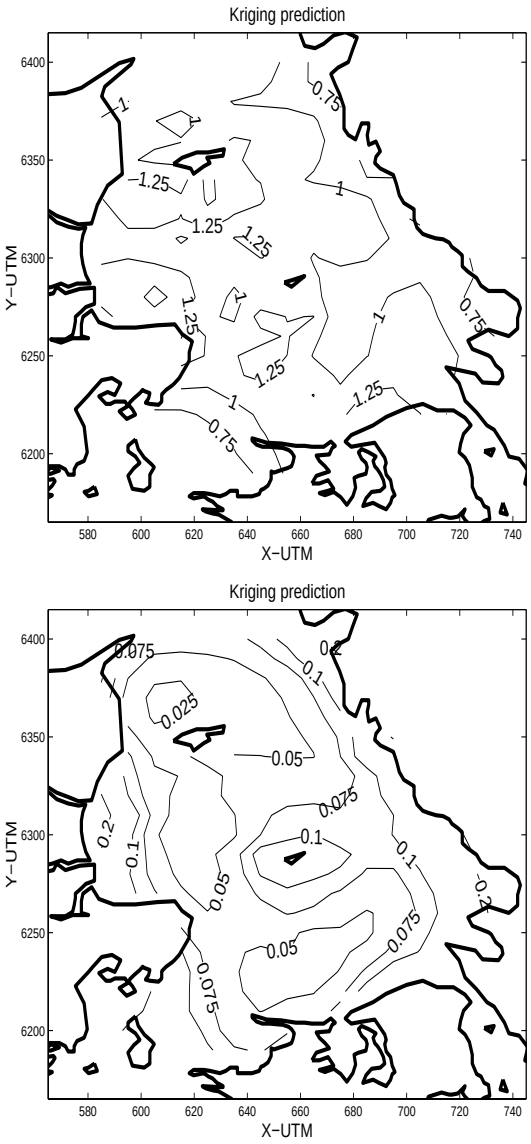

Figure 4.5: *Spatial distribution of DIP, computed by three dimensional ordinary kriging. Upper: A week in the middle of January 1994. Lower: A week in the middle of July 1995.*

4.3 Locally weighted regression in three dimensions

Locally weighted regression can also be used in three dimensions by including time in the prediction. The theory of locally weighted regression is described in section 2.3. Polynomias of second order are used in the estimation, and the predicted value, $\hat{y}_i$, can therefore be written as

$$\begin{aligned} \hat{y}_i = \hat{g}(x_i) &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2} + \beta_{11} x_{i1}^2 \\ &+ \beta_{22} x_{i2}^2 + \beta_{13} x_{i1} x_{i3} + \beta_{23} x_{i2} x_{i3} \\ &+ \beta_{123} x_{i1} x_{i2} x_{i3} + \beta_{33} x_{i3}^2 \end{aligned} \tag{4.3}$$

where $x_i = (x_{i1}, x_{i2}, x_{i3})$ are independent variables in the 3-dimensional space. Only original, measured values are used in the analysis. x_{i1} and x_{i2} are the UTM coordinates given in kilometers. The values of these are within the range from 578 to 736 for the x-coordinates, and from 6190 to 6416 for the y-coordinates. The values of the 260 weeks in the five year period are given as the number of weeks after the first of January 1960, i.e. going from 1722 to 1981. When computing predictions in 3 dimensions we are faced with the problem that one kilometer is not the same as one week. In the 3 dimensional kriging model this is dealt with by including anisotropy. In this case the values of weeks will be divided by a constant a.

When using locally weighted regression for prediction of concentrations in 1 dimension, the local area is given by the timeinterval including a certain fraction, defined by the bandwidth, of the total number of measurements. In 3 dimensions the local area is given by the cube, which includes a certain fraction of the total number of measurements. As in the 1 dimensional case, the size of the local area is defined by the bandwidth.

4.3.1 Dissolved Inorganic Nitrogen

The optimization of the two parameters, the constant a and the bandwidth, is done by computing Akaike's information criterion (AIC), for different combinations of the parameters, to find the minimum value of AIC. The results are shown on the greyscale map in figure 4.6, where light colours represent high values. It is seen that AIC is almost independent of a for

values of the bandwidth higher than 0.16. The minimum of AIC is found for a bandwidth of 0.2 and a value of the constant a of 10. It has also been attempted to compute AIC for higher values of the bandwidth and other values of the constant a, which resulted in higher values of AIC.

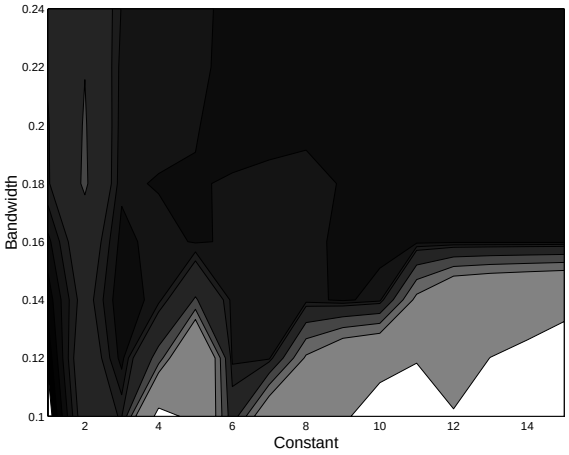

Figure 4.6: *Greyscaled map of Akaike's information criterion (AIC), used for optimization of the two parameters, i.e. bandwidth and constant, in the 3 dimensional locally weighted regression for log-transformed DIN. Light colours represent high values of AIC.*

The spatial distribution of DIN, for a week in the middle of January 1994, computed by locally weighted regression, is shown in figure 4.7. Very low concentrations are found in the southern part of Kattegat. These are much too low for the wintertime. Furthermore the spatial distribution is very scattered, i.e. the model does not compute the smooth map of the concentrations like the kriging model. The same is found when mapping the spatial distribution for the other weeks, and these results will therefore not be shown. The map in figure 4.7 indicates that the local areas, defined by the bandwidth, are too small. The size of the local areas has therefore been increased by increasing the bandwidth to 0.4. However this does not improve the results considerably. The main problem of the method seems to be that we are predicting the concentration of DIN in a 3 dimensional grid of 350,000 points, based on less than 2,000 measurements.

The optimal combination of the two parameters results in the plot of resid-

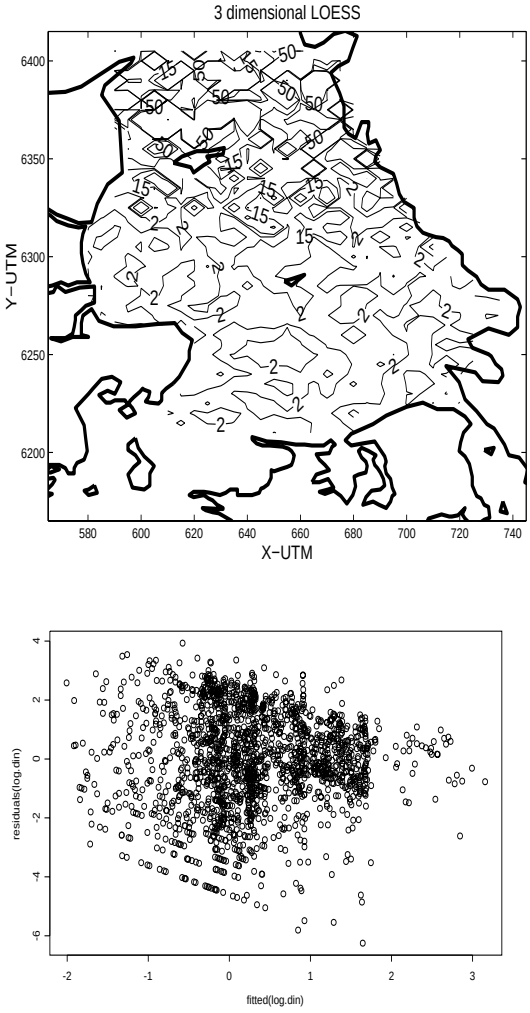

Figure 4.7: *Spatial distribution and corresponding plot of residuals for DIN for a week in the middle of January 1994, computed by locally weighted regression in 3 dimensions. The following values of the two parameters are used: Bandwidth=0.2, a=10.*

uals against the predicted value, shown in figure 4.7. The residuals seem to be higher for low predicted values, i.e. a trend seems to be present.

4.3.2 Dissolved Inorganic Phosphorus

As for DIN, the optimization of the two parameters, the constant a and the bandwidth, is done by computing Akaike's information criterion (AIC) for different combinations of the parameters, to find the minimum value of AIC. The results are shown on the greyscale map in figure 4.8, where light colours represent high values. The minimum of AIC is found for a bandwidth of 0.28 and a value of the constant a of 10. However the value of AIC seems to be almost independent of a, for values higher than 5. The

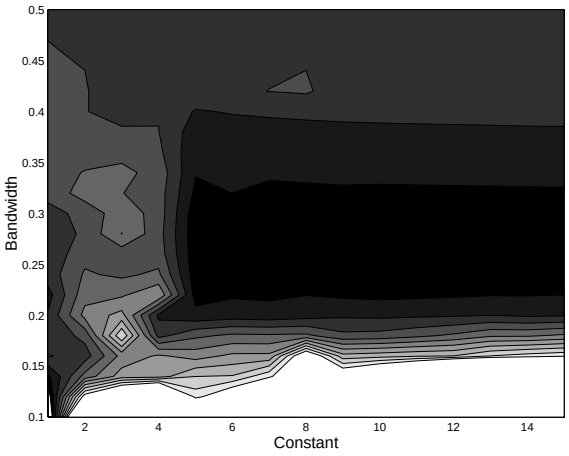

Figure 4.8: *Greyscaled map of Akaike's information criterion (AIC), used for optimization of the two parameters, i.e. bandwidth and constant, in the 3 dimensional locally weighted regression for log-transformed DIP. Light colours represent high values of AIC.*

spatial distribution of DIP, for a week in the middle of January 1994, computed by locally weighted regression, is shown in figure 4.9. As for DIN the lowest concentrations are found in the southern part of Kattegat. The computed spatial distribution of DIP is very scattered, which is the same as found when modelling the concentration of DIN. The same is found when

mapping the spatial distribution for the other weeks, and these results will therefore not be shown.

The optimal combination of the two parameters results in the plot of residuals against the predicted value, shown in figure 4.9. As for DIN a trend in the residuals seems to be present.

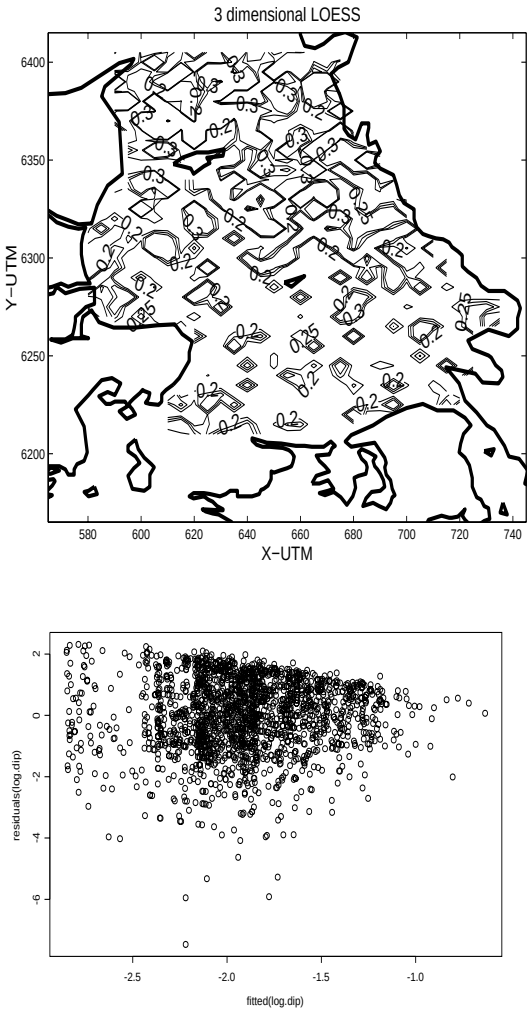

Figure 4.9: *Spatial distribution and corresponding plot of residuals for DIP for a week in the middle of January 1994, computed by locally weighted regression in 3 dimensions. The following values of the two parameters are used: Bandwidth=0.28, a=10.*

4.4 ARIMA processes

This section is about modelling of ARIMA processes, and is included in the spatiotemporal part of the thesis, because it uses the temporal correlation in data in an improved way, compared to what has been done until now. The model will be applied to different stations, and the model parameters can be seen as regionalized variables, thus kriging can be used to determine the parameters at any location in Kattegat.

ARIMA processes are linear, and can be illustrated as shown in figure 4.10.

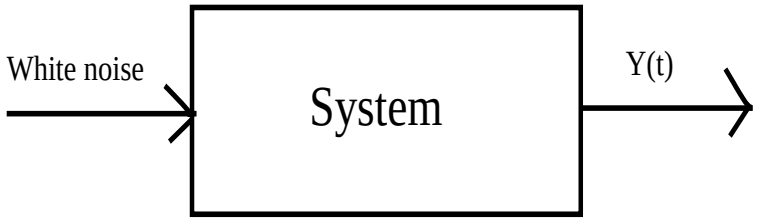

Figure 4.10: *Sketch of a linear stochastic process.*

The linear stochastic process is defined as

$$Y_t - \mu = \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i} \tag{4.4}$$

where ϵ_t is white noise, i.e. it is a sequence of uncorrelated stochastic variables, which are identically distributed with a mean of 0 and a constant variance σ_ϵ^2. A class of linear stochastic processes, which are often applied to time series, is ARMA and ARIMA processes. An ARMA(p,q) process is a combination of an autoregressive process of order p, AR(p), and a moving average process of order q, MA(q). This is given as

$$(Y_t - \overline{Y}) + \phi_1(Y_{t-1} - \overline{Y}) + \cdots + \phi_p(Y_{t-p} - \overline{Y}) = \epsilon_t + \theta_1\epsilon_{t-1} + \cdots + \theta_q\epsilon_{t-q} \tag{4.5}$$

where ϵ_t is white noise. By introducing the back shift operator, B, equation (4.5) can be written as.

$$\phi(B)(Y_t - \overline{Y}) = \theta(B)\epsilon_t \tag{4.6}$$

When analyzing time series it is often found that data cannot be described by a stationary process. This is overcome by computing differences. The model is called ARIMA, and can be written as

$$\phi(B)\nabla^d(Y_t - \overline{Y}) = \theta(B)\epsilon_t, \quad d \in \mathbb{N} \tag{4.7}$$

where $\nabla^d = (1-B)^d$ and ϵ_t is white noise. A special class of ARIMA models are used to model time series containing periodic components, which tend to repeat with a period of S. These are referred to as ARIMA models with seasonality, and can be written as

$$\phi(B)\Phi(B^s)\nabla^d\nabla_s^D(Y_t - \overline{Y}) = \theta(B)\Theta(B^s)\epsilon_t \tag{4.8}$$

where ϕ, Φ, θ and Θ are polynomias, and $\nabla_s^D = (1 - B^s)^D$ is the seasonal difference operator. The order of the ARIMA model is identified from the autocorrelation (ACF) -and partial autocorrelation function (PACF), and afterwards the parameters of the model can be estimated. Different methods for estimation of the parameters exist, e.g. maximum likelihood, unconditional least squares or conditional least squares. In this thesis conditional least squares will be used. For further information about identification and estimation, see [Shumway, 1988]. It should always be tested whether or not the model parameters are significantly different from 0. If the parameters are denoted $\hat{\theta}_i$, with a variance $\hat{\sigma}^2_{\theta_i}$, then the hypotheses

$$H_0 : \ \theta_i = 0 \quad \text{against} \quad H_1 : \ \theta_i \neq 0 \tag{4.9}$$

may be tested using the statistic

$$T = \frac{\hat{\theta}_i}{\hat{\sigma}_{\theta_i}} \tag{4.10}$$

which is t-distributed with $f = N - p - q - 1$ degrees of freedom, where N is the number of observations, that the estimation of $\hat{\theta}_i$ is based on.

4.4.1 Results

In the following the results of the application of ARIMA models will be shown for DIN. The method could be applied to DIP as well, but this has not been done. Prior to applying ARIMA models to different stations in Kattegat, the General Linear Model is used to fill out gaps in the time

series, in order to obtain a resolution of one week.

The estimation of ARIMA models will be exemplified for station 1001. The autocorrelation function (ACF) and partial autocorrelation function (PACF) for log-transformed DIN at station 1001 are shown in figure 4.11 for 60 lags, representing 60 weeks. ACF indicates that a seasonal non-

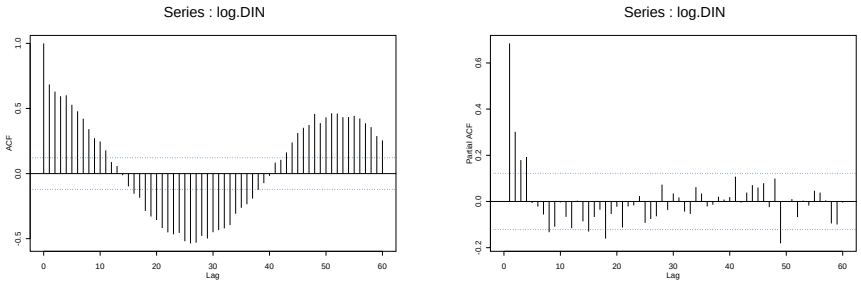

Figure 4.11: *Left: ACF for log-transformed DIN at station 1001. Right: PACF for log-transformed DIN at station 1001.*

stationarity with a period of 52 is present. This means that we have yearly variations of log-transformed DIN, and that we should compute ACF and PACF for $\nabla_{52} Y_t = \nabla_{52} \log(DIN)_t$. These are shown in figure 4.12. Both

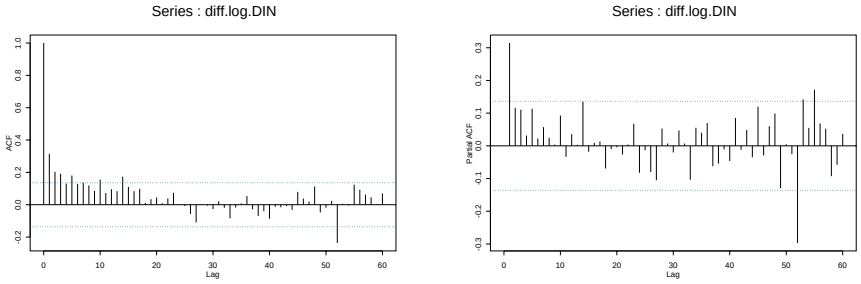

Figure 4.12: *Left: ACF for seasonal differenced log-transformed DIN at station 1001. Right: PACF for seasonal differenced log-transformed DIN at station 1001.*

ACF and PACF are strongly significant in lag 52, indicating that this seasonality should be included in the model. This can be done in different ways, and the models can be compared by computing Akaike's information criterion. It has been found that the following model results in a good description of log-transformed DIN.

$$(1 + \phi_1 B)\nabla_{52}(Y_t - \overline{Y}) = (1 + \Theta_1 B^{52})\epsilon_t \tag{4.11}$$

For station 1001 and 20004, representing an open sea -and a coastal station, respectively, the parameters are shown in table 4.5. The results of applying

Station	$\overline{Y}$	ϕ_1	Θ_1
1001	0.0204	-0.3577	-0.4725
20004	-0.06609	-0.2369	-0.5295

Table 4.5: *Parameters of ARIMA models for station 1001 and 20004.*

the model to DIN at station 1001 and 20004 are shown in figure 4.13. The model is not able to compute estimates for the first year, i.e. 1993, because of the seasonality.

The order of the model is assumed to be the same for all stations, i.e. all the time series can be described by (4.11). This assumption has been tested for three other stations, and seems to be reasonable. The parameters of the model for all the stations are shown in appendix J.

Now that we have estimated the parameters at different stations, kriging can be used to compute these at any given location in Kattegat. In order to calculate the back transformation from log-transformed DIN to DIN, we also have to determine the standard deviation by kriging. The semivariograms, of the parameters and the standard deviation, have been estimated and modelled. These are shown in figure 4.14, and the parameters of the semivariogram models are shown in table 4.6. Isotropy is assumed, and the semivariograms have therefore been estimated by including all pairs of data.

The results of kriging the parameters as well as the standard deviation are shown in figure 4.15 and 4.16.

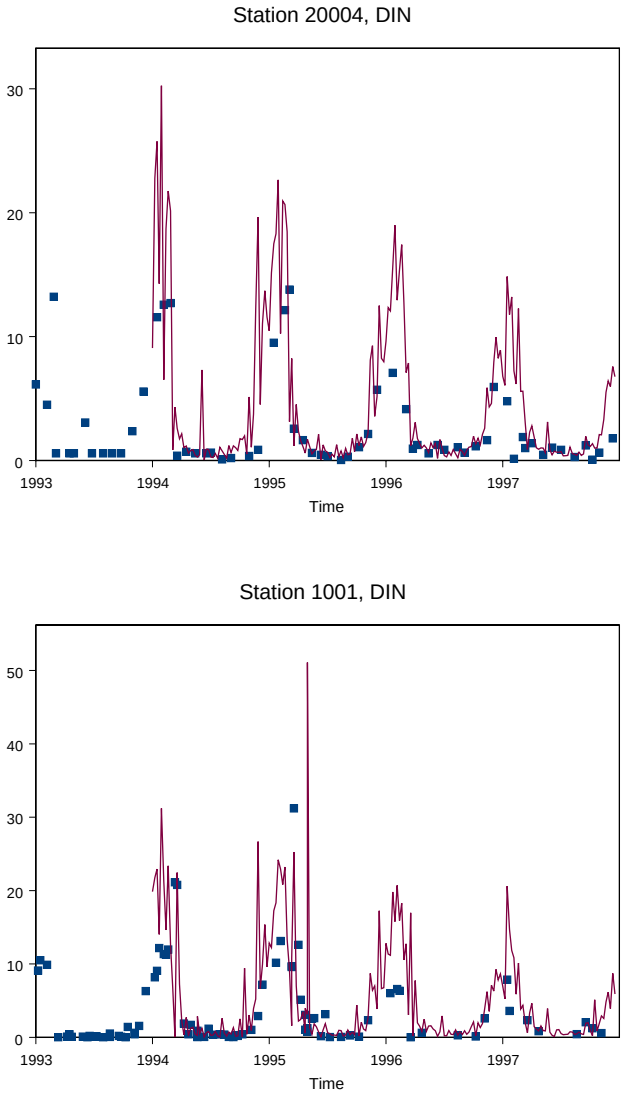

Figure 4.13: *The results of applying ARIMA models to DIN. Upper: Station 20004. Lower: Station 1001.*

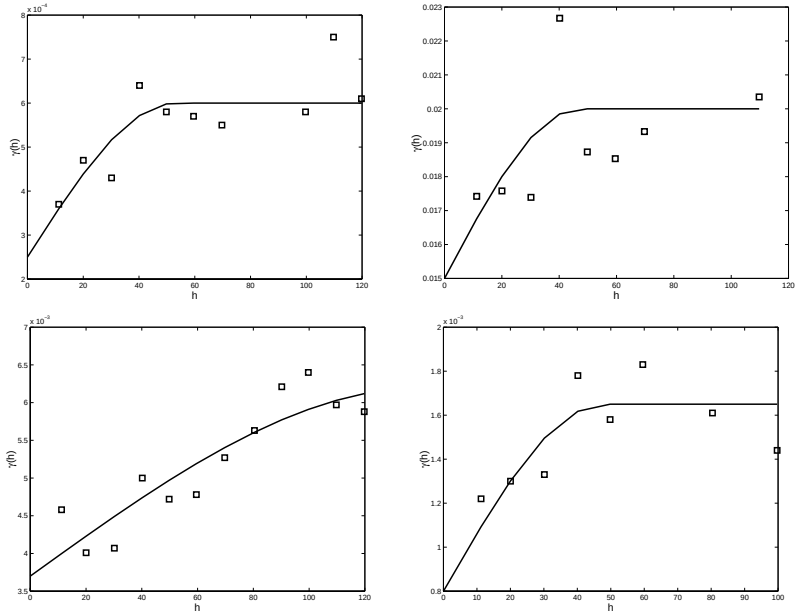

Figure 4.14: *Estimated and modelled semivariograms for parameters of ARIMA models and standard deviation. Upper left: Semivariogam for* $\overline{Y}$*. Upper right: Semivariogam for standard deviation. Lower left: Semivariogram for* ϕ_1*. Lower right: Semivariogram for* Θ_1*.*

Parameter	Model	Range	Sill	Nugget effect
$\overline{Y}$	Spherical	53	0.0006	0.00025
ϕ_1	Spherical	141	0.0062	0.0037
Θ_1	Spherical	48	0.00165	0.00080
Standard Deviation	Spherical	47	0.020	0.015

Table 4.6: *Type of semivariogram model and the range, sill and nugget effect for the parameters of ARIMA models.*

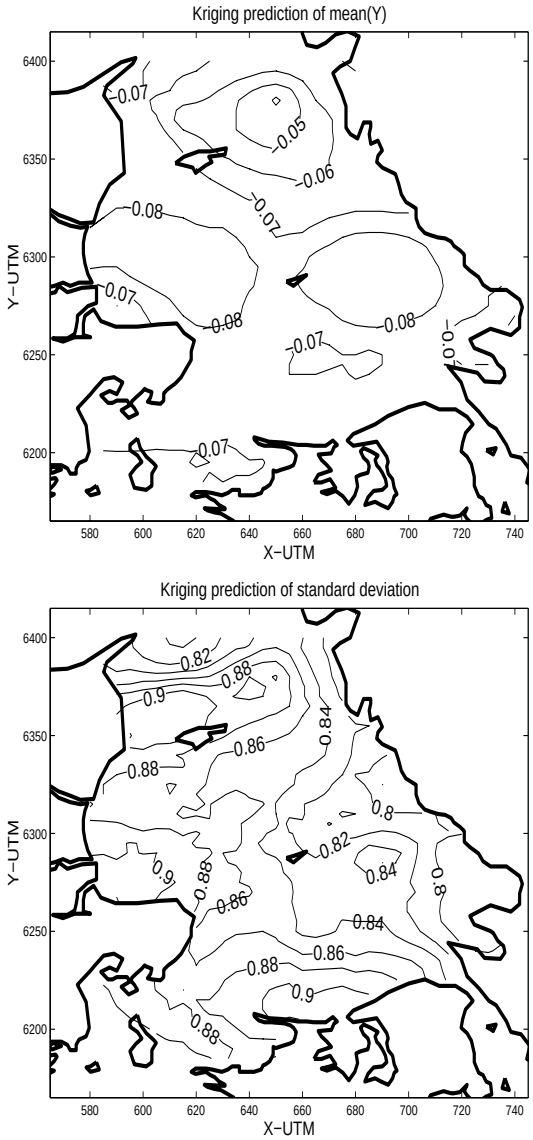

Figure 4.15: *Kriging of parameters of ARIMA models. Upper:* $\overline{Y}$*. Lower: Standard deviation.*

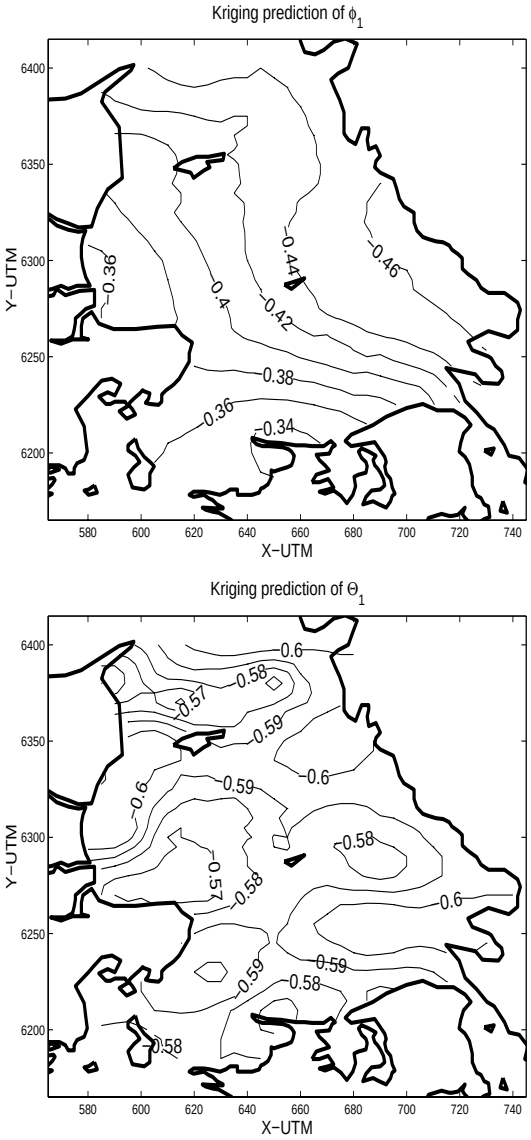

Figure 4.16: *Kriging of parameters of ARIMA models. Upper:* ϕ_1*. Lower:* Θ_1.

4.5 Other statistical methods

This section describes 4 different methods, found in the literature, which have been successfully applied to spatiotemporal data. This is done to show examples of other statistical methods which could have been applied in this thesis. When describing the different methods, different notations are used, according to the notation used by the authors.

[Geer and Zuur, 1997] apply the transfer function model to groundwater heads (output), by using the precipitation excess as input. The model is applied to individual time series. The model parameters can be seen as regionalized variables, thus kriging can be used to determine the parameters at any location in the area under consideration. It is seen that this method is very similar to the ARIMA approach, described in the previous section. It was the intension of the present author to describe concentrations of DIN by the transfer function model, using wind energy as input. However after prewhitening and filtration of the input and output series no cross correlations were found to be significant. The method might be usefull if other input variables are used, and the principle of modelling of transfer functions is therefore described in the following.

The transfer function model describes the relationship between an input, X_t, which is a stochastic variable, and an output, Y_t, which is encumbered with noise N_t. The process can be illustrated as shown in figure 4.17. The

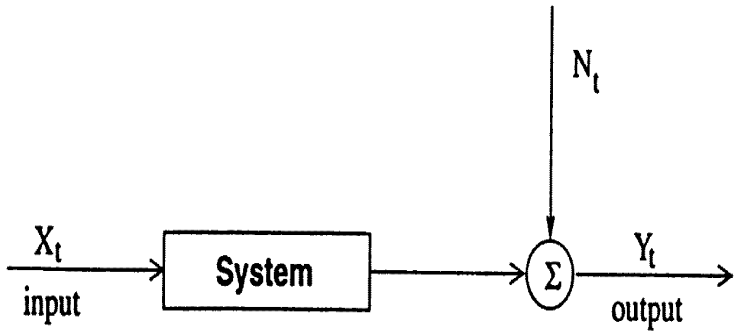

Figure 4.17: *Sketch of an input-output process.*

transfer function relates an output series to a certain input by

$$(Y_t - \overline{Y}) = \frac{\omega(B)}{\delta(B)} B^b (X_t - \overline{X}) + \frac{\theta(B)}{\varphi(B)} \epsilon(t) \tag{4.12}$$

where ϵ_t is a white noise process, b is a delay parameter, and B is the back shift operator. Furthermore;

$$\begin{aligned} \delta(B) &= 1 + \delta_1 B + \cdots + \delta_r B^r \\ \omega(B) &= \omega_0 + \omega_1 B + \cdots + \omega_s B^s \\ \varphi(B) &= 1 + \varphi_1 B + \cdots + \varphi_p B^p \\ \theta(B) &= 1 + \theta_1 B + \cdots + \theta_q B^q \end{aligned}$$

The output of the system is assumed to be stable, and independent on future values of the input series. Moreover no feedback is included, which means that the input does not depend on the output. It is assumed that the noise N_t, can be described by an ARMA model, given as

$$\begin{aligned} N_t + \varphi_1 N_{t-1} + \cdots + \varphi_p N_{t-p} &= \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} \Longrightarrow \\ \varphi(B) N_t &= \theta(B) \epsilon_t \end{aligned} \tag{4.13}$$

where ϵ_t is a white noise process, [Geer and Zuur, 1997] and [Shumway, 1988]. The impulse response function is given by

$$\begin{aligned} h(B) &= \frac{\omega(B)}{\delta(B)} B^b \\ &= h_0 + h_1 B + h_2 B^2 + \cdots \end{aligned} \tag{4.14}$$

where h_i is the impulse weight for lag i. If we use

$$N_t = \frac{\theta(B)}{\varphi(B)} \epsilon(t) \tag{4.15}$$

the transfer function can be written as

$$(Y_t - \overline{Y}) = h(B)(X_t - \overline{X}) + N_t \tag{4.16}$$

The impulse response function is estimated by using the Cross Covariance Function (CCF) as an estimate of the impulse response function, when the input series is white noise, i.e. we have to ensure that this is the case. This is done by filtering the input and output series, in a way which makes the input series a white noise process. This method is called prewhitening. Thus the estimation of the impulse response function is made in different steps, as shown in the following;

1. The input series is fitted by an ARMA model;

   $$\varphi(B)(X_t - \overline{X}) = \theta(B)\alpha_t$$

2. The input series is prewhitened;

   $$\alpha_t = \theta(B)^{-1}\varphi(B)(X_t - \overline{X})$$

3. The output series is prewhitened by the same model;

   $$\beta_t = \theta(B)^{-1}\varphi(B)(Y_t - \overline{Y})$$

4. The impulse response function, $\hat{h}_k$, is estimated by;

   $$\hat{h}_k = \frac{C_{\alpha\beta}(k)}{C_{\alpha\alpha}(0)} = \frac{C_{\alpha\beta}(k)}{S_\alpha^2}$$

By estimating the impulse response function the integer values r, s and b can be identified, and the parameters of the model estimated. Afterwards the noise, N_t, can be fitted by an ARIMA model. This section will not go into details about the estimation. For this discussion the reader is referred to [Shumway, 1988]. Some examples of the impulse response function and corresponding transfer function are shown in figure 4.18.

[Rouhani and Myers, 1990] and [Rouhani and Wackernagel, 1990] describe problems within the area of spatiotemporal analysis of data. The problems, by extending the estimation to the spatiotemporal space, are among other things caused by the fundamental difference between spatial and spatiotemporal processes. The one dimensional temporal data is ordered, with past, present and future. The two or three dimensional spatial data does usually not exhibit such an order. Further more the spatial and temporal scales are different, and can not be compared in a physical sense. A solution to this problem is to split the spatiotemporal correlation into a sum of the spatial and temporal components. Another problem of analysis of spatiotemporal data is lack of procedures for modelling of semivariograms and covariances.

[Huang and Cressie, 1996] use a spatiotemporal model to predict snow water equivalents (SWE). The Kalman filter is used to incorporate past and

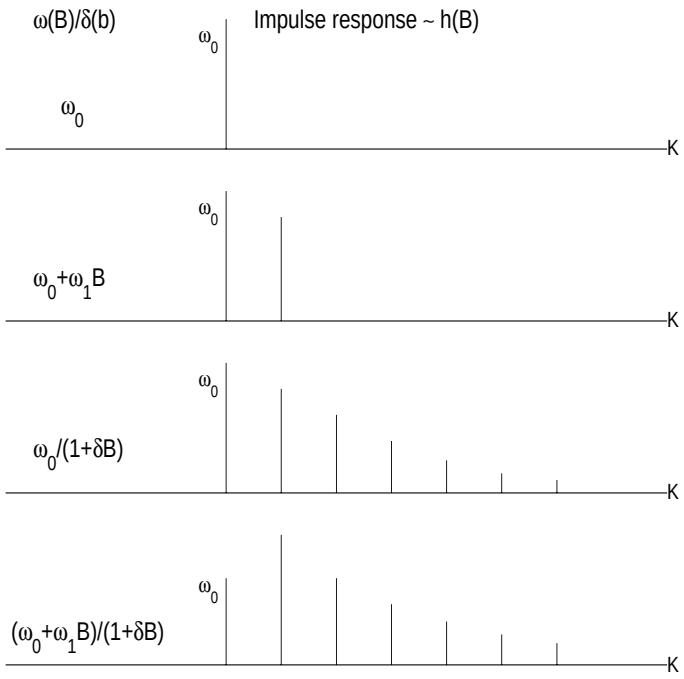

Figure 4.18: *Examples of the impulse response and transfer function.*

current data in the prediction. The model is given by

$$Z_t = S_t + \epsilon_t \tag{4.17}$$

S_t is assumed to be a temporally and spatially stationary normally distributed process with mean zero. Z_t, S_t and ϵ_t are vectors with dimension $n \times 1$, where n is the number of locations. S_t is described by an autoregressive process of order p

$$S_t(s) = \alpha_1 S_{t-1}(s) + \alpha_2 S_{t-2}(s) + \cdots + \alpha_p S_{t-p}(s) + \eta_t(s), \qquad s \in \mathcal{D} \tag{4.18}$$

The model is now given by 4.17 and 4.18, where $Z_t = (Z_t(s_1), \cdots, Z_t(s_n))^T$ is a vector containing observations of SWE for the n locations in $\mathcal{D}$ at time t, and $S_t = (S_t(s_1), \cdots, S_t(s_n))^T$ is the corresponding vector of the state variable for the n locations. ϵ_t is a white-noise process, independent of S_t, and $\eta_t(s)$ is assumed to be a temporally and spatially stationary normally distributed process with mean zero. The Kalman filter is used to update the state variable $S_t = (S_t(s_1), \cdots, S_t(s_n))^T$, when new observations are available. The optimal spatiotemporal predictor for $S_t(S_0)$ is computed for locations where measurements have been done, as well as for locations where no observations are available.

[Carroll et al., 1997] describe a spatiotemporal model, which is applied to measurements of ozone at eleven different stations, where the time series contain a substantial number of missing observations. The model is given by

$$Y(x,t) = g(t) + \epsilon(x,t) \tag{4.19}$$

where $g(t)$ is a deterministic function while $\epsilon(x,t)$ is a random function, which takes into account both the spatial and temporal variation. $g(t)$ is in this case modelled by

$$g(t) = \alpha_{hour} + \beta_{month} + \gamma_1 temp(t) + \gamma_2 temp^2(t) \tag{4.20}$$

where $temp$ is the temperature at time t, α_{hour} accounts for the overall hourly level and β_{month} for the overall monthly level of ozone. The random function $\epsilon(x,t)$ is modelled by the spatiotemporal covariance function, given by

$$Cov(\epsilon(x_1,t_1), \epsilon(x_2,t_2)) = \sigma^2 \rho(d,v) \tag{4.21}$$

where d is the distance between location x_1 and x_2, and $v = |t_2 - t_1|$ is the time lag between two times. The correlation function in (4.21) is given by

$$\rho(d,v) = \begin{cases} 1 & \text{if } d = v = 0 \\ \phi_v^d \psi_v & \text{otherwise} \end{cases} \tag{4.22}$$

where

$$\log(\psi_v) = a_0 + a_1 v + a_2 v^2 \quad (4.23)$$
$$\log(\phi_v) = b_0 + b_1 v + b_2 v^2 \quad (4.24)$$

[Meiring et al., 1998] use an approach which is very similar to the one described by [Carroll et al., 1997], but [Meiring et al., 1998] incorporate the spatial variation in the function describing the trend, i.e. the function which corresponds to $g(t)$.

[Haas, 1995] and [Haas, 1998] describe and apply a spatiotemporal model for analyzing wet sulfate deposition data, which the authors call **MCSTK**[1]. The spatiotemporal locations are given by $\boldsymbol{X} = (x, y, t)$. The prediction is done locally, e.g. for a point $\boldsymbol{X_0} = (x_0, y_0, t_0)$, which means that only the observations which are closest to the point $\boldsymbol{X_0}$ in space and time, are used in the prediction. The neighborhood is defined in the following way

1 $t_{earliest}$ and t_{latest} is the time of the earliest and latest observation in the dataset. The temporal interval of the neighborhood is given by $m_T = t_u - t_L$, where the upper limit t_u is $\{\min(t_{latest}, t_0 + m_T/2)\}$, and the lower limit is $\{\max(t_{earliest}, t_u - m_T)\}$. This selection results in n_I observations within the temporal interval.

2 The n_I observations are sorted according to their spatial distance from (x_0, y_0). After this initial sorting, observations from a particular location are sorted according to their temporal distance from t_0. The two sortings result in a list of sorted observations numbered from 1 to n_I.

3 The observation set of the neighborhood is then defined as the first n_c observations in the sorted list.

The temporal interval m_T is chosen so that second order stationarity along the temporal dimension within the neighborhood holds. Within the neighborhood a random variable of the observed spatiotemporal process is modelled by

$$Y_c(\boldsymbol{X}) = \mu(\boldsymbol{X}, \boldsymbol{\beta_c}) + \psi(\mu(\boldsymbol{X}, \boldsymbol{\beta_c}), \boldsymbol{X}) R_c(\boldsymbol{X}) \quad (4.25)$$

$\mu(\boldsymbol{X}, \boldsymbol{\beta_c})$ is the spatiotemporal drift parameterized by components of the vector $\boldsymbol{\beta_c}$. $R_c(\boldsymbol{X})$ is the spatiotemporal residual process, which is second

[1] Moving Cylinder SpatioTemporal Kriging

order stationary within the neighborhood, and $\psi(\mu(\boldsymbol{X}, \boldsymbol{\beta_c}), \boldsymbol{X})$ is a model of heteroscedasticity. $\psi(\cdot)$ is estimated only at the prediction location, $\boldsymbol{X_0}$, with an nonparametric estimator. Consequently no parametric form of $\psi(\cdot)$ is modelled. For the wet sulfate deposition data the model is given by

$$\begin{aligned} Y_c(\boldsymbol{X}) &= (\beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 y^2 + \beta_5 xy)^{-1} \\ &+ \beta_6 t + \beta_{j+6}\psi(\mu(\boldsymbol{X}, \boldsymbol{\beta_c}), t) R_c(\boldsymbol{X}) \quad j = 1, \cdots, 4 \end{aligned} \tag{4.26}$$

The covariance function within the neighborhood is given by

$$Cov[R_c(\boldsymbol{X_1}), R_c(\boldsymbol{X_2})] = C_{S,T}(g((x_1, y_1)', (x_2, y_2)'), h(t_1, t_2)) \tag{4.27}$$

The function g is the spatial lag, while h is the temporal lag. The spatiotemporal semivariogram is related to the covariance function in (4.27) in the following way

$$\gamma_{S,T}(g, h) = \frac{1}{2} var[R_c(\boldsymbol{X_1}) - R_c(\boldsymbol{X_2})] = C_{S,T}(0, 0) - C_{S,T}(g, h) \tag{4.28}$$

The product of the spatial, $C_S(\cdot)$, and temporal, $C_T(\cdot)$, covariance functions is a separable spatiotemporal covariance function with associated spatiotemporal semivariogram

$$\gamma_{S,T}(g, h) = C_S(0)\gamma_T(h) + \gamma_S(g)C_T(0) - \gamma_S(g)\gamma_T(h) \tag{4.29}$$

4.6 Summary of spatiotemporal data analysis

In this chapter three different approaches for reconstruction of concentrations of DIN and DIP have been tested. The methods combine the temporal and spatial dimensions, i.e. they are 3 dimensional.

The main problem of using such methods, is to find the optimal relationship between the correlation of data in the different dimensions. In three dimensional ordinary kriging this is dealt with by computing semivariograms in different directions, and use these to include anisotropy in the model. The results obtained by applying the method to DIP seem to be good, and the maps of the weekly distribution of this parameter are very similar to those computed in 2 dimensions, based on data temporally reconstructed by the General Linear Model. However for DIN the 3 dimensional kriging approach is not able to compute predictions in winter and spring in an area by the Swedish coast, and the results obtained for this variable do not seem to be rational from a physical point of view.

Other approaches, which have been attempted, is locally weighted regression in 3 dimensions and modelling of the transfer function. Locally weighted regression did not result in a reasonable reconstruction, while the ARIMA model seems to give good results. This method has been applied to DIN at different stations, and kriging is used to determine the parameters of the model at any location in Kattegat. The General Linear Model is used to reconstruct time series of DIN. This is done in order to obtain a resolution of one week, and means that the method cannot be applied to raw data, like the other 3 dimensional methods, and it seems to be the main disadvantage of the method.

In the end of the chapter other spatiotemporal methods, found in the literature, are presented. These have been successfully applied to other sets of data measured in both time and space.

Chapter 5
Discussion

This chapter discusses the different methods which have been applied in this thesis. The discussion is based on the Goodness Of Model values which are computed by cross validation, but also on the fact that, from the measurements, we have some physical knowledge of the level and spatial distribution of concentrations of DIN and DIP in Kattegat. The spatial distribution has been determined by both 2 -and 3 dimensional statistical methods.

Data for the 2 dimensional methods are temporally reconstructed by GLM or LOESS, when no measurement has been carried out. For weeks where measurements are available, these are used instead of the reconstructed values. Both GLM and LOESS are used for temporal reconstruction in order to examine whether or not the 24 stations with the highest sampling frequencies are sufficient to compute weekly maps of the spatial distribution of DIN and DIP.

From the cross validation of the 2 dimensional methods it is found that temporal reconstruction using LOESS in general results in lower values of the Goodness Of Model, i.e. from this point of view it seems to be better to use LOESS prior to the computation of the spatial distribution of DIN and DIP. However, from a computational and physical point of view, some disadvantages have been found:

- When ordinary kriging is used to compute the spatial distribution of DIN and DIP, too high concentrations of the species are found in the open sea.

- Universal kriging results in a dramatic overestimation of the concentration of DIN and DIP in coastal areas.

- Data temporally reconstructed by LOESS cannot be used to compute the spatial distribution of DIN and DIP by cokriging.

These disadvantages are explained by the lack of observations when LOESS is used for temporal reconstruction of data. It could also be due to the locations of the 24 stations with the highest sampling frequencies. These locations are shown in figure 5.1. It is seen that a lot of these stations are located in coastal areas, and that only three stations, marked with crosses ($\times$), are located in a large area in the eastern part of Kattegat. The stations which are located very close to the coast, e.g. the 5 stations by the coast of Zealand, are not very useful. For this purpose it would have been an advantage if they were located in the open sea.

The above mentioned disadvantages of using LOESS for temporal reconstruction of data, are not only due to lack of observations, but also caused by the skewed geographical distribution of stations with a high sampling frequency. This also explains the high values of the ranges found when modelling the semivariogram for data temporally reconstructed by LOESS. This set of data is highly dominated by coastal stations, which are expected to have very similar levels of concentrations of DIN and DIP. A large fraction of the coastal stations are located far from each other, e.g. stations by the coast of Zealand and stations in the northern part of Kattegat. These stations far from each other, with similar levels of concentrations, lead to the high values of the semivariance for high lag numbers, and consequently leading to large ranges of the semivariograms.

With the present locations of these stations, it means that we have to use a method for temporal reconstruction of data, which uses the information from surrounding stations, i.e. a method which can be used to estimate the time series for stations where only a few measurements have been carried out, in order to compute weekly maps of the spatial distribution of DIN and DIP.

Figure 5.1: *Location of stations with high sampling frequencies. LOESS for temporal reconstruction can be applied to these stations.*

The General Linear Model is such a method, which is able to compute estimations for all weeks and stations where measurements have been carried out. The major disadvantage of the model is that it does not include any temporal or spatial correlation. Furthermore it tends to overfit, i.e. both the week -and station effect can be influenced dramatically by extreme observations, shown as peaks on the reconstructed time series. It might improve the temporal reconstruction if the week effect is computed, not only by observations from a single week, but also by observations from surrounding weeks. Similarly the station effect should be computed as a function of surrounding stations.

From the above discussion it is seen that, if we want to compute weekly maps of the spatial distribution of DIN and DIP, a method for temporal reconstruction of data, which includes information from surrounding stations, should be used prior to the computation of spatial predictions. When the General Linear Model is used, the three different kriging methods, which have been applied, give results that are very similar, and understandable from a physical point of view. The values of the Goodness Of Model can therefore be used to find the most suitable method. From this it is seen that cokriging, with the depth of water as secondary variable, should be used for computing the weekly spatial distribution of DIN and DIP.

It has also been attempted to model the spatial distribution of DIN and DIP by 3 dimensional methods, which include time in the prediction. Three dimensional kriging seems to work very well, from a statistical and physical point of view, when applied to DIP, while it does not give rational results, from a physical and computational point of view, when applied to DIN. The main problem when using these kinds of models, is to find the optimal relationship between the spatial correlations in the different dimensions. Moreover, when modelling the 3 dimensional semivariogram, the distance between observations are difficult to interpret, because it includes a time component. In 2 dimensions this distance is simply given in kilometers. Methods for reconstruction of data in three dimensions are not widely used, and consequently only a sparse amount of literature about the subject exists.

In this thesis maps of the spatial distribution of DIN and DIP have been shown for 4 different weeks representing the 4 seasons of the year. Such

maps could be computed for any of the 260 weeks in the five year period, and presented as a "movie". Such a "movie" could be used to identify major differences between weeks.

Chapter 6
Conclusion

This thesis describes, applies and compares statistical methods for reconstruction of measured concentrations of Dissolved Inorganic Nitrogen (DIN) and Dissolved Inorganic Phosphorus (DIP) in Kattegat. The methods are general, thus the application, for reconstruction of these two species, can be seen as examples. They could have been applied to other parameters. Two types of methods are considered. These are:

- Methods, where the reconstruction is computed in two steps. Firstly a method is used to fill out gaps in time series by estimated values, and these values are used to compute weekly maps of the spatial distribution of the variables.

- Three dimensional methods, where the reconstruction is computed in one step, by including all three dimensions, i.e. both time and the two spatial dimensions.

In order to obtain a weekly frequency, the General Linear Model and locally weighted regression have been used for reconstruction of time series. The General Linear Model uses the information from the surrounding stations, and it can be used for temporal reconstruction at 65 different stations. On the other hand locally weighted regression only uses data from the station under consideration, and it can therefore only be applied to the 24 stations with the highest sampling frequencies.

Based on the estimations, computed by the two methods, weekly maps of the spatial distribution of DIN and DIP have been computed. This is done by using three different variants of kriging, i.e. ordinary kriging, universal kriging and cokriging, and the results are compared from a statistical and a physical point of view. This showed that the spatial predictions of DIN and DIP should be computed by cokriging, based on data temporally reconstructed by the General Linear Model.

Locally weighted regression cannot be used, when the aim is to compute weekly maps of the spatial distribution. This is due to lack of observations, as well as the skewed geographical distribution of stations with a high sampling frequency.

It has also been attempted to reconstruct in time and space using 3 dimensional methods, referred to as spatiotemporal methods. Three dimensional kriging can be used to reconstruct DIP, as a good alternative to the General Linear Model together with cokriging. On the other hand 3 dimensional kriging of DIN did not give good results.

Chapter 7
Future work

This chapter shortly describes how some of the statistical methods used in this thesis can be used for design of monitoring programs.

A lot of programs for monitoring of environmental data are characterized by a huge number of measurements. These measurements do not neccesarily result in the information, which is the aim of the monitoring. One reason is that calculations of the amount of data, which is neccesary to obtain the desired information, have not been done.

Such calculations could be based on the methods presented in this thesis, e.g.

- The General Linear Model, as described in this thesis, or a more advanced version of it, as suggested in the discussion, can be used to determine the sampling frequency, which is neccesary to compute estimates with a desired certainty. Cross validation can be used to examine the gain of additional sampling.

- Kriging can be used to determine the optimal locations of stations in Kattegat, as well as the number of stations, neccesary for obtaining a given certainty. Cross validation can be used to examine the gain of including additional stations in the monitoring program.

- Gaussian sequential conditional simulation seems to be a good method for this purpose, e.g. a measure of the total variation can be computed from a given number of stations.

The information and knowledge obtained, from a study of the use of statistical methods for design of monitoring programs, can be used both by Danish institutions and by the many countries, which are about to start up environmental monitoring programs. Furthermore, the information can be applied within other areas than the marine, which has been the topic of this thesis, e.g. monitoring of pollution of air.

Appendix A
Software and programming

This appendix is a list of the software that has been used for the computations in this thesis. The programs are not shown, but people who want to use these, can get them by contacting the author. Temporal reconstruction of time series is computed using SAS, all geostatistics, except cross validation of cokriging, has been done by GSLIB[1], which is a free package of Fortran-programs. Executable files are available at www.gslib.com. Plotting of maps and time series has been done in Matlab and S-plus.

[1] Geostatistical Software Library

- SAS-macro for optimization of the bandwidth in 1 dimensional LOESS.
- SAS-macro for temporal reconstrucion by LOESS.
- SAS-macro for temporal reconstrucion by GLM.
- SAS-program for cross validation of GLM.
- SAS-macro for computing locally weighted regression in 3 dimensions.
- SAS-program for computing ordinary kriging predictions.
- SAS-program for estimation of semivariograms.
- SAS-macro for modelling of transfer functions.
- SAS-macro for identification and estimation of ARIMA models.
- Various Matlab-programs for plotting of semivariograms and contour-maps.
- Matlab-program for optimization of parameters in semivariogram models.
- Matlab-program for cross validation of cokriging.
- S-plus function for optimization of the bandwidth in 3 dimensional LOESS.
- S-plus function for non-linear optimization of parameters in semivariogram models.
- Various parameter files for GSLIB.

Appendix B
Bandwidth for 1 dimensional loess

Bandwidth for 1 dimensional loess determined from Akaike's Information Criteria (AIC). The computation has been done for log-transformed DIN and DIP.

Station	log(DIN)	log(DIP)
1001	0.28	0.23
1004	0.25	0.19
1007	0.24	0.24
1008	0.26	0.25
1009	0.25	0.25
190004	0.24	0.12
1937	0.26	0.27
1939	0.24	0.27
1993	0.21	0.30
20004	0.24	0.21
3302	0.24	0.24
3310	0.31	0.21
403	0.22	0.21
409	0.19	0.20
413	0.12	0.12
415	0.23	0.27
418	0.28	0.22
4402	0.26	0.19
4410	0.21	0.17
905	0.21	0.15
921	0.22	0.21
922	0.29	0.27
925	0.26	0.28
HALVAD	0.38	0.33

Table B.1: *Bandwidth for 1 dimensional loess.*

Appendix C
Semivariogram surface

To determine the directions of anisotropy, maps of the semivariogram surface have been computed. In this appendix these are shown for log-transformed DIN and DIP, for a week in the middle of March 1995, a week in the middle of October 1996 and a week in the middle of July 1995. The corresponding maps for a week in the middle of January 1994 are shown in the text.

C.1 Log-transformed DIN

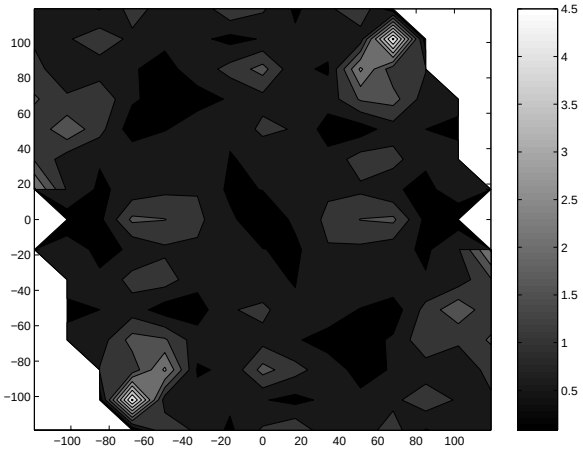

Figure C.1: *Maps of the semivariogram surface of log-transformed DIN for a week in the middle of October 1996.*

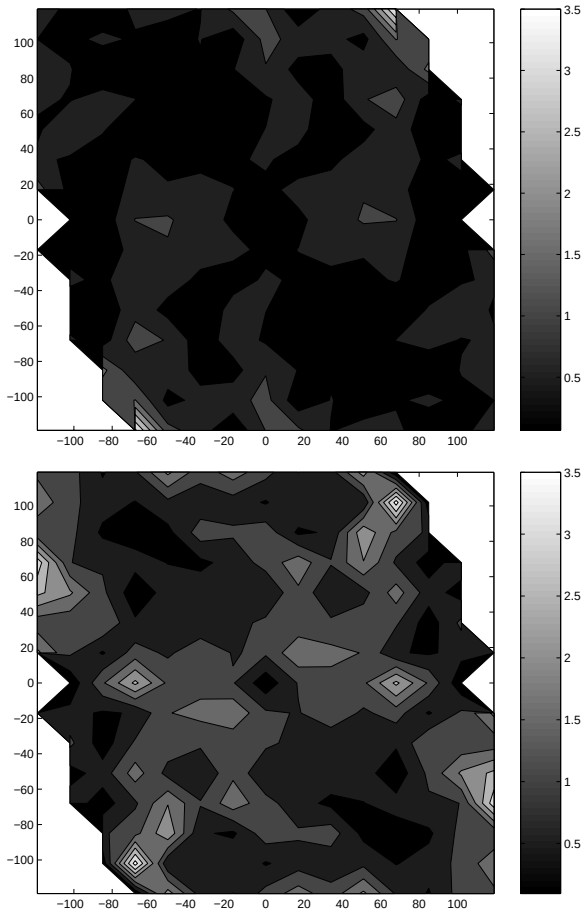

Figure C.2: *Maps of the semivariogram surface of log-transformed DIN. Upper: A week in the middle of March 1995. Lower: A week in the middle of July 1995.*

C.2 Log-transformed DIP

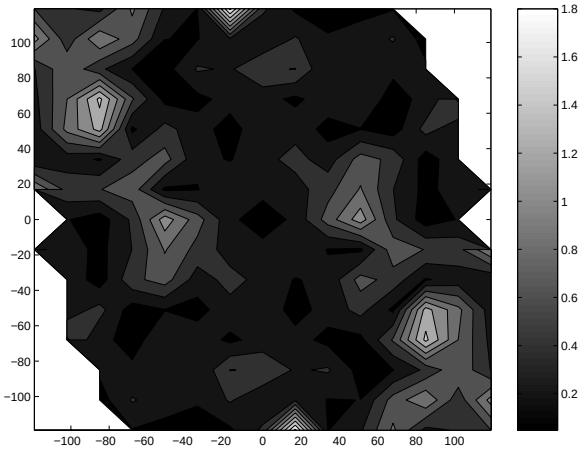

Figure C.3: *Maps of the semivariogram surface of log-transformed DIP for a week in the middle of October 1996.*

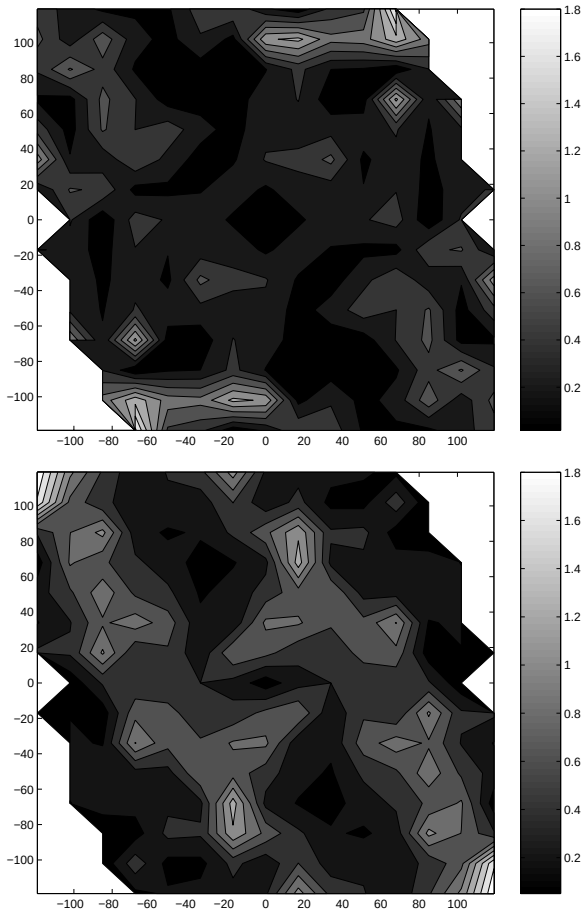

Figure C.4: *Maps of the semivariogram surface of log-transformed DIP. Upper: A week in the middle of March 1995. Lower: A week in the middle of July 1995.*

Appendix D
The system of ordinary kriging equations

This appendix shows in greater details than in the text how the system of ordinary kriging equations is calculated.

The goal is to minimize the error of the estimation, i.e.

$$\text{Minimize} \quad E[(\hat{Z}(s_0) - Z(s_0))^2] \tag{D.1}$$

To obtain an unbiased estimate, the sum of the weights has to be 1, i.e.

$$\sum_i \lambda_i = 1 \tag{D.2}$$

The expression in (D.1) can be rewritten

$$\begin{aligned} E[(\hat{Z}(s_0) - Z(s_0))^2] &= E[(\sum_i \lambda_i Z(s_i) - Z(s_0))^2] \\ &= E[(\sum_i \lambda_i Z(s_i) - \sum_i \lambda_i Z(s_0))^2] \\ &= E[(\sum_i \lambda_i (Z(s_i) - Z(s_0)))^2] \\ &= E[\sum_i \lambda_i (Z(s_i) - Z(s_0)) \sum_j \lambda_j (Z(s_j) - Z(s_0))] \\ &= \sum_i \sum_j \lambda_i \lambda_j E[(Z(s_i) - Z(s_0))(Z(s_j) - Z(s_0))] \end{aligned} \tag{D.3}$$

From the definition of the semivariogram it is known that

$$\begin{aligned} \gamma(s_i - s_j) &= \frac{1}{2} E[(Z(s_i) - Z(s_j))^2] \\ &= \frac{1}{2} E[((Z(s_i) - Z(s_0)) - (Z(s_j) - Z(s_0)))^2] \\ &= \frac{1}{2} E[(Z(s_i) - Z(s_0))^2] + \frac{1}{2} E[(Z(s_j) - Z(s_0))^2] \\ &\quad -E[(Z(s_i) - Z(s_0))(Z(s_j) - Z(s_0))] \\ &= \gamma(s_i - s_0) + \gamma(s_j - s_0) \\ &\quad -E[(Z(s_i) - Z(s_0))(Z(s_j) - Z(s_0))] \end{aligned} \tag{D.4}$$

By combining (D.3) and (D.4) we get

$$\begin{aligned} E[(\hat{Z}(s_0) - Z(s_0))^2] &= -\sum_i \sum_j \lambda_i \lambda_j \gamma(s_i - s_j) + \sum_i \sum_j \lambda_i \lambda_j \gamma(s_i - s_0) \\ &\quad + \sum_i \sum_j \lambda_i \lambda_j \gamma(s_j - s_0) \\ &= \sum_i \sum_j \lambda_i \lambda_j \gamma(s_i - s_j) + 2 \sum_i \lambda_i \gamma(s_i - s_0) \end{aligned} \tag{D.5}$$

The last equal-sign is true according to (D.2) and because

$$\sum_i \lambda_i \gamma(s_i - s_0) = \sum_j \lambda_j \gamma(s_j - s_0) \tag{D.6}$$

To minimize equation (D.5) with the constraint (D.2) a Lagrange multiplier m is introduced, and the expression that will be minimized becomes

$$\frac{1}{2} E[(\hat{Z}(s_0) - Z(s_0))^2] - m[\sum_i \lambda_i - 1] \tag{D.7}$$

The minimization is done by calculating the partial derivatives with respect to m and λ_i, and we get

$$\begin{aligned} \sum_j \lambda_j \gamma(s_i - s_j) + m &= \gamma(s_i - s_0) \\ \sum_i \lambda_i &= 1 \end{aligned} \tag{D.8}$$

Appendix E
Spatial distribution using ordinary kriging

This appendix shows the spatial distribution of DIN and DIP, for a week in the middle of March 1995 and a week in the middle of October 1996, when ordinary kriging is applied to data, which are temporally reconstructed by GLM or LOESS. The corresponding results for a week in the middle of January 1994 and a week in the middle of July 1995 are shown in the text.

E.1 Dissolved Inorganic Nitrogen

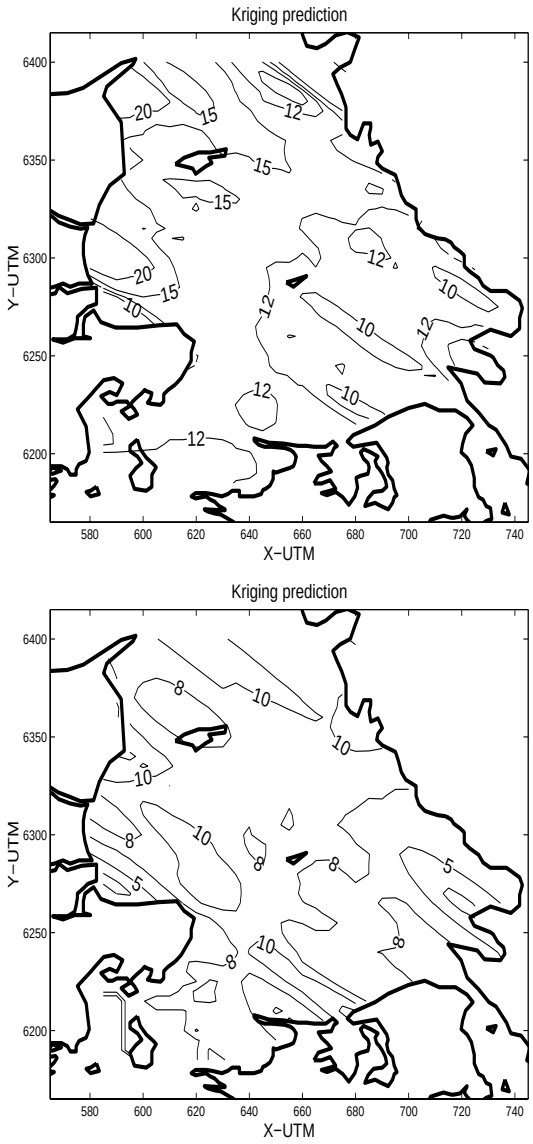

Figure E.1: *Spatial distribution of DIN for a week in the middle of March 1995 computed by ordinary kriging. Upper: Temporal reconstruction by GLM. Lower: Temporal reconstruction by LOESS.*

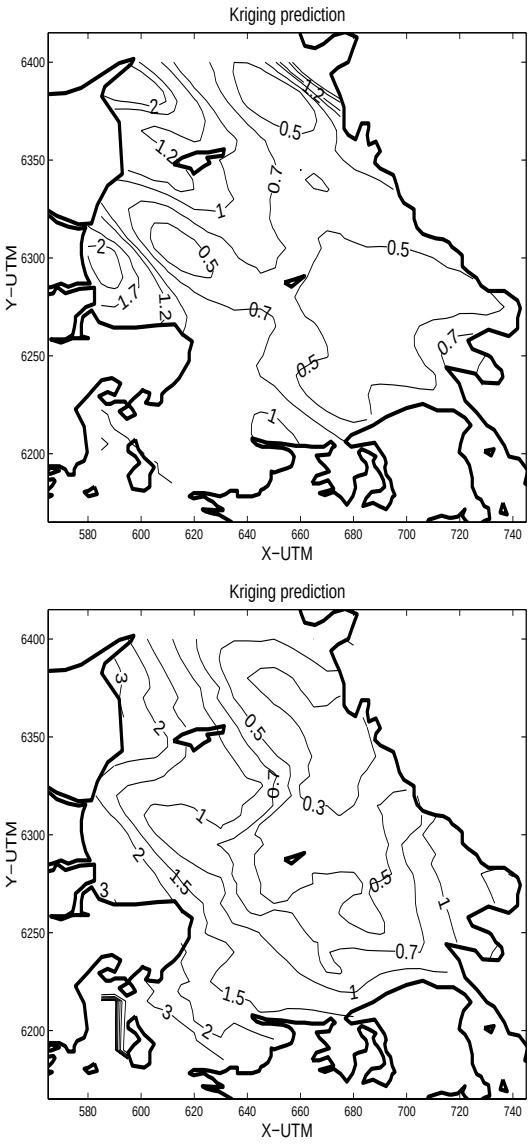

Figure E.2: *Spatial distribution of DIN for a week in the middle of October 1996 computed by ordinary kriging. Upper: Temporal reconstruction by GLM. Lower: Temporal reconstruction by LOESS.*

E.2 Dissolved Inorganic Phosphorus

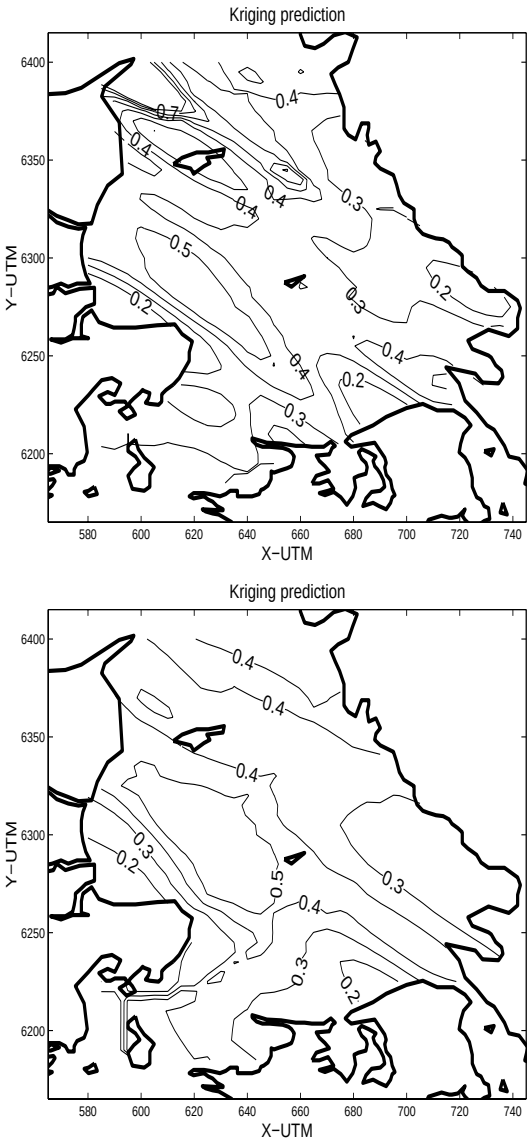

Figure E.3: *Spatial distribution of DIP for a week in the middle of March 1995 computed by ordinary kriging. Upper: Temporal reconstruction by GLM. Lower: Temporal reconstruction by LOESS.*

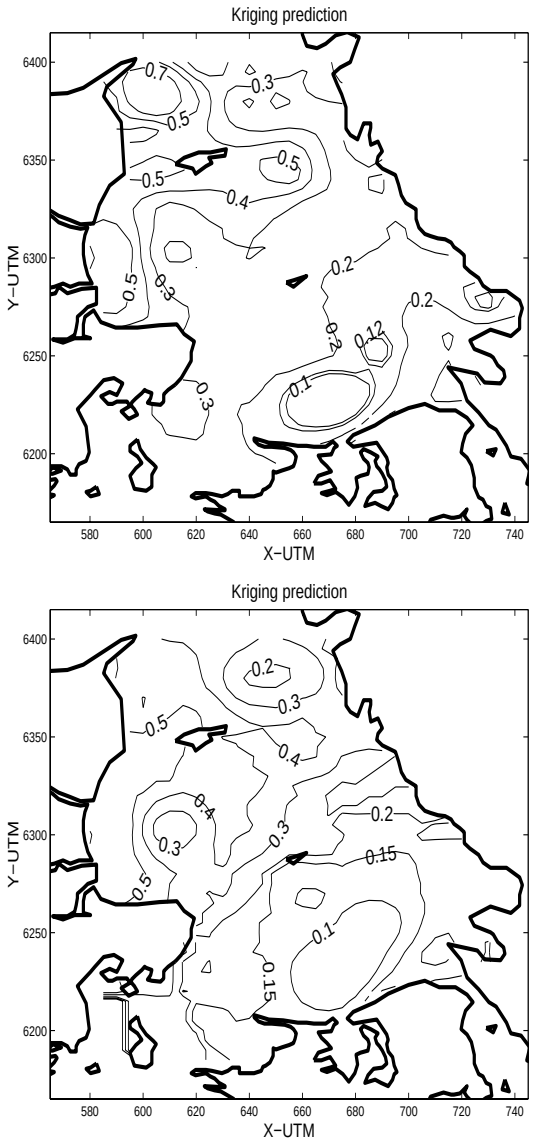

Figure E.4: *Spatial distribution of DIP for a week in the middle of October 1996 computed by ordinary kriging. Upper: Temporal reconstruction by GLM. Lower: Temporal reconstruction by LOESS.*

Appendix F
Spatial distribution using universal kriging

This appendix shows the spatial distribution of DIN and DIP, for a week in the middle of March 1995 and a week in the middle of October 1996, when universal kriging is applied to data, which are temporally reconstructed by GLM or LOESS. The corresponding results for a week in the middle of January 1994 and a week in the middle of July 1995 are shown in the text.

F.1 Dissolved Inorganic Nitrogen

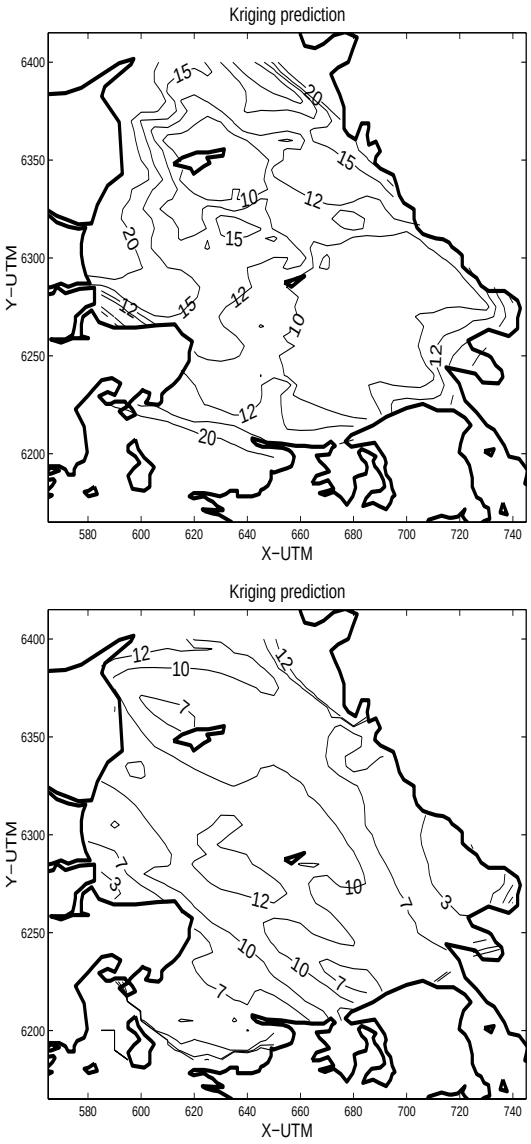

Figure F.1: *Spatial distribution of DIN for a week in the middle of March 1995 computed by universal kriging. Upper: Temporal reconstruction by GLM. Lower: Temporal reconstruction by LOESS.*

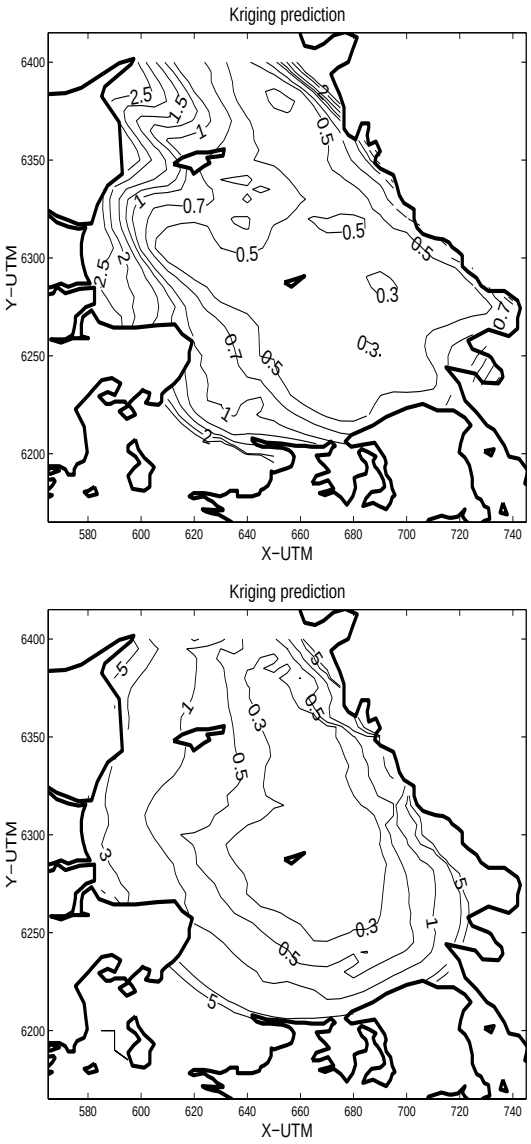

Figure F.2: *Spatial distribution of DIN for a week in the middle of October 1996 computed by universal kriging. Upper: Temporal reconstruction by GLM. Lower: Temporal reconstruction by LOESS.*

F.2 Dissolved Inorganic Phosphorus

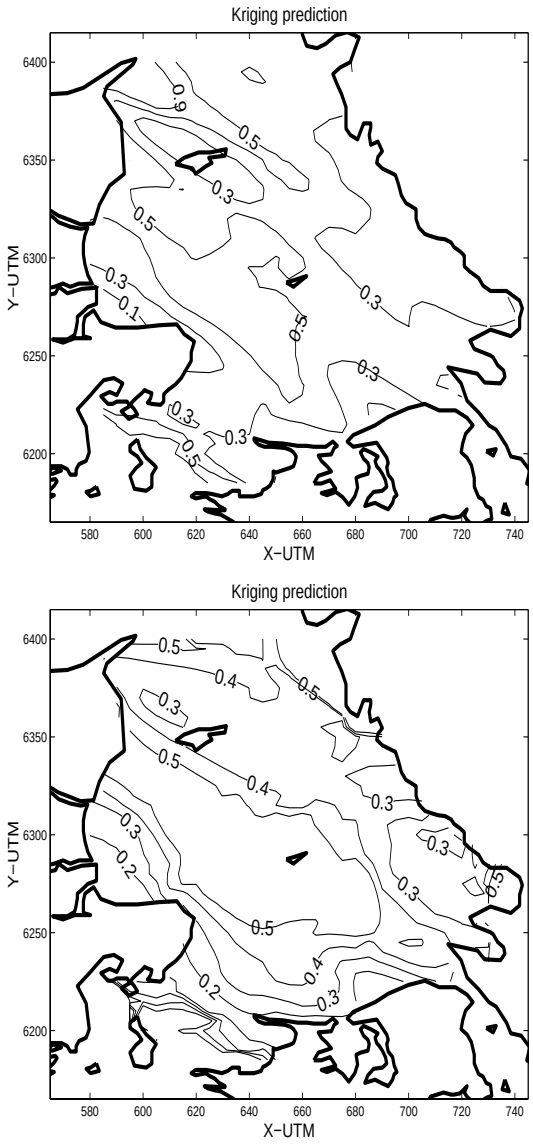

Figure F.3: *Spatial distribution of DIP for a week in the middle of March 1995 computed by universal kriging. Upper: Temporal reconstruction by GLM. Lower: Temporal reconstruction by LOESS.*

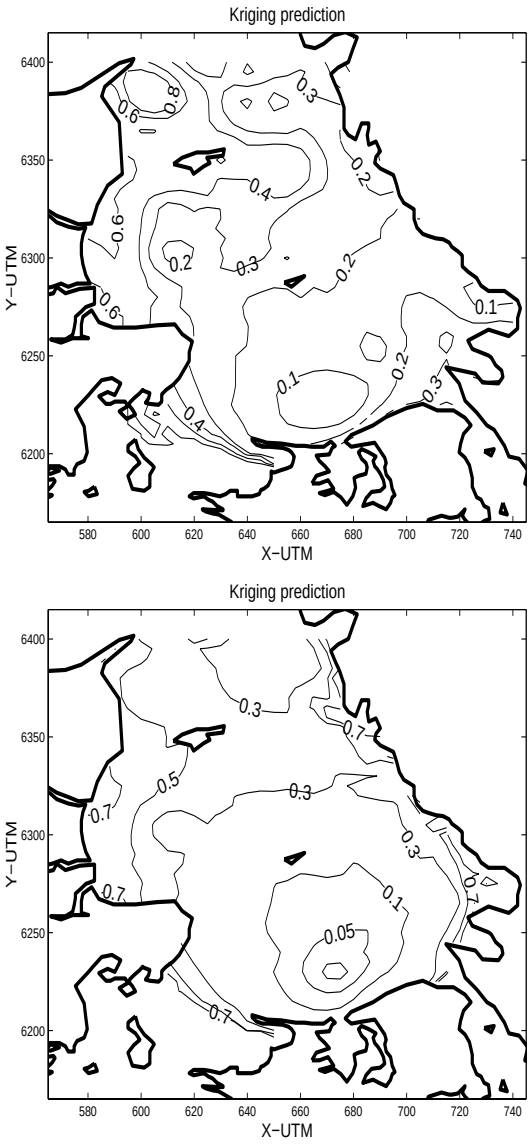

Figure F.4: *Spatial distribution of DIP for a week in the middle of October 1996 computed by universal kriging. Upper: Temporal reconstruction by GLM. Lower: Temporal reconstruction by LOESS.*

Appendix G
Spatial distribution using cokriging

This appendix shows the spatial distribution of DIN and DIP, for a week in the middle of March 1995 and a week in the middle of October 1996, when cokriging is applied to data, which are temporally reconstructed by GLM. The corresponding results for a week in the middle of January 1994 and a week in the middle of July 1995 are shown in the text.

G.1 Dissolved Inorganic Nitrogen

G.2 Dissolved Inorganic Phosphorus

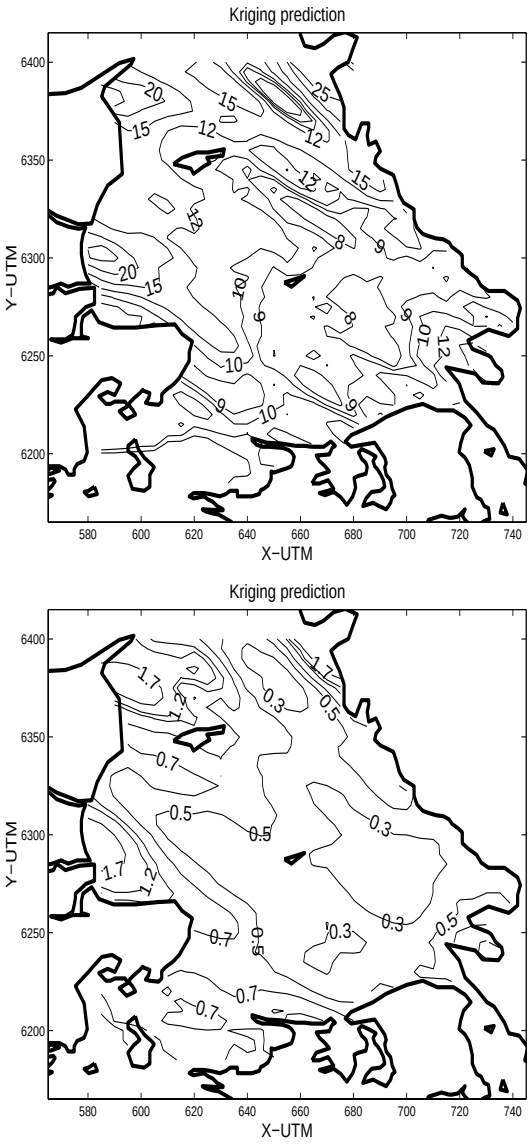

Figure G.1: *Spatial distribution of DIN computed by cokriging of data temporally reconstructed by GLM. Upper: A week in the middle of March 1995. Lower: A week in the middle of October 1996.*

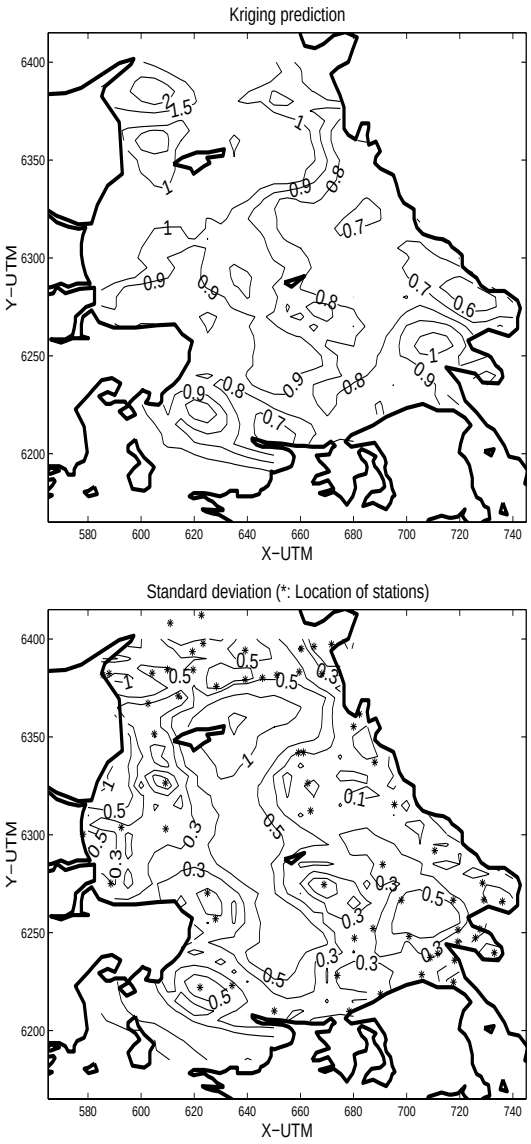

Figure G.2: *Spatial distribution of DIP computed by cokriging of data temporally reconstructed by GLM. Upper: A week in the middle of March 1995. Lower: A week in the middle of October 1996.*

Appendix H
Sequential conditional simulation

This appendix shows the results of applying Gaussian sequential conditional simulation to DIP in wintertime, i.e. week 1776. Data for the computation are measured or temporally reconstructed by GLM. The figures presented in the appendix correspond to those in section 3.6.1, and the reader is therefore referred to this section for an explanation of the figures. The use of the method aim at exemplifying how knowledge of uncertainty of prediction at a given location in Kattegat can be obtained. This is done by computing 100 realizations, and calculating the coefficient of variance (CV), given as the standard deviation divided by the mean, for two different locations, which are shown in figure 3.36. The coefficients of variance for the two locations are:

Southern location: CV=0.39/0.84=0.46
Northern location: CV=0.36/0.83=0.43

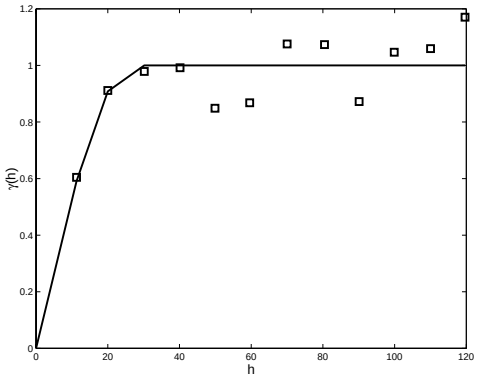

Figure H.1: *Semivariogram for log-transformed normal score DIP data.*

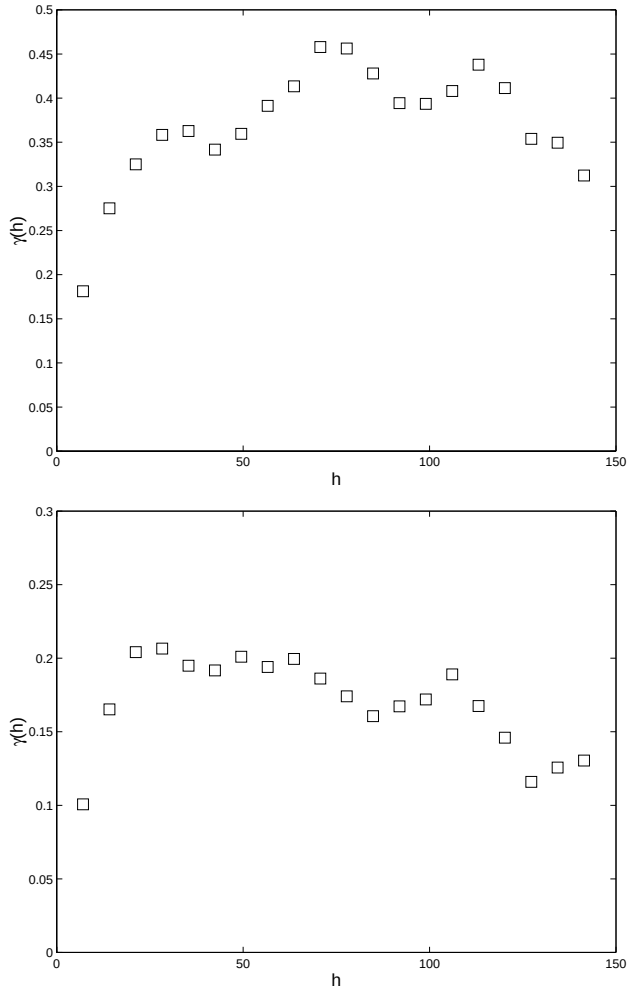

Figure H.2: *Semivariogram of simulated log-transformed DIP data for two different realizations.*

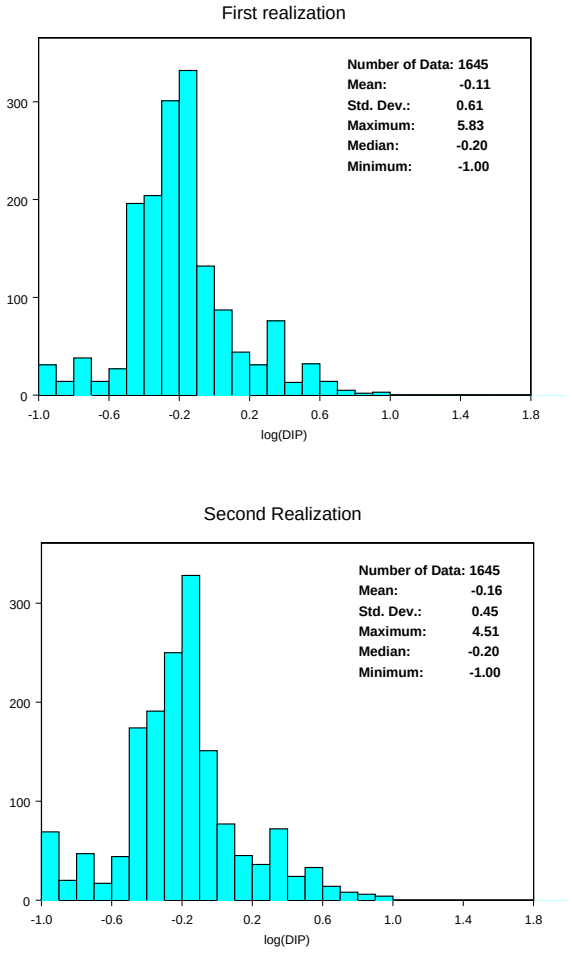

Figure H.3: *Histogram of simulated log-transformed DIP data for two different realizations.*

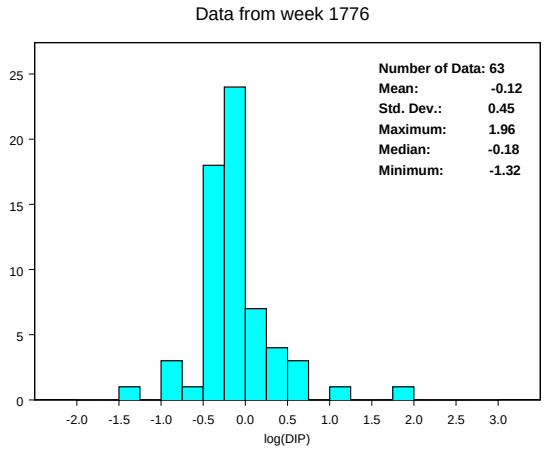

Figure H.4: *Histogram of log-transformed observations of DIP. Data are measured or reconstructed by GLM.*

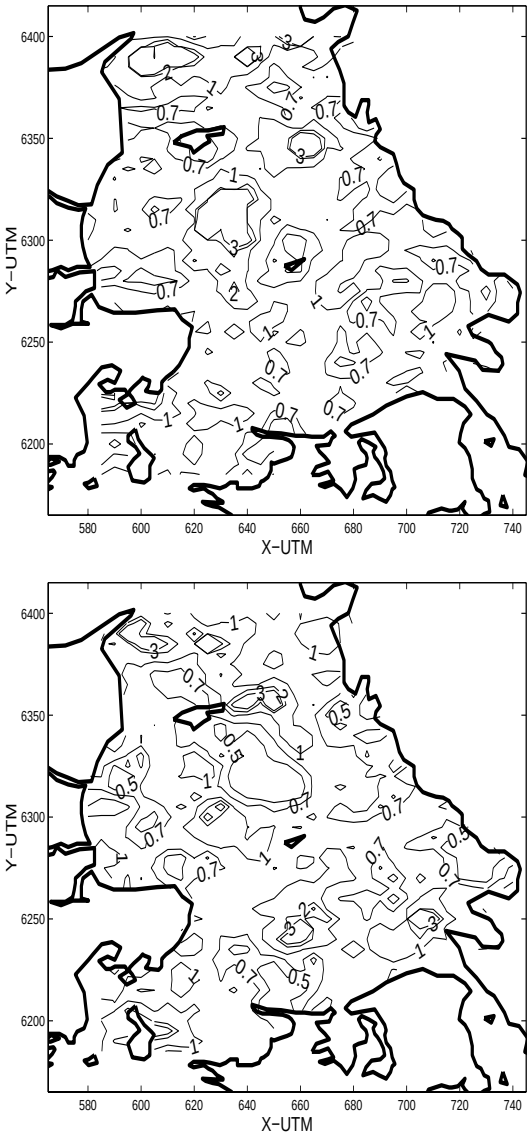

Figure H.5: *Mapping of two different realizations of DIP.*

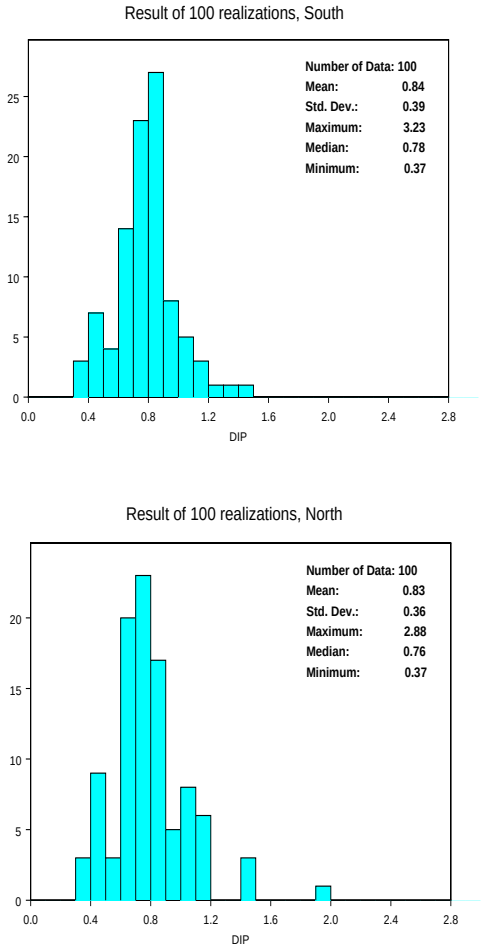

Figure H.6: *Histograms of DIP for 100 realizations. Two different locations in Kattegat.*

Appendix I
Kriging in three dimensions

This appendix shows the spatial distribution of DIN and DIP, for a week in the middle of March 1995 and a week in the middle of October 1996, when 3 dimensional ordinary kriging is applied to data. The corresponding results for a week in the middle of January 1994 and a week in the middle of July 1995 are shown in the text.

I.1 Dissolved Inorganic Nitrogen

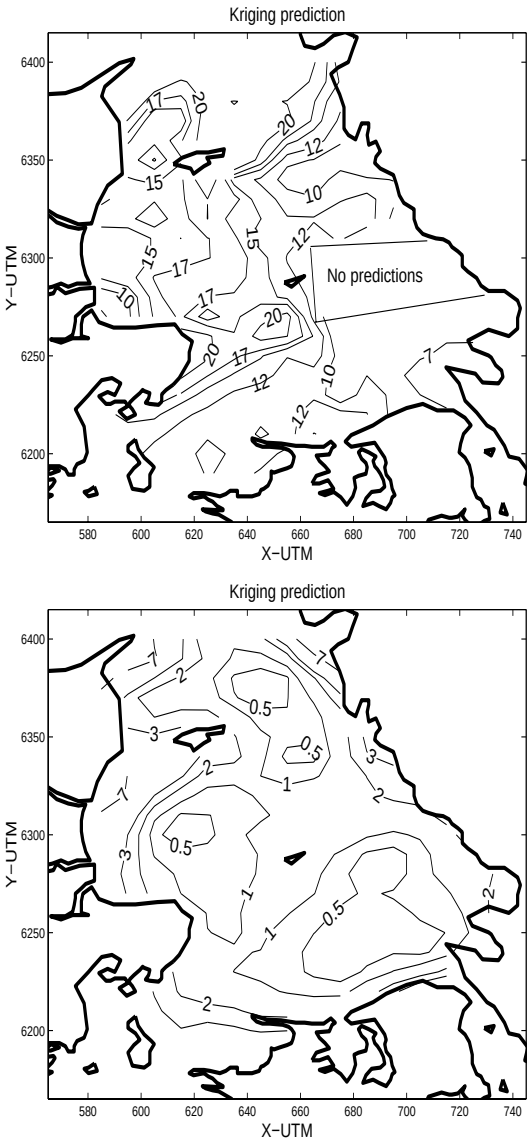

Figure I.1: *Spatial distribution of DIN, computed by three dimensional ordinary kriging. Upper: A week in the middle of March 1995. Lower: A week in the middle of October 1996.*

I.2 Dissolved Inorganic Phosphorus

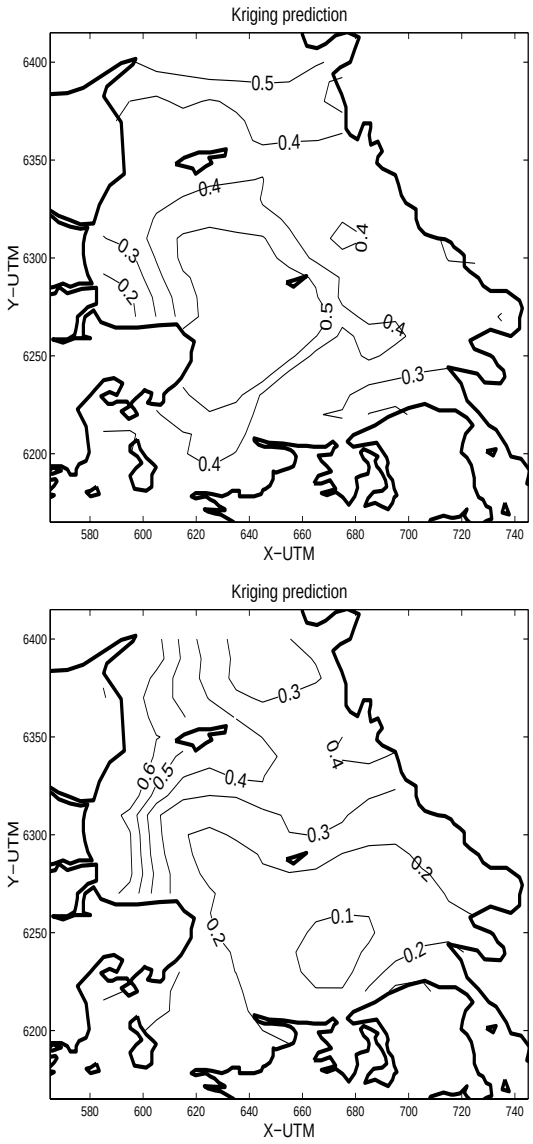

Figure I.2: *Spatial distribution of DIP, computed by three dimensional ordinary kriging. Upper: A week in the middle of March 1995. Lower: A week in the middle of October 1996.*

Appendix J
Parameters of ARIMA models

Station	$\overline{Y}$	ϕ_1	Θ_1
HALVAD	-0.065853144	0.4491854983	0.6074326427
L3	-0.069390589	0.4755571523	0.6039239959
L4	-0.070243257	0.4746997095	0.6051380139
L9	-0.070522971	0.4796859482	0.5930382353
N13	-0.06998247	0.4774552891	0.5949711052
N14	-0.069828729	0.4763022492	0.5957045618
N15	-0.070522971	0.4796859482	0.5930382353
N6	-0.070522971	0.4796859482	0.5930382353
N7	-0.064132579	0.4516444718	0.6240987142
P23	-0.064408115	0.459618356	0.6002236392
SI1	-0.069919182	0.4764689321	0.6018432374
SI2	-0.06206171	0.4301832339	0.5979627799
SI3	-0.070989451	0.4776637555	0.5979386351
SI4	-0.068626445	0.4764916063	0.6102238303
SI5	-0.071491614	0.4808991244	0.5972848729
SKALKORG	-0.070602001	0.4772147383	0.5994974422
VALO	-0.070536548	0.4658915657	0.6000686432
1001	0.0204411423	0.3576985125	0.4725439007
1002	-0.072490194	0.4783844372	0.5891424642
1004	-0.082972796	0.350435192	0.6968119704
1005	-0.070680788	0.4705666323	0.602467007
1007	-0.031682435	0.3063432442	0.5870834593
1008	-0.017803566	0.3091940073	0.4662645754
1009	-0.092823066	0.3587425331	0.5876345988
1141	-0.0699125	0.4751962962	0.6245302466
1142	-0.067943467	0.4718834796	0.5841680901
1143	-0.067676024	0.4795756285	0.5933581213
1230	-0.070994874	0.4761576129	0.5916681022
1234	-0.068926184	0.4797879837	0.5969140113
1257	-0.070522971	0.4796859482	0.5930382353
159092	-0.066099942	0.4751909984	0.6037267766

Table J.1: *Parameters of ARIMA models for different stations.*

Station	$\overline{Y}$	ϕ_1	Θ_1
190004	-0.042884484	0.3509501125	0.5410606363
1937	-0.069938841	0.3664904373	0.5553456235
1939	-0.076105158	0.3293855848	0.573568466
1993	-0.07941904	0.2664234561	0.5980550835
2000	-0.066094167	0.2369445433	0.5294629416
3302	-0.081331235	0.4295985908	0.5233236363
3310	-0.055173947	0.3969640934	0.5184890495
4002	-0.070522971	0.4796859482	0.5930382353
4010	-0.070522971	0.4796859482	0.5930382353
4017	-0.067481942	0.4851601328	0.5862742778
4023	-0.07208619	0.4756070877	0.599310311
403	-0.092094125	0.363244175	0.6560973097
4039	-0.070522971	0.4796859482	0.5930382353
406	-0.070522971	0.4796859482	0.5930382353
407	-0.070748842	0.4804300145	0.5945983934
4080	-0.077135757	0.4318895026	0.6070128598
4088	-0.071342241	0.473607908	0.5995557558
4089	-0.069570573	0.4703660415	0.6015347508
409	-0.117966171	0.3731339527	0.5365301495
413	-0.159254609	0.4013042786	0.5414548325
415	-0.097019308	0.3200963346	0.5360907285
416	-0.069618523	0.4813871707	0.5955561
418	-0.061632269	0.3925385411	0.6624372443
421	-0.070522971	0.4796859482	0.5930382353
4402	-0.10010693	0.2812385037	0.6113284623
4410	-0.076981892	0.3372927095	0.6639191491
727	-0.070511943	0.4791584578	0.591744479
905	-0.03030104	0.3101930785	0.6342209964
921	-0.106397629	0.2197114094	0.6102793882
925	-0.078876367	0.2878338342	0.6237615613
ALVSBROA	-0.070670962	0.4458692581	0.6031289154
OVF1	-0.062625474	0.4588513382	0.6144431682

Table J.2: *Parameters of ARIMA models for different stations.*

Bibliography

[Armstrong, 1984] Armstrong, M. (1984). Problems with universal kriging. *Mathematical Geology*, 16:101–108.

[Awad, 1996] Awad, A. (1996). Properties of Akaike information criterion. *Microelectronic Reliability*, 36:457–464.

[Caers, 2000] Caers, J. (2000). Adding local accuracy to direct sequential simulation. *Mathematical Geology*, 32:815–850.

[Carroll et al., 1997] Carroll, R., Chen, R., George, E., Li, T., Newton, H., Schmiediche, H., and Wang, N. (1997). Ozone exposure and population density in Harris County, Texas. *Journal of the American Statistical Association*, 92(438):392–404.

[Carstensen et al., 2000] Carstensen, J., Conley, D., Danielsson, Å., Rahm, L., Toompuu, A., and Karulis, B. (2000). *STAMP. Phase B Technical Report - Spatial heterogeneity and temporal reconstruction of data for phytoplankton modelling*. Nordisk Ministerråd.

[Carstensen et al., 1999] Carstensen, J., Conley, D., Danielsson, Å., Rahm, L., Toompuu, A., Yurkovskis, A., and Ikauniece, A. (1999). *STAMP. Phase A Technical Report - Temporal and spatial coverage of data for phytoplankton modelling*. Nordisk Ministerråd.

[Clausen, 1980] Clausen, F. (1980). *Using universal kriging for automatic contouring. Description of programs PTUK and CONTOUR*. IMSOR - DTH, Lyngby Denmark.

[Cleveland, 1979] Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836.

[Cleveland, 1988] Cleveland, W. (1988). Locally weighted regression: An

approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596–610.

[Cressie, 1985] Cressie, N. (1985). Fitting variogram models by weighted least squares. *Mathematical Geology*, 17:563–586.

[Cressie, 1993] Cressie, N. (1993). *Statistics for spatial data*. Wieley & Sons Inc.

[Danish-EPA, 2000] Danish-EPA (2000). *Aquatic Environment 1999, State of the Danish Aquatic Environment*. Environmental Investigations, No. 3 2000.

[Deutsch and Journel, 1992] Deutsch, C. and Journel, A. (1992). *GSLIB - Geostatistical Software Library and user's guide*. Oxford University Press.

[Edwards, 1984] Edwards, A. (1984). *An introduction to linear regression and correlation*. W.H. Freeman and company.

[Edwards, 1985] Edwards, A. (1985). *Multiple regression and the analysis of variance and covariance*. W.H. Freeman and company.

[Ersbøll, 1994] Ersbøll, A. (1994). *On the spatial and temporal correlations in experimentation with agricultural applications*. Ph.D. Thesis No. 7. IMM - DTU, Lyngby Denmark.

[Geer and Zuur, 1997] Geer, F. and Zuur, A. (1997). An extension of box-jenkins transfer/noise models for spatial interpolation of groundwater head series. *Journal of Hydrology*, 192:65–80.

[Haas, 1995] Haas, T. (1995). Local prediction of a spatio-temporal process with an application to wet sulfate deposition. *Journal of the American Statistical Association*, 90(432):1189–1199.

[Haas, 1998] Haas, T. (1998). Statistical assessment of spatio-temporal pollutant trends and meterological transport models. *Atmospheric Enviroment*, 32(11):1865–1879.

[Hansen et al., 2000] Hansen, J., Pedersen, B., Carstensen, J., Conley, D., Christiansen, T., Dahl, K., Henriksen, P., Josefson, A., Larsen, M., Lisbjerg, D., Lundsgaard, C., Markager, S., Rasmussen, B., Strand, J., Ærtebjerg, G., Krause-Jensen, D., Laursen, J., Ellermann, T., Hertel, O., Skjøth, C., Ovesen, N., Svendsen, L., and Pritzl, G. (2000). *Marine areas - State of the environment in 1999 (In Danish)*. Technical Report No. 333. National Environmental Research Institute, Denmark.

[Huang and Cressie, 1996] Huang, H. and Cressie, N. (1996). Spatio-temporal prediction of snow water equivalent using the Kalman filter. *Computational Statistics & Data Analysis*, 22:159–175.

[Hurvich and Simonoff, 1998] Hurvich, C. and Simonoff, J. (1998).

Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society*, pages 271–293.

[Isaaks and Srivastava, 1989] Isaaks, E. and Srivastava, R. (1989). *An introduction to applied geostatistics*. Oxford University Press.

[Journel and Huijbregts, 1978] Journel, A. and Huijbregts, C. (1978). *Mining geostatistics*. Academic Press, London.

[Journel and Rossi, 1989] Journel, A. and Rossi, M. (1989). When do we need a trend model in kriging? *Mathematical Geology*, 21(7):715–739.

[Kaluzny et al., 1998] Kaluzny, S., Vega, S., Cardoso, T., and Shelly, A. (1998). *S+ Spatial Stats. User's manual for Windows and UNIX*. Springer.

[Marcotte, 1991] Marcotte, D. (1991). Cokriging with Matlab. *Computers and Geosciences*, 17(9):1265–1280.

[Marsily, 1986] Marsily, G. (1986). *Quantitative hydrogeology*. Academic Press.

[McQuarrie et al., 1997] McQuarrie, A., Shumway, R., and Tsai, C. (1997). The model selection criterion AICu. *Statistics & probability letters*, 34:285–292.

[Meiring et al., 1998] Meiring, W., Guttorp, P., and Sampson, P. (1998). Space-time estimation of grid-cell hourly ozone levels for assessment of a deterministic model. *Enviromental and Ecological Statistics*, 5:197–222.

[Myers, 1982] Myers, D. (1982). Matrix formulation of co-kriging. *Mathematical Geology*, 14(3):249–257.

[Nielsen, 1994] Nielsen, A. (1994). *Analysis of regularly and irregularly sampled spatial, multivariate, and multi-temporal data*. Ph.D. Thesis No. 6. IMM - DTU, Lyngby Denmark.

[Nielsen, 1997] Nielsen, H. (1997). *LFLM Version 1.0. An S-PLUS/R library for locally weighted fitting of linear models*. Technical Report No. 22. IMM - DTU, Lyngby Denmark.

[Rossi et al., 1993] Rossi, R., Borth, P., and Tollefson, J. (1993). Stochastic simulation for characterizing ecological spatial patterns and appraising risk. *Ecological Applications*, 3:719–735.

[Rouhani and Myers, 1990] Rouhani, S. and Myers, D. (1990). Problems in space-time kriging og geohydrological data. *Mathematical Geology*, 22(5):611–623.

[Rouhani and Wackernagel, 1990] Rouhani, S. and Wackernagel, H. (1990). Multivariate geostatistical approach to space-time data analysis. *Water Resources Research*, 26(4):585–591.

[Shumway, 1988] Shumway, R. (1988). *Applied statistical time series analysis*. Prentice Hall.

[Zimmerman, 1993] Zimmerman, D. (1993). Another look at anisotropy in geostatistics. *Mathematical Geology*, 25:453–471.

.

.